

Volume 33 Number 3 October 2009

ISSN 0350-5596

Informatica

**An International Journal of Computing
and Informatics**

Special Issue:

Information Reuse and Integration

Guest Editors:

Reda Alhajj

Kang Zhang



EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatika is a journal primarily covering the European computer science and informatics community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatika is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatika is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Executive Editor – Editor in Chief

Anton P. Železnikar
Volaričeva 8, Ljubljana, Slovenia
s51em@lea.hamradio.si
<http://lea.hamradio.si/~s51em/>

Executive Associate Editor - Managing Editor

Matjaž Gams, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 251 93 85
matjaz.gams@ijs.si
<http://dis.ijs.si/mezi/matjaz.html>

Executive Associate Editor - Deputy Managing Editor

Mitja Luštrek, Jožef Stefan Institute
mitja.lustrek@ijs.si

Executive Associate Editor - Technical Editor

Drago Torkar, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 251 93 85
drago.torkar@ijs.si

Editorial Board

Juan Carlos Augusto (Argentina)
Costin Badica (Romania)
Vladimir Batagelj (Slovenia)
Francesco Bergadano (Italy)
Marco Botta (Italy)
Pavel Brazdil (Portugal)
Andrej Brodnik (Slovenia)
Ivan Bruha (Canada)
Wray Buntine (Finland)
Hubert L. Dreyfus (USA)
Jozo Dujmović (USA)
Johann Eder (Austria)
Vladimir A. Fomichov (Russia)
Maria Ganzha (Poland)
Marjan Gušev (Macedonia)
Dimitris Kanellopoulos (Greece)
Hiroaki Kitano (Japan)
Igor Kononenko (Slovenia)
Miroslav Kubat (USA)
Ante Lauc (Croatia)
Jadran Lenarčič (Slovenia)
Huan Liu (USA)
Suzana Loskovska (Macedonia)
Ramon L. de Mantras (Spain)
Angelo Montanari (Italy)
Deepak Laxmi Narasimha (Malaysia)
Pavol Návrat (Slovakia)
Jerzy R. Nawrocki (Poland)
Nadja Nedjah (Brasil)
Franc Novak (Slovenia)
Marcin Paprzycki (USA/Poland)
Gert S. Pedersen (Denmark)
Ivana Podnar Žarko (Croatia)
Karl H. Pribram (USA)
Luc De Raedt (Belgium)
Dejan Raković (Serbia)
Jean Ramaekers (Belgium)
Wilhelm Rossak (Germany)
Ivan Rozman (Slovenia)
Sugata Sanyal (India)
Walter Schempp (Germany)
Johannes Schwinn (Germany)
Zhongzhi Shi (China)
Oliviero Stock (Italy)
Robert Trappl (Austria)
Terry Winograd (USA)
Stefan Wrobel (Germany)
Konrad Wrona (France)
Xindong Wu (USA)

Editorial: Special Issue on Information Reuse and Integration

The increasing volumes and dimensions of information have dramatic impact on effective decision-making. To remedy this situation, Information Reuse and Integration (IRI) seeks to maximize the reuse of information by creating simple, rich, and reusable knowledge representations and consequently explores strategies for integrating this knowledge into legacy systems. IRI plays a pivotal role in the capture, representation, maintenance, integration, validation, and extrapolation of information; and applies both information and knowledge for enhancing decision-making in various application domains. This special issue includes six papers.

The six papers cover an interesting range of topics from software reuse and measurements to information and knowledge integration. The first two papers discuss issues and techniques in knowledge-based approaches to software measurements and evaluation. The next two papers investigate the development of Web-based intelligent systems for image tagging and user dialog. The last two papers present optimization issues and mappings between OO databases and XML.

The first paper by Folleco, Khoshgoftaar, Hulse and Napolitano presents their investigation into the robustness of a variety of common-used learning algorithms relative to low quality, class imbalanced measurement data. The authors identify learners from a total of 11 classification algorithms with robust performance in the presence of low quality imbalanced measurement data. Four performance metrics suitable for class imbalanced data were used to measure learner performance. Based on their results, the results demonstrate that the quality of the measurement data can impact classification performance significantly. The authors recommend using the random forest ensemble learning technique for building classification models from software measurement data, regardless of the quality and class distribution of the data.

The paper by Far, Mudigonda and Elamy focuses on the design and development of a General Purpose Software Evaluation (GPSE) system that uses statistical methods based on Multidimensional Weighted Attribute Framework (MWAF) for the evaluation of software systems. The architectural elements of MWAF are essentially survey questionnaire which gathers information from several domain experts. The GPSE system then applies principles of Analysis of Variance (ANOVA) and Tukey's pairwise comparison tests on the collected data to arrive at selection of the best suited alternative for the given problem. The authors have fully implemented the GPSE system and tested it on several projects including evaluation of multi-agent development methodologies and selection of COTS products.

The paper of Özyer proposes a collaborative infrastructure that helps users store pictures that they have seen on Web pages during surfing. The system

consists of two modules: an add-on application that will work on Firefox Mozilla browser and the Web page itself. The add-on application helps user tag and store pictures that will remind of something meaningful to him. The author has implemented a system that lets users bookmark pictures they want to keep with their tag info according to image content. Later on the users may reorganize the tags and may also perform search by using tag information rather than image content.

The paper by Shibata, Nishiguchi, and Tomiura discusses a type of open-ended dialog system that generates appropriated responses based on the vast amount of Web documents. Generated from Web documents, the candidate corpus in Japanese could maintain surface cohesion and semantic coherence. The authors have prototyped the system that is capable of meaningful conversation on various topics. They have experimented on a conversation about movies, with the result showing that the system could generate 66% appropriate responses.

In their paper, Faraz Rafi, Zaidi and Levis present two algorithms for the optimal selections of a sequential evolution of actions, in conjunction with the preconditions of their environment and their effects, depicted by Activation Timed Influence Nets, given a set of preconditions. A special case for the two algorithms is also considered where the selection of actions is further constrained by the use of dependencies among them. The algorithms are based on two different optimization criteria: one maximizes the probability of a given set of target effects, while the other maximizes the average worth of the effects' vector.

The last paper by Naser, Alhadjj, and Ridley presents a novel approach for mapping an existing object-oriented database into XML and vice versa. They first derive the so-called object graph based on characteristics of the schema to be mapped. For object-oriented schema, the object graph simply summarizes and includes all nesting and inheritance links. The inheritance is then simulated in terms of nesting to get a simulated object graph. Doing so, everything in a simulated object graph can be directly represented in XML, and thus the mapping of the actual data from the object-oriented database into corresponding XML document(s) is easily achieved. The mapping from XML into object-oriented database is similarly achieved, where everything in a simulated object graph could be directly represented in object-oriented database.

Finally, we would like to thank all the authors for their efforts in enhancing and extending their papers for this special issue. Thanks also the Informatica team for their assistance in making this special issue published in a timely fashion.

*Reda Alhadjj, University of Calgary, Canada
Kang Zhang, University of Texas at Dallas, USA*

Identifying Learners Robust to Low Quality Data

Andres A. Folleco, Taghi M. Khoshgoftaar, Jason Van Hulse and Amri Napolitano
 Florida Atlantic University,
 Boca Raton, Florida, USA
 E-mail: {andres, taghi}@cse.fau.edu , {jvanhulse, amrifau}@gmail.com

Keywords: quality of data, class imbalance, random forest, robust learning

Received: October 12, 2008

Low quality or noisy data, which typically consists of erroneous values for both dependent and independent variables, has been demonstrated to have a significantly negative impact on the classification performance of most learning techniques. The impact on learner performance can be magnified when the class distribution is imbalanced or skewed. Unfortunately in real world environments, the presence of low quality imbalanced data is a common occurrence. In most scenarios, the actual quality of such datasets is unknown to the data mining practitioner. In this study, we identify learners (from a total of 11 classification algorithms) with robust performance in the presence of low quality imbalanced measurement data. Noise was injected into seven imbalanced software measurement datasets, initially relatively free of noise. Learners were evaluated using analysis of variance models based on their performance as the level of injected noise, the number of attributes with noise, and the percentage of minority instances containing noise were increased. Four performance metrics suitable for class imbalanced data were used to measure learner performance. Based on our results, we recommend using the random forest ensemble learning technique for building classification models from software measurement data, regardless of the quality and class distribution of the data.

Povzetek: Predstavljena je metoda za identificiranje robustnih klasifikatorjev pri šumnih podatkih.

1 Introduction

Not only are real-world datasets often class imbalanced, but typically their attributes (including the class) can contain erroneous values that may negatively impact learning performance [17, 29, 32]. Consequently, it is not uncommon for empirical software engineering practitioners to construct learners using suboptimal or low quality imbalanced data. Note that in this work¹, only binary classification problems were considered. In software quality classification, class imbalance occurs when the number of fault-prone (fp) modules is significantly outnumbered by the number of not fault-prone (nfp) program modules. No other related work in the software quality prediction domain was found that evaluated the robustness of learning techniques, using four performance metrics, relative to low quality imbalanced measurement data. In this study, simulated noise was injected in both the independent attributes as well as the class (i.e., labeling errors) of seven class imbalanced software engineering measurement datasets. The substantial and significant scope of our experiments make this study truly unique for the domain of empirical software engineering.

The robustness of 11 distinct algorithms trained using low quality imbalanced real-world measurement data is

¹This is an expanded version, by invitation, of the work accepted and presented at the 2008 IEEE International Conference on Information Reuse and Integration- IRI'08 [11]

evaluated by analyzing four metrics related to learner performance. Noisy data was simulated by injecting artificially induced, domain realistic noise into the independent attributes and class of seven real-world class imbalanced software measurement datasets as explained in Section 4.3. Four performance metrics particularly well suited to deal with class imbalanced data were selected for this study. The area under the ROC curve (AUC), the area under the Precision-Recall curve (PRC), the Kolmogorov-Smirnov statistic (KS), and the F-measure (FM) test statistic were selected (see Section 4.2) to measure the impact on learning performance. The overall impact of noise on each learner was measured as the level of noise, the number of attributes with injected noise, and the percentage of minority instances containing noise increased across all the datasets (see Section 5). Furthermore, we illustrate the impact of the factor interaction between the overall noise levels and the percent of minority instances with noise. Conclusions are presented in Section 6.

2 Related work

Regardless of the perceived soundness of a preferred classification algorithm, learning from low quality class imbalanced data will very likely result in biased and suboptimal performance. Several studies in classification initiatives have demonstrated that the presence of noisy values

(mainly corrupted class labels) in the training dataset will likely impact the predictive accuracy of a learning algorithm [17, 32]. Arguably, the main factors determining data quality include independent attribute noise [27, 32], dependent attribute or class noise [16, 5], and missing or omitted values [20]. The data quality characteristics considered in this study include the presence of noise in both the independent and dependent (class) attributes.

In addition, the real-world measurement datasets used in this work are inherently class imbalanced. A software measurement dataset selected for binary classification tasks (we only consider binary classification in this work) is said to be *imbalanced* if the number of positive (fault-prone or fp) class modules is less than the number of negative (not fault-prone or nfp) class modules. Typically, minority instances make up the positive class, while the negative class is composed of the majority instances. Frequently in real-world scenarios, the practitioner is primarily interested in identifying examples from the minority group. In the context of software engineering, this is frequently seen in mission critical applications, where a premium is placed on identifying fault-prone modules during the test phase. Low quality real-world datasets containing class imbalance distributions pose a great challenge to any data mining and machine learning effort. In fact, researchers contend that data quality and class imbalance can significantly impact the reliability of machine learning models in real-world scenarios, and consequently demand empirical consideration and experimental evaluation [10, 35]. Data with such characteristics can be found in a wide variety of application domains besides software quality classification, including for example network intrusion detection [19] and fraud detection [13].

In our study, we investigate the robustness of 11 learning algorithms in the presence of low quality class imbalanced real-world software measurement datasets, initially relatively free of noise. Weiss [29] performed a preliminary study of the effects of class and attribute noise on classification with simulated datasets using the C4.5 learner. However, no real world datasets or additional learners were used. Weiss and Provost [30] investigated the effect of class distribution and training set size on classification performance. Their results imply that the natural class distribution generates higher overall accuracy. However, the overall accuracy rate is often not considered an appropriate measurement method when dealing with class imbalanced datasets. Furthermore, they also suggested that a more balanced class distribution can result in higher AUC values. Van Hulse et al. [28] conducted an investigation of the impact on classification performance from class imbalanced data injected with simulated class noise. The AUC metric was used to measure the learning performance. The independent attributes were not considered for noise injection in their study. Studies have been conducted to investigate the impact of noise using classification performance-enhancing techniques like cost sensitive learning [33, 34] with various cost ratios [9]. In general, relatively few stud-

ies have evaluated the impact of noise (in the class only) in imbalanced data. To our knowledge, no related works have investigated the impact of class and attribute noise on learners constructed from class skewed measurement datasets using several performance metrics suitable for class imbalanced distributions.

3 Learning algorithms

The open-source Java based Weka data mining and machine learning tool [31] was used to implement the learners in this study. The learners used were selected because most of them are commonly used in class imbalance scenarios and several are also used in the software engineering and software quality classification domain. The default parameter values of some of the techniques were changed when their respective classification performances improved substantially.

3.1 Random forest

The *random forest* (RF) classifier was developed by Breiman [3]. RF is a powerful, relatively new approach to data exploration, data analysis, classification, and predictive modeling. A random forest is a collection of unpruned, CART-like trees [4] following specific rules for tree growing, tree combination, self-testing, and post-processing. Trees are grown using binary partitioning in which each parent node is split into no more than two children. Each tree is grown on a different random subsample of the training data. Randomness is also injected into the tree split selection process. RF selects a relatively small subset of available attributes at random. In the Weka tool, the default for the *numFeatures* parameter uses $\lfloor \log_2 M + 1 \rfloor$ attributes selected at random for each node in the tree where M is the original number of independent attributes in the data. Attribute selection significantly speeds up the tree generation process. Once a node is split on the best splitter attribute, the process is repeated entirely on each child node. Then, a new list of predictive attributes is selected at random for each node. The trees must remain unpruned to their absolute maximum size in order to maximize the chances of including important attributes into the trees.

Bootstrapping is a process of random sampling with replacement from the training dataset. By applying bootstrapping during the tree induction process, approximately 37% of the observations in the training dataset are not used and form the out-of-bag samples. Another important parameter in the RF algorithm is the number of trees *numTrees* in the ensemble. The default value for *numTrees* in Weka is 10, however previous research by our group [15] found 100 to be a more appropriate value, so 100 was used instead. Combining results from multiple models (trees) generally yields better performance results than those obtained from a single model. Combining trees by averaging the votes will only be beneficial if the trees are different from each other. RF induces vastly more

between-tree variation by forcing random splits on different predictive attributes. Having a diverse collection of robust trees lowers the overall error rate, avoids over-fitting training data, imbues a substantial resilience to noisy values, and therefore enhances the performance of the RF [3] ensemble classifier.

3.2 k Nearest Neighbors (Two versions)

K nearest neighbors [1] (kNN) is called IBk in the Weka implementation of an instance-based classification technique using k nearest neighbors. The class of a test case is predicted by majority voting of the k nearest neighbors. If only one nearest neighbor is selected to predict the class of a test instance, especially in the presence of outliers and/or low quality data, it may lead to increased inaccuracy. The Euclidean distance is often used as a similarity function to determine the potential candidate nearest neighbors. A possible disadvantage of IBk is that its computation time depends on the size of the number of nearest neighbors. As the number of nearest neighbors increases, so do the computational resources and time needed. In our experiments, kNN classifiers were built with changes to two parameters: The ‘distanceWeighting’ parameter was set to ‘Weight by 1/distance’ and two different ‘kNN’ classifiers were built using $k = 2$ and $k = 5$ neighbors. These were denoted ‘2NN’ and ‘5NN,’ respectively.

3.3 C4.5 Decision Tree (Two versions)

C4.5 [23] is a benchmark decision tree classification algorithm. J48 is Weka’s implementation of this algorithm. It is an inductive supervised classifier that uses decision trees to represent the underlying structure of the input data. The algorithm has four major components: the decision tree generator, the production rule generator, the decision tree interpreter, and the production rule interpreter. These modules are used for constructing and evaluating the classification tree models. The algorithm begins with an empty tree, to which is added decision and leaf nodes, starting at the root (top) node. In the next step, using one of the attributes x_j ($j = 1, \dots, m = \#attributes$) the instances in the root node are split into two (or more) child nodes N_l and N_r . For example, if x_j is a continuous attribute and $i = 1, \dots, n = \#instances$, we define $N_l = \{ \mathbf{x}_i \in D \mid x_{ij} < t \}$, and $N_r = \{ \mathbf{x}_i \in D \mid x_{ij} \geq t \}$ for some value of t from x_j . C4.5 evaluates each of the attributes x_j to determine the best split at each tree node. The splitting process is recursively applied to each of the resulting child/leaf nodes until some stopping criteria is met. After the tree is fully built, C4.5 provides the option to prune sections of the tree to avoid over-fitting. Two different versions of the C4.5 classifier were used in our experiments. The version we call C4D uses the default parameter settings in Weka to build the tree(s). The version of C4.5 we call C4N disables decision-tree pruning and enables Laplace smoothing. These settings were recommended for speed

and performance by Weiss [30].

3.4 Support Vector Machine

The *support vector machine* (SVM) classifier, called SMO in Weka, can be used to solve two-class (binary) classification problems [24]. These classifiers find a maximum margin linear hyperplane within the instance space that provide the greatest separation between the two classes. Instances that are closest to the maximum margin linear hyperplane form the support vectors. Once the instances that form the support vector have been identified, the maximum margin linear hyperplane can then be constructed. We consider the following linear hyperplane separating two classes [31] from:

$$x = b + \sum \alpha_i y_i a(i) \cdot a \quad (1)$$

where i is a support vector, y_i is the class of the training instance $a(i)$, b , and α_i are numeric parameters that are adjusted based on the classification algorithm. The term $a(i) \cdot a$ represents the dot product of the test instance with one of the support vectors. Identifying a solution to the linear hyperplane by Equation 1 is the same as solving a constrained quadratic optimization problem. In this study, the SVM classifier had two changes to the default parameters: the complexity constant ‘c’ was set to 5.0 and ‘buildLogisticModels’ was set to ‘true.’ By default, a linear kernel was used.

3.5 Logistic Regression

Logistic regression (LR) is a statistical regression model [14] that can be used to estimate two-class classification problems. Using the training data instances as input, a logistic regression model is created which is used to decide the class membership of the test data instances. The logistic function used for modeling may be defined as follows:

$$f(z) = \frac{1}{1 + e^{-z}}$$

where z denotes the input instances from the training data and e denotes the base of the natural logarithm. The logistic function can take as input z any negative or positive value, while the output of the function is always in the range of zero to one. The output of the logistic regression classifier expresses the probability of an instance belonging to a certain class. An instance of a training dataset that is used as input for the logistic regression model can be:

$$z = \omega_0 + (\omega_1 \times x_1) + (\omega_2 \times x_2) + (\omega_3 \times x_3) + \dots + (\omega_k \times x_k)$$

where ω_0 is known as the intercept, $\omega_1, \omega_2, \omega_3, \dots, \omega_k$ are called the regression coefficients or model weights, $x_1, x_2, x_3, \dots, x_k$ denote the corresponding instance attribute values from the training set, and k is the total number of

attributes considered. Each of the weights describes the impact of the corresponding attribute value on z . Recall that z is used to determine the class membership of the instance. The weights must be adjusted in order to optimize the logistic regression model on the training data. The log-likelihood of the model is used to estimate the goodness of fit and the weights for the model are chosen to maximize this log-likelihood function. In this study, the Weka default parameter settings were used for this classifier.

3.6 Naive Bayes

Naive Bayes (NB) is a simple and fast algorithm based on the Bayesian rule of conditional probability [12]. NB assumes that attributes are independent of each other within a given class. Even though this condition may not be realistic in real-world data, NB has been known to perform well with this assumption of attribute independence. The algorithm estimates the class probabilities $P(fp|\mathbf{x})$ using the Bayes theorem by considering the following expression,

$$\begin{aligned} P(fp|\mathbf{x}) &= \frac{P(fp, \mathbf{x})}{P(\mathbf{x})} = \frac{P(\mathbf{x}|fp)P(fp)}{P(\mathbf{x})} \\ &= \frac{P(x_1|fp)\dots P(x_m|fp)P(fp)}{P(\mathbf{x})} \end{aligned} \quad (2)$$

where m is the number of attributes and fp is the fault-prone (or the minority/positive) class. The independence of attributes can be used to factor the class conditional probability $P(\mathbf{x}|fp)$ into $P(x_1|fp)\dots P(x_m|fp)$. This transformation allows for the estimation of m one-dimensional distributions $P(x_j|fp)$, $j = 1, \dots, m$, instead of estimating the joint distribution of $P(\mathbf{x}|fp)$ from the data. In our experiments, the default parameter values were used within the Weka implementation of this algorithm.

3.7 Rule-Based Classifier

RIPPER (Repeated Incremental Pruning to Produce Error Reduction - RIP) is a rule-based classifier and is named JRip [6] in Weka. This algorithm was introduced by Cohen [6] as a fast classifier of "If-Then" classification rules. Initially, the algorithm splits the training dataset into two parts. One part is used to induce rules, while the second part of the training dataset is used to validate the induced rules. If a rule's classification accuracy falls below a minimum accuracy threshold, the rule is then eliminated from the model. RIP imposes a rule induction ordering, minority class rules first, followed by the majority class. Once all of the instances in the minority class have been covered, a default rule is generated to classify the majority data. This feature reduces the description length of a rule set. The default Weka parameters for this classifier were not changed in our experiments.

3.8 Multilayer Perceptron Networks

Multilayer Perceptron (MLP) is a network of perceptrons [21]. A perceptron is the simplest neural network representing a linear hyperplane within instance space. MLPs can be used to solve complex problems. Every MLP contains an input and output layer and at least one hidden layer. A layer is an arrangement of neurons that include hidden ones which do not have any connections to external sources or environments. MLPs are typically implemented as a back-propagation neural network. In a back-propagation neural network, the error from an output neuron is fed back to the same neuron. The neuron output is the thresholded weighted sum of all its inputs from the previous layer. This process is continued iteratively until the error can be tolerated or reaches a specific threshold. MLPs map the instances in the input data onto a set of output values using three or more layers of neurons. Activation functions are used to calculate the output from the input into the neurons, which is comprised of weighted sums of the outputs from the previous layer. Two parameters from MLP were changed from their default values. The 'hiddenLayers' parameter was changed to '3' to define a network with one hidden layer containing three nodes, and the 'validationSet-Size' parameter was changed to '10' to cause the classifier to leave 10% of the training data aside to be used as a validation set to determine when to stop the iterative training process.

3.9 Radial Basis Function Networks

Radial basis function networks (RBF) are another type of artificial neural network [21]. They are similar to MLPs except in the method used for processing the data within a single hidden layer. The hidden layer is of high enough dimension which provides a nonlinear transformation from the input space. The output layer in these networks provides a linear transformation from the hidden-unit space to the output space. When using RBF neurons, a category of patterns can be regarded as a Gaussian distribution of points in pattern space. The neuron fires when its input is sufficiently close to activate the Gaussian. Inputs are encoded by computing a measure of how close they are to a receptive field, e.g., distance between the input vector and the centroid of that neuron. In this study, the only parameter change for RBF was to set the parameter 'numClusters' to 10.

4 Empirical methodology

4.1 Software Measurement Datasets

The JM1, CM1, MW1, PC1, KC1, KC2, and KC3 datasets were obtained from the NASA Metrics Data Program (MDP). Learners were built using 13 basic metrics [16] as independent variables. The dependent variable was a binary module-class label, i.e., fault-prone or not fault-

prone. The minority class is represented as the positive (fault-prone) class, while the majority class is represented as the negative (not fault-prone) class. All the instances in the data represent measurements taken from the software modules.

The KC1, KC2, and KC3 projects comprise a mission control system and were developed and implemented by different personnel with no overlapping software components. The KC1 system, implemented in C++, is a software component of a large ground system. The KC2 system, implemented in C++, is the science data processing component of a storage management system used for ground processing data. The KC3 system, written in Java, is software developed for collection, processing, and delivery of satellite meta-data. The PC1 system, implemented in C, is flight control software from an earth orbiting satellite. The JM1 project, implemented in C, is a real-time ground system that uses simulation to generate predictions for space missions. The MW1 project, implemented in C, is the control software of a zero gravity experiment related to combustion. The CM1 project, implemented in C, is a science instrument system used for mission measurements. The software modules fault data obtained for the software projects indicated the number of faults detected during the corresponding software development cycles.

A rule-based noise filter was applied to the CM1, MW1, PC1, KC1, KC2, and KC3 datasets to identify and remove noisy instances [18]. Table 1 provides details about the seven initial² datasets and their respective cleansed versions. See Van Hulse and Khoshgoftaar [26] for a detailed discussion of the cleansing process for JM1. The ‘*i/c*’ sub-headers indicate the initial and cleansed number of instances of a dataset, listed as ‘#initial / #cleansed’. The row labeled ‘P’ indicates the number of positive examples in the initial and cleansed datasets. Likewise, the row labeled ‘N’ indicates the number of negative examples. The ‘P + N’ row contains the total number of instances in the initial and cleansed datasets, respectively. The ‘%P’ row contains the level of imbalance present in the initial and cleansed datasets. For example, PC1 initially contained 1107 instances, of which 6.9% were positive. After cleansing, 703 total instances remained of which 7.5% were positive. Only the cleansed datasets were used in this work because they were subjected to a methodical and carefully designed noise cleansing process developed by Khoshgoftaar et al. [18] (see also Van Hulse [25] for a discussion of the noise cleansing procedure). The motivation for using relatively cleansed datasets before actually injecting the noise is to ensure the reliability of the results. Adding noise to inherently low quality data can significantly bias and compromise the reliability of any results derived from using such data.

Table 2 contains the overall classification performance obtained across all eleven classifiers for each cleansed dataset. According to the AUC, PRC, KS, and FM val-

ues (described in Section 4.2), the best classification performance was obtained using the largest (and relatively cleanest) dataset, JM1. The second best performance was obtained using KC1 (second largest), and the worst performance was obtained using MW1 (the most imbalanced and nearly the smallest dataset). The average values shown in the ‘Avg’ row were calculated across all seven datasets.

4.2 Performance Metrics

In a binary decision problem, a learner labels examples as either positive or negative. If very few examples belong to the positive class (as few as 1% or less), a learner could obtain an overall accuracy of 99% by just classifying all instances as negative. This method is useless in a domain like software quality classification because the examples of interest are typically from the positive class. Thus, performance metrics such as accuracy or the misclassification rate are inappropriate for substantially class imbalanced data.

The Receiver Operating Characteristic curve [22] (ROC) graphs true positive rates on the *y*-axis versus the false positive rates on the *x*-axis. The resulting curve illustrates the trade-off between detection and false alarm rates. Often, performance metrics consider only the default decision threshold of 0.5. ROC curves illustrate the performance across all decision thresholds. The threshold independent nature of ROC curves makes them well suited for describing the classification performance of models built on class imbalanced data. For a single numeric measure, the *area under the ROC curve* (AUC) is widely used, providing a general idea of the predictive potential of the classifier. Two classifiers can be evaluated by comparing their AUC values. Provost and Fawcett [22] give an extensive overview of ROC curves and their potential use for optimal classification.

The Precision-Recall curve [8] (PR) provides a different perspective regarding a classifier’s performance on class imbalanced datasets. Precision measures that fraction of instances classified as positive that are truly positive. Recall measures the fraction of positive instances that have correct labels. The PR curve graphs recall on the *x*-axis and precision on the *y*-axis. A single numeric measure for the PR curve is the *area under the PR curve* (PRC). Often, a large change in the number of false positives can lead to a small change in the false positive rate which is used in ROC analysis. On the other hand, PR analysis typically compares false positives to true positives rather than true negatives, encapsulating the impact of the large number of negatives instances on classification performance. However, a classifier that optimizes the area under the ROC is not guaranteed to optimize the area under the PR curve [8]. The expressions for precision and recall (which is the same as the true positive rate) are as follows:

²JM1 is an exception to these initial datasets because it is a subset of the much larger original JM dataset [25].

Table 1: Datasets *P* Positive & *N* Negative Instance Distributions

Instance	JM1 (i/c)	PC1 (i/c)	CM1 (i/c)	MW1 (i/c)	KC1 (i/c)	KC2 (i/c)	KC3 (i/c)	Total
<i>P</i>	470/235	76/53	48/39	31/20	325/271	106/82	43/38	1099/738
<i>N</i>	2393/2210	1031/650	457/277	372/291	1782/1093	414/333	415/264	6864/5118
<i>P + N</i>	2863/2445	1107/703	505/316	403/311	2107/1364	520/415	458/302	7963/5856
% <i>P</i>	16.4/9.6	6.9/7.5	9.5/12.3	7.7/6.4	15.4/19.9	20.4/19.8	9.4/12.6	13.8/12.6

Table 2: Classification Performance by Cleansed Dataset

Data	AUC	Data	PRC	Data	KS	Data	FM
JM1	0.9987	JM1	0.9956	JM1	0.9974	JM1	0.9972
KC1	0.9977	KC1	0.9923	KC1	0.9763	KC1	0.9739
KC2	0.9922	KC2	0.9774	PC1	0.9607	KC2	0.9471
PC1	0.9915	PC1	0.9650	KC2	0.9532	CM1	0.9350
KC3	0.9865	CM1	0.9595	KC3	0.9521	PC1	0.9301
CM1	0.9837	KC3	0.9580	CM1	0.9487	KC3	0.9244
MW1	0.9767	MW1	0.9266	MW1	0.9428	MW1	0.9040
Avg	0.9897		0.9678		0.9616		0.9445

$$Precision = \frac{\#TP}{\#TP + \#FP} \quad (3)$$

$$Recall = \frac{\#TP}{\#TP + \#FN}$$

Similarly to the PRC curve, the F-Measure (FM) is derived from *recall* and *precision*. There is a decision threshold parameter required for this metric, and the default decision threshold of 0.5 was used in this work. Further, the expression defining the FM metric has a tunable parameter β used to indicate the relative importance of recall and precision. Typically, one can alter β to place more emphasis on either *recall* or *precision*. In this study, $\beta = 1$.

$$FM = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$$

The Kolmogorov-Smirnov significance test [7, 13] measures the maximum difference between the empirical distribution function of the posterior probabilities $p(x)$ of instances in each class. Let $i \in \{positive | negative\}$ and $F_i(t) = P(p(x) \leq t | i)$, $0 \leq t \leq 1$. The $F_i(t)$ can be estimated by the proportion of class (*positive | negative*) instances $\leq t$.

$$F_i(t) = \frac{\#class(i) \text{ instances with posterior probability} \leq t}{\#class(i) \text{ instances}}$$

Therefore, the KS statistic is defined as follows:

$$KS = \max_{t \in [0,1]} |F_{positive}(t) - F_{negative}(t)|$$

As the separation between the two distribution functions becomes larger, the distinction between the two classes a

classifier has made will also improve. The maximum possible value for the KS is one (representing perfect separation), with a minimum of zero. The KS statistic is a commonly used metric of classifier performance in the credit scoring application domain [13].

4.3 Noise Injection Procedure

This section describes the noise injection procedure employed in our empirical study. Note that attribute noise was injected into all selected instances in the derived datasets, while class noise was only injected into the training instances, so that the class labels for the test instances were left uncorrupted. In order to add attribute noise into the datasets, the level of attribute noise (L^a) and the number of significant attributes to be injected with the noise (N^a) were the parameters considered. In the case of class noise, the level of class noise (L^c) and the percentage of instances with class noise injected that were originally from the positive (fault-prone) class (L^m) were the parameters considered. In this study, the noise injection procedure consisted of both attribute and class noise and therefore all four parameters (L^a , N^a , L^c , L^m) were considered.

4.3.1 Class Noise

The class was injected with noise by swapping the respective class label of some training set instances, e.g., positive \rightarrow negative or negative \rightarrow positive. The number of training set instances injected with class noise was a function of two parameters, L^c and L^m . Five levels of class noise, $L^c \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$, were used. The actual number of training dataset instances to be injected with class noise was calculated as $2 \times L^c \times |P|$, where $|P|$ is the number of positive instances in the dataset. For example, a

dataset with 1000 instances, 10% fault-prone modules (100 positive modules), and a L^c of 30% would have a total of 60 instances ($2 \times 30\% \times 100$) injected with noise.

Further, five percentages of positive instances, $L^m \in \{0\%, 25\%, 50\%, 75\%, 100\%\}$, were corrupted from the positive to the negative class. That is, five different relative proportions of positive \rightarrow negative versus negative \rightarrow positive were used in this study. Continuing with the previous example, suppose that the percentage of class noise corrupted from the positive class was $L^m = 75\%$. Then of the 60 noisy training set instances, 45 (or 75% of 60) of these will be from the positive class. The remaining 15 instances with injected noise will be from the negative class. Therefore, a randomly selected group of 45 instances from the 100 positive instances in the initial training dataset will have their respective class labels switched from positive to negative. In a similar fashion, a randomly selected 15 instances from the 900 negative instances in the initial training dataset, will have their class labels changed from negative to positive. Once the noise injection process is completed, the corrupted dataset of this example would have a total of $100 - 45 + 15 = 70$ positive and $900 + 45 - 15 = 930$ negative instances.

4.3.2 Independent Attribute Noise

Independent attributes (software metrics) were injected with noise at five levels, with $L^a \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$. Noise was first injected into the most significant predictive attribute (based on the two-sample KS significance test - see Section 4.2), creating a total of 35 derived datasets (seven initial datasets and five levels of L^a). The next 35 datasets were obtained by corrupting both the most and second-most significant attributes. This procedure was repeated until the seven most significant attributes ($N^a \in \{1, 2, 3, 4, 5, 7\}$) were corrupted. The results of six attributes injected with noise were excluded from this study because of similarities to the results obtained when $N^a = 7$. A noise level of 10% implied that the values for the selected attributes were corrupted for 10% of the instances. Noise was injected by replacing the selected attribute value with a randomly selected attribute value reflecting an instance of the opposite class. For a given injected noise level, the negative and positive proportions of the instances injected with noise was approximately the same as the negative and positive proportions of the given dataset. For example, if the given dataset had a proportion of 80:20 for negative:positive instances and if 180 instances were injected with noise, then the set of instances to be corrupted with attribute noise would have contained 144 negative and 36 positive instances.

Since we are evaluating the impact of noise on classifier performance, it is sensible to inject noise into attributes that are useful for differentiating between *fp* and *nfp* instances. In the case of noise injected into a single attribute, if the chosen attribute was not useful for prediction, then

the classifiers can easily circumvent the effects of attribute noise. For the purposes of this study, attribute significance was evaluated using a two-sample KS significance test. Other studies can use other attribute significance algorithms and noise injection methodologies.

4.4 Experimental Design Summary

Ten-fold cross validation was applied to build and test the learning models. Additionally, 10 independent repetitions of each experiment were performed to avoid any bias that may occur during the random selection process. The results reported in this work represent the average of these repetitions. A total of 5,544,000 learning models were built and evaluated from 11 learners \times 6 (total number of significant independent attributes) \times 5 (levels of attribute noise) \times 7 datasets \times 100 (10 runs of 10-fold CV) \times 24 (levels of class noise and percentages of minority instances with noise³).

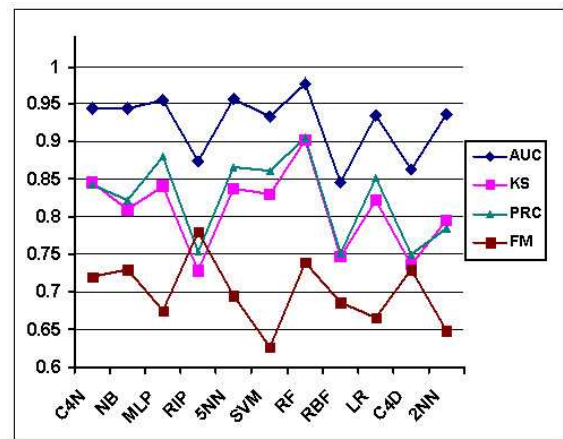


Figure 1: Overall Noise Impact on Learners

5 Learning performance

This section is organized as follows: Section 5.1 illustrates the overall impact on learning performance across all levels of noise and all datasets; Section 5.2 presents the results of the analysis of learner performance aggregated by increasing levels of injected noise; Section 5.3 presents the results of the analysis grouped by the percentage of the minority class injected with noise; Section 5.4 tabulates the impact on learner performance as the number of attributes with injected noise increases; Section 5.5 illustrates with figures the cross-effect of the increasing levels of noise and the percentage of positive instances injected with noise on learning performance; and Section 5.6 contains the analysis of

³The class noise case excluded (one out of 25 possible combinations) was for the 50% level of noise to be injected into 100% of the minority class. This case would cause all minority instances to be relabeled, leaving no minority class instances in the training data.

variance models (ANOVA) which statistically corroborated all the observations made from the analysis of the results.

5.1 Overall Classifier Performance

Figure 1 illustrates the impact of both class and attribute noise across all datasets on each learner. The average value obtained by the learners for each of the four performance metrics is provided. According to the AUC, KS, and PRC metrics, the learner with the best and most robust performance is RF, closely followed by 5NN and MLP. RIP, however, is the best performing learner as measured by FM, followed closely by RF, C4D, and NB. The worst performing learners are RIP, RBF, and C4D according to the AUC, KS, and PRC metrics. In contrast, the FM metric shows 2NN and SVM as the worst performing learners. Notice that SVM (in particular) and 2NN obtained above average performance according to the AUC, KS, and PRC metrics. In summary, the performances of eight of the learners (C4N, MLP, RIP, 5NN, SVM, LR, C4D, 2NN) according to the FM are significantly different from those obtained by the AUC, KS, and PRC metrics. These differences emphasize the importance of using an appropriate performance metric when evaluating model performance. The performance metric should most closely match the intended use of the learner during post-development deployment.

5.2 Impact of Increasing Levels of Noise on Learning

Table 3 presents the impact of increasing noise levels (10%, 20%, 30%, 40%, 50%) on learner performance across all datasets and all levels of L^m and N^a . More specifically, Table 3 considers the performance of the learners built from training datasets with L^a equal to L^c , and averaged over all datasets and all levels of L^m and N^a . 10% noise, for example, considers the datasets with both 10% attribute noise (L^a) and 10% class noise (L^c). Regardless of the metric used, the performance of all learners generally decreases as the level of noise increases. At the highest level of noise (50%), some learners exhibited a slight increase in performance. This is due to the fact that the scenario with 50% class noise and 100% of the injected noise coming from the positive class could not be implemented. In this case, there would be no instances left in the minority class. This same effect can be observed in Table 4 as well.

In Table 3, the highest value in each row is bolded. The row labeled ‘Avg’ contains the average of the five levels of noise for each learner. The average value of the best performing learner is underlined if it is significantly better, at the 95% confidence level, than the value from the second best learner (Tables 4 and 5 also have these enhancements). The AUC, KS, and PRC metrics agree that RF is the best and most robust learner. On the other hand, RIP is the most robust learner at lower levels of noise as measured by the FM metric, while NB is the best performing learner

at higher levels of noise. In addition, RIP has the highest average FM, while RF has the second best.

5.3 Impact of Minority Class Noise on Learning

The impact of the percentage of instances from the positive class injected with class noise across all datasets and all levels of noise is presented in Table 4. This table presents the impact of minority class noise on the classification performance of each learner. Once again, according to the AUC, KS, and PRC metrics, RF is the most robust learner relative to increasing levels of L^m . Furthermore, the average performances of these three metrics (AUC, KS, PRC) showed RF as the best and most robust learner.

The FM metric, on the other hand, shows RIP as the top performing learner for $L^m = 0\%$, 25%, and 50%. When $L^m = 75\%$ or 100%, the top performing learner was NB. However, the learner with the best averaged performance was RIP, followed by RF. Incidentally, none of the learners showed an increase in FM when 100% of the positive instances had noise. The results based on the FM metric continue to be quite different from the results obtained using the AUC, KS, and PRC metrics.

5.4 Impact of the Number of Significant Attributes with Noise

Table 5 illustrates the impact of increasing the number of attributes with noise on learner performance. For each learner, the performance is averaged over all datasets and all values of L^m , L^c , and L^a . Once again according to the same three metrics (AUC, KS, and PRC), RF is the best performing learner. Relative to the FM metric, RIP is the most robust learner, while RF is the second best performing learner. Clearly, the selection of a learning technique can be dramatically influenced by the choice of performance metric used for measuring the results. It is also apparent that the learning performance of all learners drop as the number of significant attributes with injected noise increased. This is particularly noticeable when N^a was 5 or 7. For the datasets used in this study, 5 attributes represent approximately 39% of all the attributes.

5.5 Impact of Noise Levels and Percentage of Noise Injected into Minority Instances

Figures 2 to 9 illustrate with line plots the impact of the factor interaction between the overall noise levels and the percentage of noisy instances injected into the positive class L^m , across all datasets. Only the best and worst performing learners are presented for each respective performance metric due to space limitations.

Figures 2 and 3 illustrate the impact of the cross-effect between the overall noise and L^m as measured by the AUC metric. RF was the best performing learner relative to this

Table 3: Impact of Increasing Noise Levels on Learning

Met	n-%	C4N	NB	MLP	RIP	5NN	SVM	RF	RBF	LR	C4D	2NN
AUC	10	0.9728	0.9702	0.9758	0.9425	0.9808	0.9891	0.9939	0.9124	0.9674	0.9320	0.9708
	20	0.9651	0.9584	0.9712	0.9279	0.9742	0.9804	0.9906	0.8782	0.9543	0.9017	0.9581
	30	0.9465	0.9458	0.9578	0.8785	0.9645	0.9505	0.9846	0.8461	0.9385	0.8644	0.9414
	40	0.9209	0.9281	0.9349	0.8095	0.9478	0.8842	0.9704	0.8106	0.9077	0.8105	0.9201
	50	0.9178	0.9209	0.9399	0.8139	0.9210	0.8627	0.9510	0.7828	0.9066	0.8093	0.8899
	Avg	0.9446	0.9447	0.9559	0.8745	0.9577	0.9334	0.9781	0.8460	0.9349	0.8636	0.9359
KS	10	0.9175	0.8638	0.8880	0.8827	0.9100	0.9314	0.9557	0.8354	0.8801	0.8733	0.8778
	20	0.8994	0.8333	0.8736	0.8474	0.8844	0.9069	0.9404	0.7863	0.8525	0.8314	0.8389
	30	0.8423	0.8083	0.8420	0.7320	0.8495	0.8483	0.9179	0.7416	0.8246	0.7411	0.7965
	40	0.7789	0.7804	0.7964	0.5902	0.8018	0.7419	0.8784	0.6928	0.7769	0.6149	0.7546
	50	0.7911	0.7670	0.8041	0.5903	0.7477	0.7231	0.8187	0.6827	0.7773	0.6210	0.7102
	Avg	0.8458	0.8106	0.8408	0.7285	0.8387	0.8303	0.9022	0.7477	0.8223	0.7363	0.7956
PRC	10	0.9155	0.8608	0.9251	0.8754	0.9407	0.9554	0.9728	0.8353	0.9085	0.8655	0.9062
	20	0.8953	0.8408	0.9123	0.8496	0.9207	0.9338	0.9584	0.7917	0.8838	0.8291	0.8618
	30	0.8444	0.8233	0.8828	0.7618	0.8892	0.8806	0.9316	0.7498	0.8545	0.7585	0.8015
	40	0.7847	0.8018	0.8382	0.6412	0.8334	0.7794	0.8779	0.7007	0.8063	0.6507	0.7259
	50	0.7725	0.7869	0.8403	0.6397	0.7489	0.7601	0.7911	0.6781	0.8037	0.6499	0.6284
	Avg	0.8425	0.8227	0.8797	0.7536	0.8665	0.8618	0.9063	0.7511	0.8513	0.7507	0.7848
FM	10	0.8838	0.7799	0.8495	0.9051	0.8669	0.8218	0.9208	0.8323	0.8280	0.8918	0.8291
	20	0.8376	0.7452	0.8018	0.8846	0.8048	0.7139	0.8564	0.7660	0.7454	0.8506	0.7375
	30	0.7120	0.7211	0.6694	0.7982	0.6939	0.6088	0.7389	0.6732	0.6503	0.7240	0.6410
	40	0.5753	0.7006	0.5225	0.6596	0.5555	0.4938	0.5891	0.5749	0.5516	0.5806	0.5313
	50	0.5956	0.7024	0.5318	0.6557	0.5603	0.4938	0.5908	0.5837	0.5514	0.6014	0.5025
	Avg	0.7209	0.7298	0.6750	0.7807	0.6963	0.6264	0.7392	0.6860	0.6654	0.7297	0.6483

Table 4: Impact of Minority Class Noise, L^m

Met	L^m	C4N	NB	MLP	RIP	5NN	SVM	RF	RBF	LR	C4D	2NN
AUC	0%	0.9658	0.9796	0.9805	0.9448	0.9703	0.9890	0.9909	0.9348	0.9801	0.9442	0.9570
	25%	0.9580	0.9603	0.9718	0.9384	0.9645	0.9844	0.9844	0.8841	0.9602	0.9114	0.9410
	50%	0.9487	0.9395	0.9614	0.9020	0.9532	0.9486	0.9756	0.8215	0.9377	0.8641	0.9246
	75%	0.9360	0.9192	0.9354	0.8021	0.9418	0.8663	0.9650	0.7735	0.9033	0.7952	0.9171
	100%	0.9136	0.9257	0.9223	0.7777	0.9679	0.8825	0.9805	0.8246	0.8900	0.8014	0.9526
	Avg	0.9444	0.9449	0.9543	0.8730	0.9595	0.9342	0.9793	0.8477	0.9342	0.8632	0.9384
KS	0%	0.9054	0.8988	0.8970	0.8859	0.8742	0.9341	0.9358	0.8794	0.9113	0.8888	0.8527
	25%	0.8783	0.8366	0.8723	0.8693	0.8545	0.9212	0.9141	0.8059	0.8630	0.8444	0.8020
	50%	0.8607	0.7909	0.8474	0.7736	0.8207	0.8497	0.8888	0.7221	0.8203	0.7639	0.7723
	75%	0.8214	0.7558	0.7996	0.5731	0.8040	0.7230	0.8648	0.6568	0.7671	0.6037	0.7581
	100%	0.7565	0.7716	0.7746	0.5281	0.8631	0.7237	0.9299	0.6724	0.7428	0.5709	0.8136
	Avg	0.8445	0.8107	0.8382	0.7260	0.8433	0.8303	0.9067	0.7473	0.8209	0.7343	0.7997
PRC	0%	0.9065	0.8790	0.9325	0.8745	0.9164	0.9552	0.9613	0.8575	0.9388	0.8606	0.8547
	25%	0.8757	0.8417	0.9112	0.8618	0.8931	0.9440	0.9277	0.8017	0.8952	0.8348	0.7990
	50%	0.8395	0.8103	0.8874	0.7961	0.8507	0.8816	0.8851	0.7309	0.8494	0.7795	0.7452
	75%	0.8115	0.7848	0.8414	0.6312	0.8073	0.7636	0.8501	0.6722	0.7936	0.6492	0.7248
	100%	0.7808	0.8005	0.8180	0.5952	0.8944	0.7660	0.9367	0.6970	0.7736	0.6246	0.8430
	Avg	0.8428	0.8233	0.8781	0.7518	0.8724	0.8621	0.9122	0.7519	0.8501	0.7497	0.7933
FM	0%	0.8550	0.8223	0.8640	0.9053	0.8501	0.8881	0.8810	0.8552	0.8766	0.8952	0.7337
	25%	0.8354	0.7583	0.8229	0.8925	0.8344	0.7720	0.8746	0.8050	0.7954	0.8677	0.7062
	50%	0.7740	0.7133	0.7041	0.8337	0.7322	0.5876	0.7816	0.6960	0.6490	0.7800	0.6546
	75%	0.5940	0.6793	0.6185	0.6499	0.5574	0.4474	0.6051	0.5504	0.5067	0.5757	0.5871
	100%	0.5336	0.6693	0.4605	0.6136	0.4939	0.4229	0.5443	0.5083	0.4861	0.5120	0.5743
	Avg	0.7184	0.7285	0.6940	0.7790	0.6936	0.6236	0.7373	0.6830	0.6627	0.7261	0.6512

Table 5: Noise Impact by Number of Significant Attributes, N^a

Met	N^a	C4N	NB	MLP	RIP	5NN	SVM	RF	RBF	LR	C4D	2NN
AUC	1	0.9558	0.9672	0.9685	0.9029	0.9727	0.9408	0.9840	0.8543	0.9334	0.8926	0.9594
	2	0.9547	0.9627	0.9674	0.8949	0.9694	0.9419	0.9844	0.8543	0.9410	0.8867	0.9541
	3	0.9497	0.9563	0.9649	0.8824	0.9635	0.9426	0.9816	0.8543	0.9448	0.8717	0.9433
	4	0.9493	0.9493	0.9635	0.8795	0.9593	0.9420	0.9823	0.8528	0.9431	0.8707	0.9358
	5	0.9375	0.9380	0.9592	0.8682	0.9533	0.9375	0.9748	0.8478	0.9418	0.8515	0.9264
	7	0.9273	0.9006	0.9160	0.8341	0.9370	0.9107	0.9682	0.8286	0.9124	0.8218	0.9082
	Avg	0.9457	0.9456	0.9566	0.8770	0.9592	0.9359	0.9792	0.8487	0.9361	0.8658	0.9379
KS	1	0.8771	0.8700	0.8745	0.7863	0.8725	0.8576	0.9251	0.7697	0.8279	0.7887	0.8493
	2	0.8721	0.8556	0.8704	0.7680	0.8622	0.8556	0.9254	0.7676	0.8393	0.7784	0.8303
	3	0.8557	0.8340	0.8614	0.7426	0.8523	0.8503	0.9140	0.7612	0.8453	0.7483	0.8046
	4	0.8549	0.8150	0.8558	0.7383	0.8448	0.8456	0.9133	0.7570	0.8401	0.7469	0.7931
	5	0.8303	0.7898	0.8447	0.7192	0.8326	0.8270	0.8917	0.7441	0.8278	0.7174	0.7765
	7	0.7987	0.7097	0.7473	0.6511	0.7904	0.7673	0.8648	0.7031	0.7644	0.6672	0.7413
	Avg	0.8481	0.8124	0.8423	0.7343	0.8425	0.8339	0.9057	0.7504	0.8242	0.7412	0.7992
PRC	1	0.8730	0.8958	0.9105	0.8103	0.9007	0.8870	0.9266	0.7824	0.8619	0.8023	0.8342
	2	0.8725	0.8787	0.9067	0.7953	0.8929	0.8845	0.9290	0.7746	0.8710	0.7944	0.8224
	3	0.8537	0.8547	0.8982	0.7654	0.8831	0.8806	0.9186	0.7654	0.8751	0.7614	0.8008
	4	0.8527	0.8228	0.8946	0.7615	0.8749	0.8771	0.9195	0.7583	0.8689	0.7598	0.7908
	5	0.8359	0.7926	0.8855	0.7476	0.8641	0.8609	0.9045	0.7409	0.8581	0.7385	0.7731
	7	0.7844	0.7006	0.7928	0.6697	0.8130	0.8018	0.8686	0.7033	0.7850	0.6732	0.7264
	Avg	0.8454	0.8242	0.8814	0.7583	0.8715	0.8653	0.9111	0.7542	0.8533	0.7549	0.7913
FM	1	0.7578	0.7855	0.7055	0.8287	0.7289	0.6478	0.7664	0.7156	0.6891	0.7743	0.6675
	2	0.7538	0.7730	0.7032	0.8158	0.7241	0.6466	0.7651	0.7080	0.6912	0.7684	0.6669
	3	0.7316	0.7555	0.6962	0.7907	0.7147	0.6424	0.7526	0.6997	0.6880	0.7402	0.6630
	4	0.7290	0.7320	0.6920	0.7874	0.7057	0.6401	0.7514	0.6926	0.6857	0.7375	0.6608
	5	0.7173	0.7063	0.6847	0.7782	0.6986	0.6311	0.7398	0.6785	0.6706	0.7238	0.6480
	7	0.6671	0.6334	0.6041	0.7145	0.6396	0.5735	0.6968	0.6473	0.5959	0.6661	0.6201
	Avg	0.7261	0.7310	0.6810	0.7859	0.7019	0.6302	0.7454	0.6903	0.6701	0.7351	0.6544

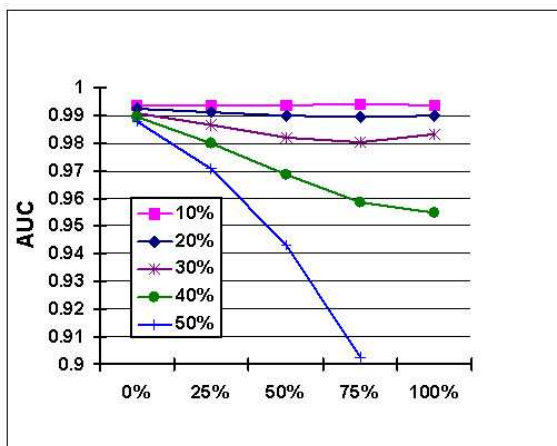


Figure 2: Cross-effect of Noise Level and L^m on RF by AUC

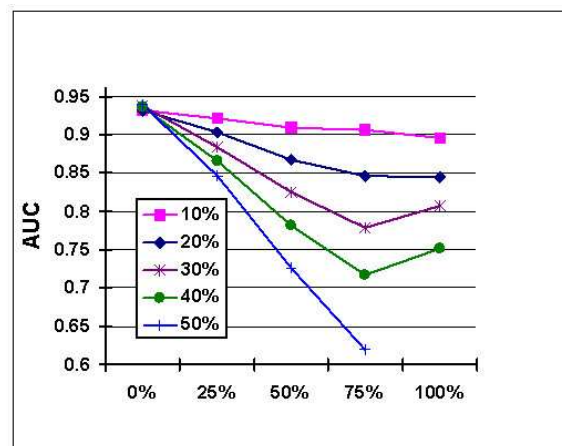


Figure 3: Cross-effect of Noise Level and L^m on RBF by AUC

noise combination. The impact on RF was only at the highest level of noise (50%) and when L^m was more than 50%. On the other hand, RBF was the learner most affected by this factor interaction. Even at the lowest level of noise of 10% and with $L^m = 50\%$ or more, the performance of RBF declined substantially.

Figures 4 and 5 illustrate the impact of the factor interaction between the overall noise and L^m as measured by the KS metric. For the RF learner, which was the best perform-

ing and most robust learner relative to the KS, significant deterioration was only seen at the 40% level of noise or higher, and when L^m was more than 50%. In contrast, RIP was the learner most affected by this interaction. The performance of RIP was affected little at the lowest level of noise of 10%. However, for all other noise levels and when L^m was 25% or more, the performance of RIP declined dramatically.

Figures 6 and 7 illustrate the impact of the factor interac-

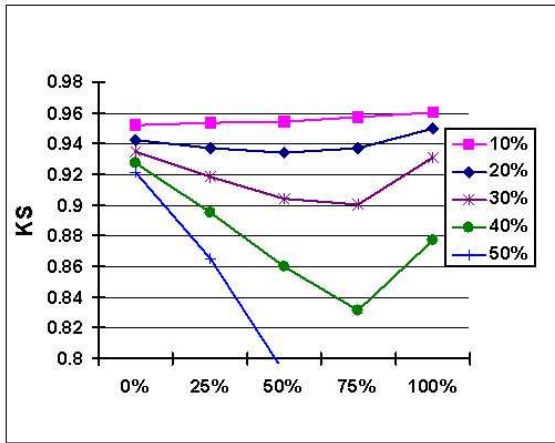


Figure 4: Cross-effect of Noise Level and L^m on RF by KS

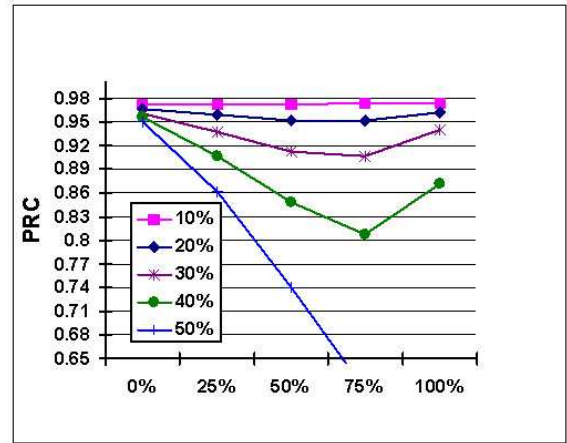


Figure 6: Cross-effect of Noise Level and L^m on RF by PRC

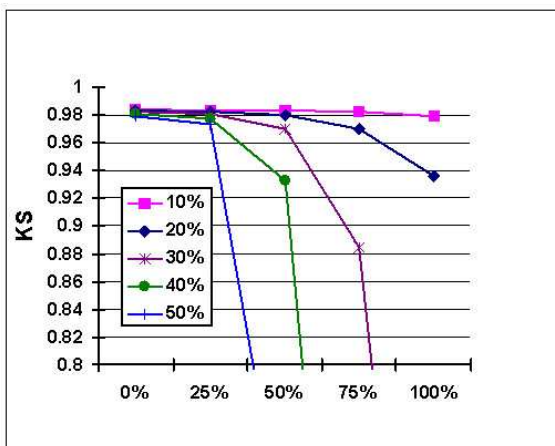


Figure 5: Cross-effect of Noise Level and L^m on RIP by KS

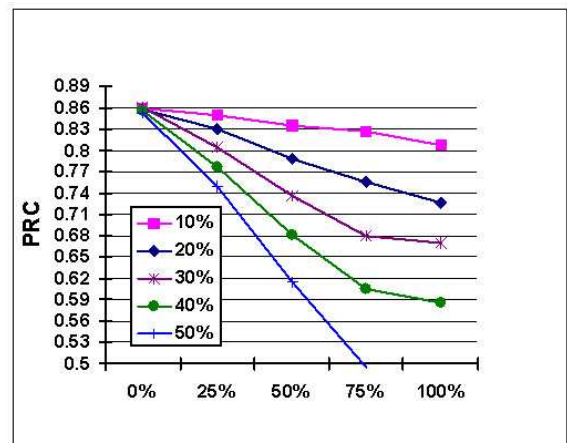


Figure 7: Cross-effect of Noise Level and L^m on RBF by PRC

tion between the overall noise and L^m as measured by the PRC metric. The best performing and most robust learner to this type of noise was RF. On the other hand, RBF was the learner most affected by this factor interaction. Even at the lowest level of noise of 10% and with $L^m = 50\%$ or more, RBF’s performance declined very significantly.

Figures 8 and 9 illustrate the impact of the cross-effect between the overall noise and L^m as measured by the FM metric. The best performing and most robust learner was RIP. Nevertheless, the performance of RIP was noticeably affected when the noise level was 30% or more and L^m was 50% or more. On the other hand, SVM was the learner most affected by this factor interaction. Even at the lowest level of noise and when L^m was 25% or more, SVM’s performance declined very significantly.

These figures clearly illustrate the very significant deterioration of learning performance when both the levels of noise and the percent of positive instances with noise increased. In real-life scenarios, this observation would imply that noise, and, in particular, the amount of noise in the

minority class specifically, are critically important factors determining the ultimate reliability and value of any classification undertaking.

5.6 Analysis of Variance (ANOVA)

The results presented in this study are also tested and validated for statistical significance at the $\alpha = 5\%$ level using five factor analysis of variance [2] models. An ANOVA model can be used to test the hypothesis that the classification performances of each level of the main factors are equal against the alternative hypothesis that at least one is different. Note that in this study, only the main factor representing the classification techniques (learners) and the corresponding cross-effects with the other factors were investigated in detail. The five factor models [2] used in this work can be represented as follows:

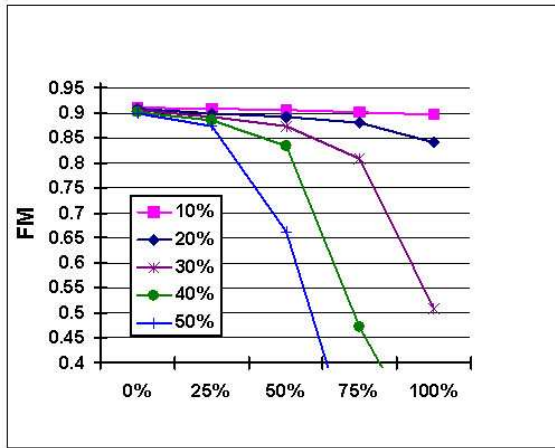


Figure 8: Cross-effect of Noise Level and L^m on RIP by FM

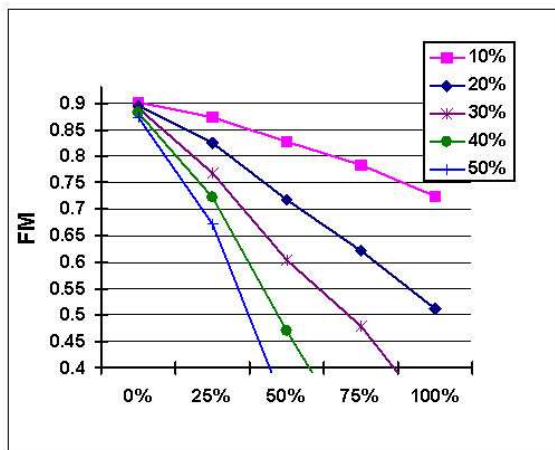


Figure 9: Cross-effect of Noise Level and L^m on SVM by FM

$$\psi_{(jklmin)} = \mu + C_j + L_k^c + L_l^m + L_m^a + N_i^a + \varepsilon_{(jklmin)}$$

where the parameters of the model are defined as:

- $\psi_{(jklmin)}$ is the response variable (AUC, KS, PRC, or FM) for the n^{th} observation of the j^{th} level of C , k^{th} level of L^c , l^{th} level of L^m , m^{th} level of L^a , and i^{th} level of N^a .
- μ is the overall mean effect on the observations of each response variable.
- C_j (resp., L_k^c , L_l^m , L_m^a , N_i^a) is the mean performance of level j (resp., k , l , m , i) for factor C (resp., L^c , L^m , L^a , N^a).
- $\varepsilon_{(jklmin)}$ is the random error.

The main factor (the learner) is tested to see if the average performance of the 11 levels (classifiers/learners) of C and the corresponding interactions with the five levels of L^c , the five levels of L^m , the five levels of L^a , and the six groups of N^a are equal, respectively. Only the most significant two or three-way interactions were reported in this article. An example of a two-way factor interaction can be $C \times L_{(jl)}^m$, which represents the interaction of the learners and the L^m factor. To determine which response variable (AUC, KS, PRC, or FM) values are significantly different, a pair-wise comparison of each response variable with the null hypothesis that they are equal (i.e., not significantly different) can be used. In this study, we apply the Tukey’s Honestly Significant Difference (HSD) test to identify which levels of the main factor are significantly different [2].

The ANOVA models were built using the AUC, KS, PRC, and FM metrics as the respective response variables. The Fisher’s distribution values (F -values) of the factor and the most significant multi-way interactions were tabulated with their corresponding p -values in Table 6. This table shows the significance of the experimental factor and its interactions by the respective p -values. The main factor and the two or three-way interaction terms are statistically significant at $\alpha = 5\%$ with p -values $\ll 0.001$, which are denoted < 0.0001 . We report the most significant two or three-way factor interactions.

In Table 6, the learners (or C factor) had the highest F -values as measured by all four metrics. This agrees with our previous observation made in the analysis that the choice of learner, particularly in the presence of low quality class imbalanced data, can be critical to classification performance. The most significant multi-way factor interaction was statistically determined to be the learners and the number of instances injected with noise from the minority class, $C \times L^m$. This observation indicates that on average, the performance of the classification techniques investigated in this study were most affected when the positive instances contained noise. This is true regardless of the performance metric used in this work. The second most significant factor interaction was with the levels of class noise, $C \times L^c$. This indicates that on average the level of class noise can also significantly impact the performance of the learning techniques. Furthermore, the three-way factor interaction of the learners, the percentage of positive instances with noise, and the level of class noise ($C \times L^c \times L^m$) also shows a significant impact on learning performance. Once again, all four performance metrics concurred on this result.

On the other hand, and according to all four metrics, the interaction with the level of attribute noise ($C \times L^a$) was relatively less significant compared to the other two-way cross-effect presented in Table 6. This also corroborates previous observations made elsewhere by our group that attribute noise is not nearly as significant as class noise to classification performance.

Table 7 provides the mean values of each metric as well

Table 6: ANOVA Models

Factor	DoF	AUC		KS		PRC		FM	
		F-val	p-val	F-val	p-val	F-val	p-val	F-val	p-val
\mathcal{C}	10	33528.0	<0.0001	18082.2	<0.0001	16223.8	<0.0001	13660.7	<0.0001
$\mathcal{C} \times N^a$	50	201.6	<0.0001	198.8	<0.0001	204.6	<0.0001	142.4	<0.0001
$\mathcal{C} \times L^a$	40	39.4	<0.0001	57.6	<0.0001	108.1	<0.0001	76.3	<0.0001
$\mathcal{C} \times L^c$	40	1191.3	<0.0001	1438.3	<0.0001	1027.2	<0.0001	1582.9	<0.0001
$\mathcal{C} \times L^m$	40	2967.4	<0.0001	3830.0	<0.0001	1838.0	<0.0001	3300.6	<0.0001
$\mathcal{C} \times N^a \times L^a$	200	9.3	<0.0001	8.2	<0.0001	9.0	<0.0001	8.2	<0.0001
$\mathcal{C} \times N^a \times L^c$	200	5.1	<0.0001	5.7	<0.0001	4.7	<0.0001	2.6	<0.0001
$\mathcal{C} \times L^m \times N^a$	200	10.5	<0.0001	16.4	<0.0001	11.5	<0.0001	4.1	<0.0001
$\mathcal{C} \times L^c \times L^a$	160	1.1	0.3	2.3	<0.0001	1.2	0.0420	0.6	1.0
$\mathcal{C} \times L^m \times L^a$	160	2.4	<0.0001	4.3	<0.0001	1.4	0.0003	1.8	<0.0001
$\mathcal{C} \times L^c \times L^m$	160	587.2	<0.0001	831.2	<0.0001	440.3	<0.0001	837.1	<0.0001

Table 7: ANOVA Factor: \mathcal{C}

\mathcal{C}	AUC		\mathcal{C}	KS		\mathcal{C}	PRC		\mathcal{C}	FM	
	Mean	HSD		Mean	HSD		Mean	HSD		Mean	HSD
RF	0.9792	A	RF	0.9057	A	RF	0.9111	A	RIP	0.7859	A
5NN	0.9591	B	C4N	0.8481	B	MLP	0.8814	B	RF	0.7454	B
MLP	0.9566	C	5NN	0.8425	C	5NN	0.8715	C	C4D	0.7351	C
C4N	0.9457	D	MLP	0.8423	C	SVM	0.8653	D	NB	0.7310	D
NB	0.9456	D	SVM	0.8339	D	LR	0.8533	E	C4N	0.7261	E
2NN	0.9379	E	LR	0.8242	E	C4N	0.8454	F	5NN	0.7019	F
LR	0.9361	F	NB	0.8124	F	NB	0.8242	G	RBF	0.6903	G
SVM	0.9359	F	2NN	0.7992	G	2NN	0.7913	H	MLP	0.6810	H
RIP	0.8770	G	RBF	0.7504	H	RIP	0.7583	I	LR	0.6701	I
C4D	0.8658	H	C4D	0.7411	I	C4D	0.7549	J	2NN	0.6544	J
RBF	0.8487	I	RIP	0.7343	J	RBF	0.7542	J	SVM	0.6302	K

as the significant HSD grouping levels for each learner. Note that if two or more instances of the factor have the same block letter, then their performances are not significantly different. Table 7 shows the overall impact of noise across all levels of noise and all datasets on the performance of each learner \mathcal{C} . According to the AUC, KS, and PRC metrics, RF performs significantly better than all of the other learners. The second best learner varies by metric (5NN for AUC, C4N for KS, and MLP for PRC). In general, RF, 5NN, and MLP perform very well as measured by AUC, KS, and PRC, while C4D, RIP, and RBF are the most affected by noise. The FM metric, however, shows RIP as the most robust learner in group ‘A’, followed by RF in group ‘B’. The two learners most impacted by noise were 2NN and SVM, according to the FM metric. Interestingly, two of the worst performing learners (C4D and RIP) according to the AUC, KS, and PRC are among the top three performing techniques according to FM.

The best performing and most robust classifier regardless of the quality of the data was RF, according to the AUC, KS, and PRC metrics. According to the FM metric, RF was the second best performing learner and was uniquely placed in group ‘B’, while RIP was the best performing learner. It is noteworthy to emphasize the unmatched robustness of RF in the presence of any type and level of noise injected in these experiments. To summarize, the ANOVA analysis presented in this section has corroborated

our previous observations and conclusions regarding the robustness of the RF learner in the presence of low quality and class imbalanced data. Further, the percent of instances injected with noise from the minority class and the level of class noise present in the data, respectively, had the most profound effect on learning. The results from the FM metric are substantially different from the other three metrics (AUC, KS, PRC) and thus reiterate the importance of determining the appropriate metric for measuring learner performance.

6 Conclusion

The objective of this study was to investigate the robustness of a variety of common-used learning algorithms relative to low quality, class imbalanced measurement data. Real-world software measurement data typically contains an imbalanced class distribution, and if erroneous attribute values are also present, the impact on learning would be more significant. Our classification study using 11 different learning techniques and low quality, class imbalanced data can be most helpful to practitioners in many application domains. In order to conduct this investigation, a comprehensive suite of experiments was designed and implemented with the use of seven real world measurement datasets, initially relatively free of noise. A novel noise injection procedure was designed and applied using sev-

eral domain realistic noise parameters. The results were measured using four distinct performance metrics appropriate for imbalanced data. The level of comprehensiveness achieved in this study can be easily seen by the mere fact that over 5.5 million classification models were built and evaluated during our experimentation.

In general, the results unequivocally demonstrated that the quality of the measurement data (both attribute and class noise) can impact classification performance significantly. All conclusions and observations made in the analysis of this study were statistically verified by constructing analysis of variance (ANOVA) models. The learning technique with the best and most consistent performance in all experiments was the random forest ensemble classification technique. Three (AUC, KS, PRC) of the four performance metrics used in this study concurred on the determination of RF as the best and most robust learner. According to the FM metric, RF was second only to the RIP rule-based learner. On the other hand, those learners most impacted by low quality class imbalanced data were C4D, RIP, and RBF, according to the AUC, KS, and PRC metrics. Curiously, C4D and RIP learners were among the top three best performing techniques as measured by FM. The worst performing learners according to the FM metric were SVM and 2NN. These results are also unusual because both learners, SVM (in particular) and 2NN, are known to generate acceptable classification performances even in the presence of low quality class imbalanced data, as reported in this study with the results from the AUC, KS, and PRC metrics.

Regardless of the quality of the measurement data and whether or not the data has significant class imbalance, we are very confident in recommending the random forest ensemble classifier for learning initiatives. To our knowledge, no other related classification study has identified a learning technique significantly robust with consistently excellent performance in the presence of low quality, class imbalanced measurement data using four distinct performance metrics. Future work will consider performance enhancing techniques such as cost-sensitive learning and boosting and will also include additional datasets and learning techniques. Additional noise injection methodologies can also be considered in future work.

References

- [1] D. W. Aha. *Lazy learning*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.
- [2] M. L. Berenson, D. M. Levine, and M. Goldstein. *Intermediate Statistical Methods and Applications: A Computer Package Approach*. Prentice-Hall, Inc., 1983.
- [3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [4] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC Press, Boca Raton, FL, 1984.
- [5] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.
- [6] W. W. Cohen. Fast effective rule induction. In *Proc. 12th International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.
- [7] W. Conover. *Practical Nonparametric Statistics*. John Wiley and Sons, NY, 1971.
- [8] J. Davis and K. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.
- [9] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 239–246, 2001.
- [10] Y. Feng, Z. Wu, and Z. Zhou. Enhancing reliability throughout knowledge discovery process. In *6th IEEE International Conference on Data Mining - Reliability Issues in Knowledge Discovery Workshop (RIKD06)*, pages 754–758, 2006.
- [11] A. Folleco, T. M. Khoshgoftaar, J. Van Hulse, and L. Bullard. Identifying learners robust to low quality data. In *Proceedings of the IEEE International Conference on Information Reuse and Integration-IRI'08*, pages 190–195, July 2008.
- [12] E. Frank, L. Trigg, G. Holmes, and I. Witten. Naive bayes for regression. *Machine Learning*, pages 1–20, 2000.
- [13] D. J. Hand. Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, 56:1109–1117, 2005.
- [14] D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley Sons, Inc, 2nd edition, 2000.
- [15] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse. An empirical study of learning from imbalanced data using random forest. In *Proceedings of 19th IEEE International Conference on Tools with Artificial Intelligence*, pages 310–317, Patras, Greece, October 2007.
- [16] T. M. Khoshgoftaar, V. Joshi, and N. Seliya. Detecting noisy instances with the ensemble filter: A study in software quality estimation. *International Journal of Software Engineering and Knowledge Engineering*, 16(1):1–24, 2006.
- [17] T. M. Khoshgoftaar and N. Seliya. The necessity of assuring quality in software measurement data. In *Proceedings of 10th International Software Metrics*

- Symposium*, pages 119–130, Chicago, IL, September 2004. IEEE Computer Society.
- [18] T. M. Khoshgoftaar, N. Seliya, and K. Gao. Detecting noisy instances with the rule-based classification model. *Intelligent Data Analysis: An International Journal*, 9(4):347–364, 2005.
- [19] W. Lee, S. J. Stolfo, and K. W. Mok. A data mining framework for building intrusion detection models. *Proceedings of IEEE Symposium on Security and Privacy*, pages 120–132, May 1999.
- [20] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, Hoboken, NJ, 2nd edition, 2002.
- [21] J. Moody and C. J. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1(2):281–294, 1989.
- [22] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42:203–231, 2001.
- [23] J. R. Quinlan. *C4.5: Programs For Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
- [24] B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, Massachusetts, 1999.
- [25] J. Van Hulse. Data quality in data mining and machine learning. *Ph.D. Dissertation, Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL USA*, May 2007. Advised by T. M. Khoshgoftaar.
- [26] J. Van Hulse and T. M. Khoshgoftaar. Class noise detection using frequent itemsets. *Intelligent Data Analysis: An International Journal*, 10(6):487–507, 2006.
- [27] J. Van Hulse, T. M. Khoshgoftaar, and H. Huang. The pairwise attribute noise detection algorithm. *Knowledge and Information Systems Journal, Special Issue on Mining Low Quality Data*, 11(2):171–190, 2007.
- [28] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Skewed class distributions and mislabeled examples. In *Seventh IEEE International Conference on Data Mining - Workshops (ICDMW'07)*, pages 477–482, October 2007.
- [29] G. Weiss. Learning with rare cases and small disjuncts. In *Proceedings of the 12th International Conference on Machine Learning*, pages 558–565. Morgan-Kaufmann, 1995.
- [30] G. M. Weiss and F. Provost. Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.
- [31] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, California, 2nd edition, 2005.
- [32] X. Zhu and X. Wu. Class noise vs attribute noise: A quantitative study of their impacts. *Artificial Intelligence Review*, 22(3-4):177–210, November 2004.
- [33] X. Zhu and X. Wu. Cost-guided class noise handling for effective cost-sensitive learning. In *4th IEEE International Conference on Data Mining (ICDM 2004)*, pages 297–304, November 2004.
- [34] X. Zhu, X. Wu, T. M. Khoshgoftaar, and Y. Shi. An empirical study of the noise impact on cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, pages 1168–1174, 2007.
- [35] L. Zhuang and H. Dai. Reducing performance bias for unbalanced text mining. In *6th IEEE International Conference on Data Mining - Reliability Issues in Knowledge Discovery Workshop (RIKD06)*, pages 770–774, 2006.

A General Purpose Software Evaluation System

Behrouz H. Far and Vani Mudigonda

Schulich School of Engineering, University of Calgary, Calgary, Alberta, Canada

E-mail: far@ucalgary.ca, vmudigon@ucalgary.ca

Abdel-Halim Elamy

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

E-mail: elamy@ualberta.ca

Keywords: SW evaluation, statistics.

Received: October 12, 2008

In the present day market situation, there are several alternatives available when a customer wants to purchase a product or adopt a software system that meets the customer's requirements. General Purpose Software Evaluation (GPSE) system uses state of the art statistical methods based on Multidimensional Weighted Attribute Framework (MWAFF) for the evaluation of the available alternatives. By using GPSE system, the user can follow the MWAFF process and design the architecture which best describes the given evaluation problem. The architectural elements of MWAFF essentially focus on survey questionnaire which involves gathering information from several domain experts. The GPSE system then applies principles of Analysis of Variance (ANOVA) and Tukey's pairwise comparison tests on the data collected to arrive at selection of the best suited alternative for the given problem. The GPSE system has been fully implemented and successfully tested on several projects including evaluation of multi-agent development methodologies and selection of COTS products.

Povzetek: Predstavljen je splošni sistem GPSE za vrednotenje programskih sistemov.

1 Introduction

Software technologies have been evolving rapidly and for a given set of functional and non-functional requirements there usually exist several competing software products. The present day users are faced with a challenging situation that requires evaluation and selection of a suitable software product that satisfies the users' operational and business needs. Unfortunately this evaluation is usually carried on in an ad-hoc and informal way and with various degree of success. The objective of this research is to develop a General Purpose Software Evaluation (GPSE) system that helps a user systematically evaluate a set of alternative products available for a given set of requirements by employing sound statistical methods. The GPSE system incorporates and implements the Multidimensional Weighted Attribute Framework (MWAFF) [10, 11]. MWAFF is a framework for creating the evaluation criteria and collecting data from subject matter experts in the form of rates and weights for each alternative included in the evaluation. The data collected is then processed and subjected to statistical analysis by using Analysis of Variance (ANOVA) and Tukey's pairwise comparison tests. The MWAFF possesses great potential in its applicability to a variety of applications. The present work focused on implementing the MWAFF in the form of the GPSE system.

The paper is organized as follows. Section 2 delivers a quick review of several software evaluation techniques.

Section 3 presents overview of the MWAFF framework and Section 4 describes the GPSE system design and implementation details. Finally, Section 5 presents the conclusions and scope for future extension of the project.

2 Related works

Software evaluation methodologies can be divided into two categories. The first category is used to evaluate software development methodologies or processes such as those used to evaluate various agent-based development methodologies. The second category is used to evaluate software products such as COTS evaluation and selection methodologies.

In the literature, there are a few studies addressing the comparison and evaluation of processes and methodologies. Available techniques merely focus on a single application domain making generalization of the method almost impossible. For instance in the domain of evaluating agent-based development methodologies, Dam et al [9] proposed an attribute-based framework for evaluation by analyzing feedback data from both the system developers as well as from end users. Juneidi and Vouros [14] utilized the evaluation criteria of Shehory and Sturm [21] and conducted a study to evaluate three agent-based development methodologies. Further, Bayer and Svantesson [2] introduced a study to compare and evaluate two agent-based methodologies by identifying their strengths and weaknesses. Another work has been presented by Sudeikat et al. [23] to evaluate three agent-based methodologies (MaSE, Tropos, and Prometheus)

against a number of evaluation criteria (e.g., internal architecture, social architecture, communication, and process-related features) that have been examined and compared qualitatively.

Tran et al. present a comparative Feature Analysis Framework [24] that includes 4 criteria: process, technical, model and support, and is tailored to evaluating agent-based methodologies. The Framework can be recommended to adopt as an analytical tool to exhibit various detailed features involving agents and multi-agent systems. Yet, it is not a purely evaluation framework.

Silva et al. proposed a Non-Functional Requirements (NFR) framework to describe the internal properties of systems and to evaluate the agent-based methodologies based on these properties [20]. As a matter of fact, non-functional requirements (NFRs) have significant impact on the process of software development [7]. When designing a system, such NFRs represent trade-offs in the design basic principles that contribute to deciding upon specific structural/behavioral aspects of the system [13]. Similar to Tran et al.'s framework, Silva et al.'s is lacking the empirical/analytical approach to quantify the subjective features of the NFRs which are qualitative in their nature and consequently, cannot be easily and accurately examined and compared.

Regarding product evaluation methodologies, there are relatively larger number of methods such as those used to evaluate and select commercial off-the-shelf (COTS) products [1, 3, 4, 7, 15, 16, 18].

To conclude, all of the above mentioned techniques show one or more of the following methodological deficiencies:

- a) unrepresentative set of responses;
- b) heterogeneous experimental subjects;
- c) using different instruments for obtaining similar responses; and
- d) mixing up the scales of measurement.

Our GPSE system is generic enough to be used for evaluation of both software products and processes. The Multidimensional Weighted Attribute Framework (MWAF), which is used in the GPSE system, follows sound statistical guidelines to design experiments and interpret data and consequently does not suffer from the above mentioned deficiencies.

3 Multidimensional Weighted Attribute Framework (MWAF)

In this section we present the Multidimensional Weighted-Attributes Framework (MWAF) for software system evaluation. MWAF is a general-purpose framework that can be adapted to evaluate software products, e.g., programming languages, operating systems, software engineering methodologies, software development toolkits and software communications protocols.

3.1 MWAF Framework

The main idea of MWAF is to define the most common and important criteria (or dimensions) of the system being evaluated, identifying the attributes that describe each of these dimensions, and then evaluating each dimension through its attributes against all the potential systems that are selected for evaluation. As shown in Figure 1, MWAF consists of the following three main components:

- 1) Dimensions: the framework comprises a number of dimensions, each of which represents one of the major evaluation criteria.
- 2) Attributes: are the different features pertaining to each criterion (i.e. dimension) to describe it using a set of definite questions.
- 3) Parameters: the numeric values that are given to measure the attributes.

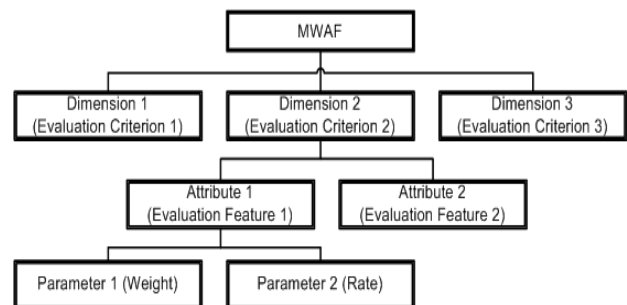


Figure 1: Hierarchy of the Multidimensional Weighted-Attributes Framework (MWAF).

For example, the following four attributes can be used for evaluating the ‘objectivity’ of public websites: a) Goal-orientation; b) Comprehensiveness; c) Fair-mindedness; and d) Independency. To perform this evaluation, we can assume ‘objectivity’ as an evaluation dimension that encompasses the above 4 attributes. Each of these attributes can be evaluated through relevant expressive questions, such as:

- a. Goal-orientation: to what degree does the website meet its announced goals?
- b. Comprehensiveness: how detailed is the information posted on the website?
- c. Fair-mindedness: to what extent would you agree with the opinions expressed by the authors of the website?
- d. Independency: to what degree would you reject the popped up advertising on this web page?

When applying MWAF, several expert users will be asked to give two parameters to each of the evaluated attributes: a *weight* to identify the importance of the attribute, and a *rate* to measure its strength or effectiveness. Weight is a subjective parameter, as it entirely relies upon the evaluator’s personal opinion. On the other hand, rate is an objective parameter because it is measured according to the degree of availability or effectiveness of the examined property as represented by the evaluated attribute. In MWAF, the values given to the two parameters are numeric and range from 0 to 10. A value of ‘0’ implies full absence of the measured

attribute, whereas a value of 10 reflects its maximum availability and strength. For instance, the ‘Comprehensiveness’ attribute may receive weights and rates by four participants as shown in Table 1.

Table 1: Sample expert user input.

Expert User	1	2	3	4
Weight	5	9	6	7
Rate	10	7	8	9

In this example, the first expert user assumes that the ‘Comprehensiveness’ is moderately important to evaluate a public website. However, in his/her view, the evaluated website is extremely comprehensive. Based on the collected data, we can determine the weighted rates by normalizing each raw rate against the average weight given to this attribute.

$$\text{Average Weight} = \frac{\sum w_i}{n} = \frac{5+9+6+7}{4} = 6.75$$

And the weighted rates are depicted in Table 2.

Table 2: Calculated weighted rates.

Expert User	1	2	3	4
Weighted Rate	6.	4.7	5.	6.0
Rate	750	25	400	75

The rest of the evaluating procedure will be carried out upon analyzing and comparing these rates, as weighted against the average importance of the evaluated attribute.

In order to take a broad view of the final conclusions and findings, each system shall be evaluated by several expert users. The number of expert users will be identified during the experiment design (See Section 3.3 step 5).

3.2 MWAFF Data Abstraction

The data abstraction process formulates blocks and replicas based on the identified dimensions and attributes (see Figure 2). A block consists of a set of treatments assigned to an expert user for evaluation. Each treatment is included in multiple blocks and hence evaluated by multiple expert users leading to multiple replicas of data. Identification of blocks and replica is part of the experiment design. For example, in the Balanced Incomplete Block Design (BIBD) model each pair of treatment must occur the same number of times as any other pair. The common choices are Completely Randomized Design (CRD), Randomized Complete Block Design (RCBD), and Balanced Incomplete Block Design (BIBD) model.

The CRD is the simplest type of randomization schemes in that subjects are assigned to treatments completely at random such that every experimental unit has an equal chance to receive any of the available treatments [19]. Various randomization techniques could be used for assigning subjects to treatment groups; the common method is to label subjects or treatments and then use a table of random numbers to select subjects at random and assign them to treatments.

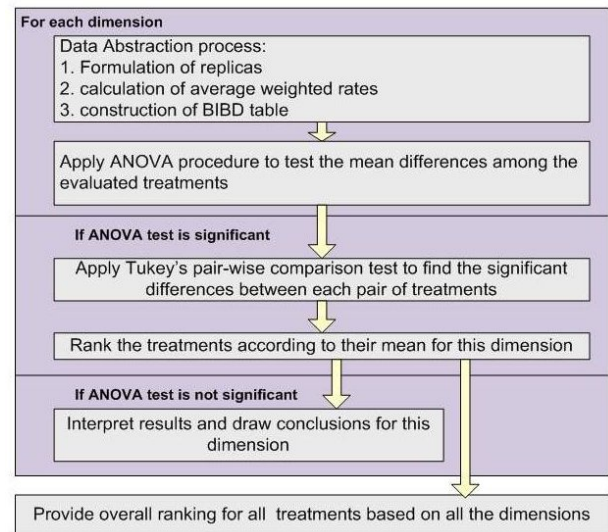


Figure 2: MWAFF process flow diagram.

Cochran and Cox [6] indicate that due to the unrestricted randomization, units that receive one treatment may be naturally different from units that receive other treatments. This heterogeneity among experimental units contributes to producing a larger experimental error as compared to other designs. However, for the same number of observations a completely randomized design has the largest degrees of freedom for error. Although the sum of squares error may be enlarged by the natural variability in units, dividing this sum of squares by larger degrees of freedom may result smaller mean square error. Given the above advantages and disadvantages, CRD is appropriate when experimental units are quite homogenous, the experiment under study is relatively small, and when there is a chance to lose some experimental units and having a missing data problem.

In RCBD, randomization is restricted and controlled such that the experimental units are arranged into homogeneous groups (called blocks), and the treatments are then assigned at random to these blocks so that each treatment occurs once in every block, or as planned if the block sizes are not the same. The rationale behind blocking is to minimize the variability among units within blocks while maximizing it among blocks. Neter argues that RCBD can potentially have disadvantages such as: more assumptions (e.g., no interactions between treatments and blocks, and constant variability among blocks) are needed to be met; missing observations are complex to handle; and precision decreases as the number of experimental units in a block increases [19].

Clarke and Kempson [5] indicate that experiments often use supplies or resources that are not homogenous, but can be arranged into blocks of similar units so that most of the heterogeneity is taken out between blocks. An incomplete block design is called “balanced” or “symmetrical” if treatment levels are binary [17]. That is, when an incomplete design is formed so that every pair of treatments occurs together the same number of times as any other pair, the design is a Balanced Incomplete Block Design (BIBD). In BIBD, all

treatment comparisons are of the same accuracy, thus, we use these designs when all treatments are equally important. Yates [25] argues that the main drawback of a BIBD is that the number of replications required is in most cases large when the number of treatments is at all large. However, we can overcome this drawback by administering with the condition of balance, but at the cost of some loss of efficiency in addition to the inconvenience of having slight variation in accuracy for different sets of treatment comparisons.

Experience shows that the BIBD is usually the appropriate design to adopt in evaluating software systems. The reason is that, evaluating software systems is usually not constrained by using sensitive resources, the situation that may limit many biological or chemical experiments from being conducted. In fact, the resources needed for software evaluation (e.g., software products and expert users) are usually manageable, or at least can be controlled at the expense of having more expert users.

3.3 MAAF Process

The MAAF is an eight step process as defined below (see Figure 2):

Step 1. Select target software products

To select the software products being evaluated, one starts with conducting a primarily survey to review a set of competing candidates and select the most qualified ones. A qualified product can be defined as the one that satisfies some generic assumptions such as: (a) has reasonable documentation to describe it; (b) is fairly known to the community; and (c) has a reasonable domain of applicability, etc.

Step 2. Identify dimensions

In this step, one identifies a set of the evaluation criteria that represent the dimensions and the hierarchical structure of our framework. Examples of the dimensions are: modeling, communication, process, support, etc.

Step 3. Identify attributes

In this step, one determines the relevant features (i.e., attributes) pertained to each dimension. This also includes constructing a hierarchy structure and validating its consistency to ensure that no redundancies exist among the attributes for all the dimensions. For example, modeling-related dimension may consist of attributes that address and examine the most common and important aspects to model the product, such as: notation, expressiveness, abstraction, consistency, concurrency, traceability, derivation, reusability, etc.

Step 4. Design questionnaire

One has to design a set of questions corresponding to the dimensions and their attributes. The questions must be understandable, unambiguous, and provide clear statements to examine the effectiveness and strength of the related attributes. When designing questionnaire, it is important to set up the appropriate scale of measurement (e.g., nominal, ordinal, interval, or ratio) based on the nature of the collected data [22].

Step 5. Select statistical model

To perform analysis, one has to select the most appropriate statistical model and procedure that can fit and treat the data. This step is also helpful to determine the proper number of observations needed (and consequently the number of expert users needed to give their feedback to the evaluation questionnaire) to achieve reasonable accuracy of the statistical analysis.

Step 6. Select expert users

After determining the proper number of expert users (aka. participants), one has to select qualified participants to deliver the questionnaire with detailed guidelines to assure clarity. It is also recommended to hold instructional sessions to explain the evaluation task, the anticipated results, and the proper way to respond to the questionnaire. The participants should receive sufficient documentation about the products being evaluated, clear instructions about the experiment, and equal amount of time to complete their tasks.

Step 7. Collect and validate responses

The collected data will be validated to assure completeness and accuracy. One way to do this is to simply run a rough test on the collected data to detect outliers, for instance, by using scatter plots. In the case that outliers are observed, it is recommended to consult with the expert users who provided the data to make sure that the meant values are correct and not mistakenly recorded.

Step 8. Perform statistical analysis

The major step for implementing MAAF is to conduct a statistical experiment to evaluate the given products (see Figure 2). Prior to this step, one has to identify the statistical hypotheses and end up by testing the statistical significance of the hypotheses, analyzing the obtained results and drawing the final conclusions. The statistical hypotheses are:

- Null hypothesis: There is no significant difference in the mean effectiveness of the examined dimension among the evaluated products.
- Alternative hypothesis: There is a significant difference in the mean effectiveness.

Then, one can analyze the data statistically by applying the analysis of variances (ANOVA) procedure to the model. The underlying idea of ANOVA is to compare the variability of the observations between groups to the variability within groups. If the variability between groups is smaller than the variability within groups, it means that different groups are not significantly different, whereas if the variability between groups is larger than the variability within groups, it implies that different groups are significantly different.

If some variability is identified among the evaluated products on a certain dimension, Tukey's test for pairwise comparison of the products is performed to test for multiple comparisons to identify which products are actually different. In contrast, if the overall ANOVA test was insignificant, applying any pairwise comparison is not necessary. In such a case, the conclusion to be made

is that all the products are statistically equal in their main effects against the attributes of the examined dimension.

The treatments are then ranked based on their means calculated for each dimension. An overall ranking of the treatments is finally calculated statistically.

It should be noted that prior to applying ANOVA, one may elect to examine the adequacy assumptions by testing the outliers, normality of residuals, and the homogeneity of residuals [10]. Table 1 shows the tests defined and used in the GPSE system.

Table 3: Suggested ANOVA tests.

Test	Test Type	Instrument Used
1	Outliers	a. Normal probability plot of residuals b. Individual value plot of residuals versus independent variable
2	Normality of residuals	Normal probability plot of residuals
3	Homogeneity of error variances	a. Residual plots against fitted values b. Bartlett's test

3.4 MWAF Advantages

Compatibility: MWAF is capable to conduct evaluation studies that are similar to many cases presented in the literature. This is because MWAF recognizes and integrates the important features of other frameworks, overcomes any obvious deficiencies, and adopts new features that generalize and extend its usability.

Structure: MWAF can be represented by an effective hierarchical structure, which derives its power from the principle of ‘divide and conquer’ that contributes to analyzing a complete taxonomy of evaluation attributes.

Scalability: MWAF is flexible to scaling up/down in order to expand or reduce its dimensions and/or attributes. In addition to its capability in supporting the conventional evaluation of software, MWAF can fit evaluation studies that are characterized by its dynamic nature; for instance, optimizing an objective function (e.g., maximizing overall performance, marketability, or minimizing costs or potential risks) by simulating potential features that can be released to a new product.

4 GPSE system analysis and design

The current stand-alone implementation of the GPSE system has a graphical user interface (GUI), a database, and a statistical analysis unit as shown in Figure 3.

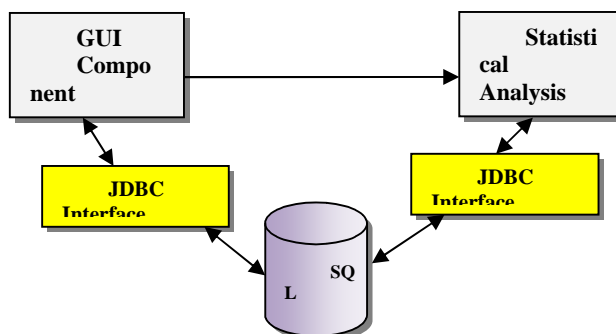


Figure 3: Overall architecture of GPSE system.

The functionalities of GUI facilitate configuration of the MWAF framework, collection of data comprising expert users’ ratings, initiation of the statistical evaluation process, and displaying of the analysis results. A database is required for storing and retrieving information pertaining to the MWAF configuration data, expert users’ ratings, and results of statistical analysis. As explained in Section 3, the key functionalities of the analysis unit include ANOVA method and Tukey’s pairwise comparison tests. Figure 4 depicts the interactions among the system components.

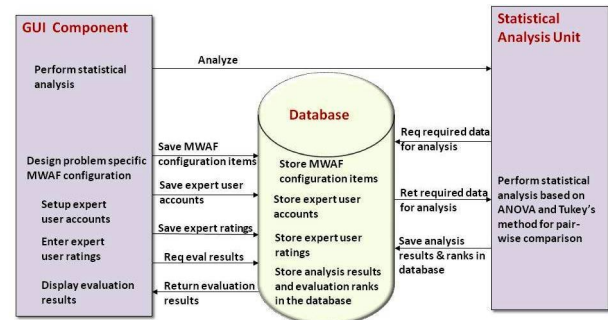


Figure 4: GPSE system component interaction.

The GPSE system is implemented using JAVA technologies (<http://java.sun.com>). MySQL database (<http://www.mysql.com/>) is used for data storage and retrieval purposes. In order to access MySQL database, Java Database Connectivity (JDBC) is used from both GUI component and statistical unit.

Current version of the GPSE system can be broken down into six main functionalities or Use Case Diagrams (UCD) as shown in Figure 5.

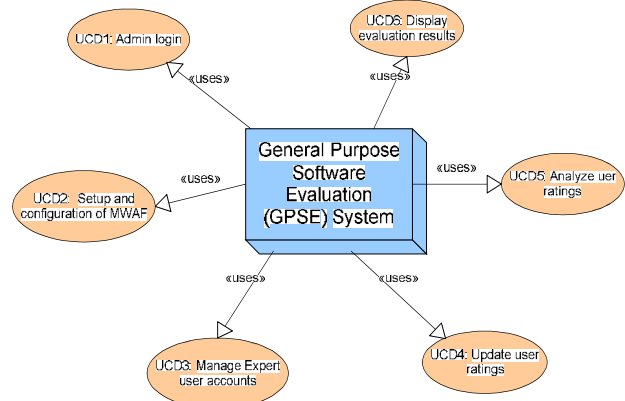


Figure 5: High level functionality diagram representing the functional decomposition of the GPSE system.

UCD1 – Admin Login: This functionality deals with logging into the GPSE system with administrator privileges. Administrator logs into the GPSE system in order to create expert users’ accounts, setup the MWAF configuration framework, initiate the statistical analysis on the ratings data provided by expert users, and display results for the software evaluation problem.

UCD2 – Setup and Configuration of MWAF: This functionality deals with the configuration of MWAF architecture for the given evaluation problem. As part of this configuration setup the administrator will use the

GPSE GUI to setup the names of dimensions and the attributes for each dimension for various products that need to be evaluated. Subsequently, the framework will be accessed by various experts to provide ratings. Along with dimensions, attributes, and treatments, the MWF framework also consists of blocks and replicas. A block consists of a set of treatments. Each user is assigned with a block of treatments to be evaluated. Managing blocks in terms of addition and deletion of blocks as well as assigning treatments to blocks is dealt by this functionality. A replica consists of the whole set of treatments considered for the particular evaluation problem. Managing the number of replicas for this evaluation problem is also included in this functionality.

UCD3 – Manage Expert User Accounts: This functionality deals with the creation of user accounts for various expert users who will use the system in order to provide their expert ratings based on the framework designed by the Administrator. During the creation of expert users of the system, each expert user is assigned with a block of treatments in order to provide their ratings to the assigned block of treatments.

UCD4 – Update User Ratings: The functionality covers the aspects related to the expert user log in, updating of ratings provided by the expert user, and saving the expert ratings to the database.

UCD5 – Analyze User Ratings: This functionality captures the statistical analysis of the evaluation process. After all the user ratings are provided, the administrator will access the system to initiate statistical analysis of the data stored in the system for the given evaluation problem. The administrator will request the system to analyze the user ratings data.

UCD6 – Display Evaluation Results: This functionality captures the aspects related to the display of the evaluation results. After the ratings from various expert users are analyzed, the administrator instructs the system to generate output results for display. Using this functionality, the administrator can also provide and save summary and recommendations for the evaluation conducted.

The stand-alone GPSE system is fully implemented, tested and verified. Figure 4 shows various screenshots of the stand-alone GPSE system. The system has been tested rigorously with the data that was collected in a set of experiments [10, 11] for the evaluation of various agent-oriented software engineering (AOSE) methodologies. The results obtained from the manual data analysis in the experiments were compared with the results from the GPSE system. Hence, the system fulfils its technical goals in that a functional GPSE system based on MWF is developed and meets the desired objectives.

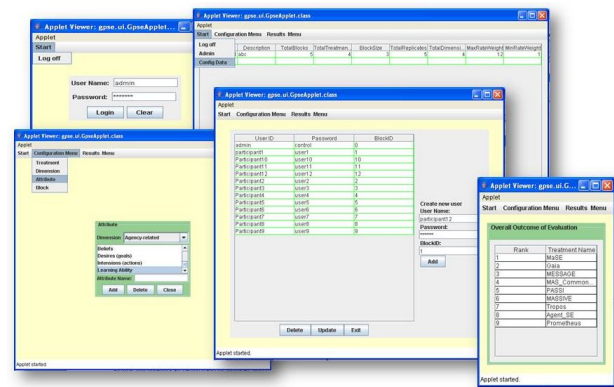


Figure 6: Screen shots of the GPSE systems.

Currently the GPSE system is being redesigned and converted to a multi-agent system application. In this case, several remote expert users may provide their ratings for the products that need to be evaluated. Furthermore, the MWF process is also being automated using a multi-agent system in which intelligent software agents are responsible for steps of the MWF process. For example, an agent, incorporated with the knowledge of designing statistical experiments, is responsible for the design of experiment (Step 5 of the MWF process) and another, incorporated with data mining capabilities, for selecting dimensions and attributes (Steps 2-3 of the MWF process). These dimensions and attributes are then sent to other agents that act as personal assistant agents for expert users for voting the alternatives based on specified evaluation criteria. These votes are statistically verified by yet another agent in order to find significant similarities among the votes to derive the rankings for the alternatives (Steps 7-8 of the MWF process).

5 Case study

The GPSE system has been applied to several cases including the followings:

- Selection of Agent-Oriented Software Engineering (AOSE) methodologies
- Selection of software testing tool
- Selection of COTS – single
- Selection (configuration) of COTS – multiple

In this case study we use the GPSE system for deciding what AOSE methodology is the best to adopt for developing a multi-agent system. So far, there is no industry-wide agreement on the kinds of features a methodology should support. Evaluation is a crucial and critical task here to identify the differences between several AOSE methodologies. The GPSE provides a reliable solution with accurate results based on applying state-of-the-art statistical procedures to evaluate AOSE methodologies and comes up with a set of measures that help in selecting the most appropriate methodology for developing prospective agent-based applications.

To select the software items being evaluated (e.g., AOSE methodologies in our case), we started by

conducting a primarily comparative survey to review a large number of AOSE methodologies and select the most qualified ones. After reviewing 31 properly-documented methodologies against the qualification assumptions (Section 3.3 Step 1), the following 9 methodologies were selected: Gaia, MaSE, Tropos, Agent-SE, MASSIVE, Prometheus, MESSAGE, MAS-Common-KADS, and PASSI [10].

Then the dimensions were identified. We studied the selected nine methodologies comprehensively to identify the most important and common measures that will be used as evaluation criteria. Consequently, we came up with six primarily criteria that we indicated by the following dimensions:

- Dimension 1: Agency-related attributes
- Dimension 2: Modeling-related attributes
- Dimension 3: Communication-related attributes
- Dimension 4: Process-related attributes
- Dimension 5: Application-related attributes
- Dimension 6: User-perception attributes

Then the relevant attributes for each dimension were identified. We broke down each dimension into a number of relational attributes that describe its main features as follows:

Dimension 1: Agency-related attributes

This dimension contains attributes that address features involving the internal properties and basic architecture of agents. The hierarchical structure of this dimension is shown in Figure 7.

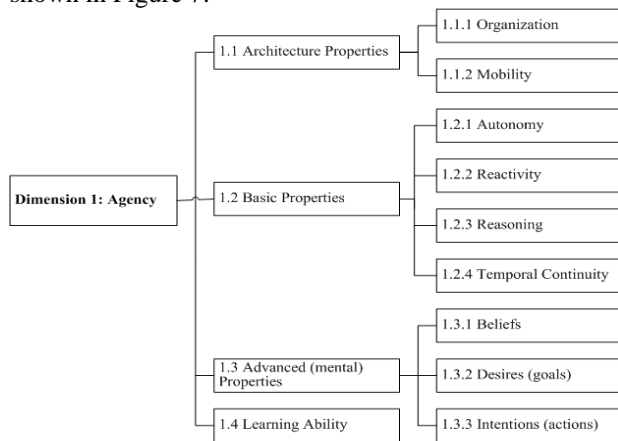


Figure 7: Hierarchical structure of Dimension 1.

Dimension 2: Modeling-related attributes

This dimension consists of the following attributes that address and examine specific features to describe the most common and important aspects to model agents.

1. Notation
2. Expressiveness
3. Abstraction
4. Consistency
5. Concurrency
6. Traceability
7. Derivation and reusability

Dimension 3: Communication-related attributes

This dimension encompasses the following attributes that address features related to the possible interactions and interfacing of agents.

1. Local Communication: cooperation; coordination; competition; negotiation.
2. Wide Communication: interaction with the external environment; agent-based user interface; subsystems interaction.

Dimension 4: Process-related attributes

This dimension encompasses a number of attributes that are given by the following hierarchy to address and examine several issues involving the development process of agents and multiagent systems.

1. Development lifecycle: architectural design; detailed design; verification and validation
2. Refinability
3. Managing complexity

Dimension 5: Application-related attributes

This dimension includes attributes that address and assess different aspects involving the methodology’s applicability, and examine some socio-economic factors that affect the decision of recommending and adopting an AOSE methodology.

1. Applicability
2. Maturity
3. Field history
4. Cost concerns

Dimension 6: User perception attributes

In order to make a decision on whether to adopt a specific AOSE methodology, perception, which is entirely a subjective feature, is important and substantial. This is due to the effect of the natural intentionality in human behavior [12]. User perceptions are assessed through the following attributes:

1. Perceived ease of use
2. Perceived usefulness
3. Intention to use

In the next step we used the GPSE system to select the appropriate statistical model and procedure that can fit to our data. In this step we have to determine the proper number of observations needed (and consequently the number of evaluators needed to give their feedback to the evaluation questionnaire) to achieve reasonable accuracy of the statistical analysis. In this case, we have nine methodologies that will be treated statistically as *treatments* and we decided to have at least 4 replicates per treatment. Balanced Incomplete Block Design (BIBD) model was selected.

By denoting the 9 methodologies with letters from A to I and assigning each block of 3 methodologies- after selecting them in such a way to be as homogeneous as possible- to one participant, we can

obtain 36 replicates that will sufficiently satisfy our goal of having 4 replicas per treatment for 12 participants. Table 4 shows this assignment.

The collected data was validated to assure completeness and accuracy using the tests suggested in Table 3. Then the statistical analysis unit was deployed. Followings are the main results of this process.

In order to determine whether significant differences exist among the evaluated methodologies, we conducted separate experiment for each of the six dimensions that characterize the nine methodologies. That is, we will conduct six individual experiments.

Table 4: BIBD tableau for blocks and treatments.

Block (Participants), j	Treatments (AOSE Methodologies), i								
	M1	M2	M3	M4	M5	M6	M7	M8	M9
	A	B	C	D	E	F	G	H	I
A			D			G	H		I
	B			E					
		C			F				I
A					F		H		
	B		D						I
		C		E		G			
A				E					I
	B				F	G			
		C	D					H	

M = Methodology; P = Participant

In this context, each set of attributes representing a specific criterion given by a dimension were investigated statistically over the nine methodologies. This helped determine whether the strength or effectiveness of this dimension differs among the evaluated methodologies. The following set of hypotheses describes this strategy.

Null hypothesis: $H_0: \tau_i=0$, for $i=1$ to 9

Indicating that there is **no** significant difference in the mean effectiveness of the examined dimension among the evaluated AOSE methodologies.

Alternative hypothesis, H_a : at least one $\tau_i \neq 0$

Implying that there is significant difference in the mean effectiveness.

The statistical analysis unit in GPSE showed that the mean effectiveness of all the evaluated dimensions (except Dimension 5: Application-related) differs among the evaluated nine methodologies. As a result, the methodologies were ranked for each dimension according to their estimated adjusted mean of effectiveness as shown in Table 5.

Table 5: Ranking evaluated methodologies for each dimension based on mean of effectiveness.

Dimension 1: Agency		
①	M1: Gaia	$\hat{\mu}_1 = 6.494$
②	M2: MaSE	$\hat{\mu}_2 = 5.861$
③	M3: Tropos	$\hat{\mu}_3 = 5.835$
④	M9: PASSI	$\hat{\mu}_9 = 5.713$
⑤	M8: MAS-Common	$\hat{\mu}_8 = 5.549$
⑥	M7: MESSAGE	$\hat{\mu}_7 = 5.524$

⑦	M4: Agent-SE	$\hat{\mu}_4 = 5.192$
⑧	M6: Prometheus	$\hat{\mu}_6 = 5.049$
⑨	M5: MASSIVE	$\hat{\mu}_5 = 4.684$

Dimension 2: Modeling

①	M2: MaSE	$\hat{\mu}_2 = 6.593$
②	M9: PASSI	$\hat{\mu}_9 = 6.428$
③	M1: Gaia	$\hat{\mu}_1 = 6.037$
④	M7: MESSAGE	$\hat{\mu}_7 = 5.777$
⑤	M8: MAS-Common	$\hat{\mu}_8 = 5.560$
⑥	M5: MASSIVE	$\hat{\mu}_5 = 5.271$
⑦	M6: Prometheus	$\hat{\mu}_6 = 5.074$
⑧	M4: Agent-SE	$\hat{\mu}_4 = 4.755$
⑨	M3: Tropos	$\hat{\mu}_3 = 4.580$

Finally, the evaluated methodologies were ranked according to the accumulated proportional order of their dimensions. For example, the *Gaia* methodology, M1, has the following accumulated proportional order: $1(\leftarrow D1) + 7/9(\leftarrow D2) + 4/9(\leftarrow D3) + 1(\leftarrow D4) + 8/9(\leftarrow D6)$, where the arrow points to the dimension contributing the proportional value. In this way, we determined the accumulated proportional order of each methodology as well as the overall ranking as shown in Tables 6, 7 and 8. Note that we discarded the proportional orders given to dimension D5 because no significant differences were detected.

Table 6: Dimension ranks for the AOSE methodologies.

Methodology, M _i	Order	Proportional Order	Dimension, D _i					
			D1	D2	D3	D4	D5	D6
1	9/9	M1	M2	M7	M1	M2	M2	
2	8/9	M2	M9	M8	M2	M1	M1	
3	7/9	M3	M1	M9	M5	M9	M7	
4	6/9	M9	M7	M2	M8	M7	M8	
5	5/9	M8	M8	M5	M9	M5	M5	
6	4/9	M7	M5	M1	M6	M8	M9	
7	3/9	M4	M6	M3	M4	M3	M4	
8	2/9	M6	M4	M4	M7	M4	M3	
9	1/9	M5	M3	M6	M3	M6	M6	

Table 7: Accumulated proportional order of the nine methodologies against the evaluated six dimensions.

Methodology	Total weight
M1: Gaia	$[9+7+4+9+8]/9 = 37/9$
M2: MaSE	$[8+9+6+8+9]/9 = 40/9$
M3: Tropos	$[7+1+3+1+2]/9 = 14/9$
M4: Agent-SE	$[3+2+2+3+3]/9 = 13/9$
M5: MASSIVE	$[1+4+5+7+5]/9 = 22/9$
M6: Prometheus	$[2+3+1+4+1]/9 = 11/9$
M7: MESSAGE	$[4+6+9+2+7]/9 = 28/9$
M8: MAS-Common	$[5+5+8+6+6]/9 = 30/9$
M9: PASSI	$[6+8+7+5+4]/9 = 30/9$

Table 8: Overall ranking of the AOSE methodologies.

Rank	1	2	3	3	4	5	6	7	8
Method	M2	M1	M8	M9	M7	M5	M3	M4	M6

6 Conclusions and future works

In the present day market, a customer is faced with various alternatives in the selection and purchase of a software product or deployment of a certain process. This work focused on designing and developing a General Purpose Software Evaluation (GPSE) system. The main objective of the GPSE system is to evaluate various software systems, that are available for a given business application, in order to select the most suitable product or process that meets the requirements of the application as well as the preferences of expert users in an effective manner. In order to obtain expert users' ratings for the products or processes, the Multi-dimensional Weighted Attribute Framework (MWAF) is proposed and adapted [10]. The framework allows the user to define and configure significant evaluation criteria in the form of dimensions and attributes for each dimension. Each treatment considered in the evaluation is rated as per the defined criteria. The GPSE system makes use of statistical analysis based on ANOVA and pairwise comparison tests for ranking the software products or processes. The user of the system is provided with overall ranking of the evaluated systems as well as ranking in all the major evaluation areas. This analysis will help the user in selecting an appropriate product or process to address his/her business needs. This system can be used for the evaluation of any software, hardware or any other product or system where there is a need for the selection to be made from among various alternatives with similar functionalities.

The stand-alone GPSE system, in its current implementation, consists of GUI, database, and a statistical analysis unit. The design of the system offers adequate flexibility to enable users to adapt and use the system for a variety of applications that require complex decision making based on the evaluation of multiple options. Furthermore, the technological choices were made with a due consideration on the portability of the system onto a variety of platforms thus enhancing the overall utility of the concept.

The GPSE system uses ANOVA method for assessing whether there are significant differences among the products being evaluated against a given set of dimensions and their attributes. As a result, if the underlying data does not support the assumptions that need to be satisfied for the application of ANOVA, the method may not be employed effectively. For such cases, usage of suitable data mining techniques will be helpful. Furthermore, the MWAF framework which is the foundation of the present GPSE system requires sufficient prior knowledge on the part of a user to determine the criteria that are considered significant for the evaluation. Eliciting and implementing this knowledge inside the system will contribute to the improved usability of the GPSE system.

References

- [1] Alves C., and Castro, J., "CRE: A Systematic Method for COTS Selection", *Proc. XV Brazilian*

Symposium on Software Engineering, Rio de Janeiro, Brazil, 2001.

- [2] Bayer, P., and Svantesson, M., "Comparison of Agent-oriented Methodologies Analysis and Design," *Programming, Blekinge Institute of Technology (BITSWAP)*, 2001.
- [3] Burgues, X., Estay, C., Franch, X., Pastor, J.A., and Quer, C., "Combined Selection of COTS Components", *Proc. 1st Int. Conf. on COTS-Based Software Systems (ICCBSS'02)*, Orlando, Florida, 2002, pp. 54-64.
- [4] Cavanaugh, B.P., and Polen, S.M., "Add Decision Analysis to Your COTS Selection Process", *Journal of Defense Software Engineering*, April 2002.
- [5] Clarke, G., and Kempson, R., "Introduction to Design and Analysis of Experiments," John Wiley & Sons, Inc., 1997.
- [6] Cochran, W., and Cox, G., "Experimental Designs," 2nd ed., John Wiley & Sons, Inc, 1992.
- [7] Chung, L., Nixon, B., Yu, E., and Mylopoulos, J., "Nonfunctional Requirements in Software Engineering," Kluwer Academic Press, 2000.
- [8] Chung, K., Cooper, K., and Courtney, C., "COTS-Aware Requirements Engineering: The CARE Process", *Proc. 2nd International Workshop on Requirements Engineering for COTS Components*, Kyoto, Japan, September 7, 2004.
- [9] Dam, K., and Winikoff, M., "Comparing Agent-Oriented Methodologies," *Proceedings of the 5th Int'l Bi-Conference Workshop on Agent-Oriented Information Systems (AOIS'03)*, Melbourne, Australia, 2003, 78-93.
- [10] Elamy, A., "A Statistical Approach for Evaluating Agent-Oriented Software Engineering Methodologies," MSc. thesis, Department of Electrical and Computer Engineering, University of Calgary, 2005.
- [11] Elamy, A., and Far, B., "A Multidimensional Weighted-Attributes Framework (MWAF) for Evaluating Agent-Oriented Software Engineering Methodologies," *Proceedings of the 19th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'06)*, Ottawa, Canada, 2006.
- [12] Fishbein, M., and Ajzen, I., "Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research," Addison-Wesley, Boston, 1975.
- [13] Gross, D., and Yu, E., "Evolving System Architecture to Meet Changing Business Goals: An Agent and Goal-Oriented Approach," *Proceedings of the ICSE-2001 Workshop: From Software Requirements to Architectures (STRAW'01)*, 2001, pp. 16-21.
- [14] Juneidi, S., and Vouros, G., "Evaluation of Agent Oriented Software Engineering Main Approaches," *Proceedings IASTED Int'l Conf. on Software Engineering (SE'04)*, Innsbruck, Austria, 2004.
- [15] Kontio, J., "A Case Study in Applying a Systematic Method for COTS Selection", *Proc. ICSE-18*, 1996, pp. 201–209.

- [16] Kunda, D., “A social-technical approach to selecting software supporting COTS-Based Systems”, *PhD Thesis*, Department of Computer Science, University of York, Oct. 2001.
- [17] Liu, M., and Chan, L., “Uniformity of incomplete block designs,” *Int’l Journal Materials and Product Technology*, vol. 20, no. 1–3, 2004, pp.143–149.
- [18] Ncube C., and Maiden, N.A.M., “Guiding Parallel Requirements Acquisition and COTS software”, *Proc. IEEE International Symposium on Requirements Engineering*, 7-11 June 1999, pp. 133 – 140.
- [19] Neter, J., Wasserman, W., and Kutner, M., “Applied Linear Statistical Models,” 5th ed., Irwin, USA, 1996.
- [20] Silva, C., Tedesco, P., Castro, J., and Pinto, R., “Comparing Agent-Oriented Methodologies Using NFR Approach,” *Proceedings of the 3rd Workshop on Software Engineering for Large-Scale Multi-Agent Systems (SELMAS’04)*, Edinburgh - Scotland, vol 1, 2004, pp. 1-9.
- [21] Shehory, O., and Sturm, A. “Evaluation of Modeling Techniques for Agent-Based Systems,” *Proceedings of the 5th Int’l Conference on Autonomous Agents*, May 2001, Montréal, pp. 624-631.
- [22] Stevens, S., “Handbook of Experimental Psychology,” John Wiley and Sons, New York, 1951.
- [23] Sudeikat, J., Braubach, L., Pokahr, A., and Lamersdorf, W., “Evaluation of Agent-Oriented Software Methodologies – Examination of the Gap between Modeling and Platform,” *Proceedings of the 5th Int’l Workshop on Agent-Oriented Software Engineering (AOSE’04)*, Jul. 2004, pp. 126-141.
- [24] Tran, Q., Low, G., and Williams, M., “Comparison of Ten Agent-Oriented Methodologies,” *Agent-Oriented Methodologies*, B. Henderson-Sellers, P. Giorgini, Eds., Idea Group Inc., PA, USA, 2005.
- [25] Yates, F., “Experimental Design: Selected Papers,” Griffin, UK, 1970.

Online WordNet Based Tagging System for Social Sharing and Retrieval of Images on Visited Pages

Tansel Özyer

Department of Computer Engineering, TOBB Economics and Technology University, Ankara, Turkey

E-mail: ozyer@etu.edu.tr

Keywords: social networks, collaborative filtering, web information retrieval, XML

Received: October 12, 2008

With the advent of internet technologies, users get involved in collective organizations that feed from collective intelligence mechanism to make further recommendations, inferences on unseen examples. In this study, as users surf on the internet, they are capable of picking pictures liked from web pages to their favorite list with definitions specific to our system. After images are tagged with definitions, they are filtered with respect to commonality measure to deduce significant definitions for storage. These tags help users to make inquiries and broaden up their search on the tags they are interested. While doing tagging process, all images are shared collaboratively and that is bolstered with Word net.

Povzetek: Opisano je sprotno označevanje slik s pomočjo WordNeta v socialnih omrežjih.

1 Introduction

1.1 Problem Definition

Researches point out that nowadays news, videos and pictures of the 7% of the sites visited is being tagged [1]. Also, 28% of people benefit from tagging.

While surfing on the internet, users are capable of bookmarking resources such as images on the internet. In a social environment people share information and exchange. Instant exchange of ideas and finding the image sought can occur by clicking on an image found at a document and by asking "What are the images similar to that image?" We are looking forward to finding similar images to that in social behavior according to the descriptions made. Otherwise, traditional image retrieval techniques will remain insufficient according to real time requirements at first.

Sites such as deli.ci.ous.com, <http://www.flickr.com> (picture sharing) and <http://www.youtube.com> (video sharing), digg, fark, StumbleUpon, Reddit, Slashdot are some of favourite ones as a social bookmarking system [16]. It has become widespread because it is very simple and practical process. However, there are also some problems with tagging in collective manner such as using synonym, polysemy words and etc. Tagging a picture depends on its interpretation by the owner. Keywords entered for the picture give the bookmarking info. User, himself can organize his own data besides, he can share his information with other users in the community.

For example, frequently used information bookmarked is stored at <http://deli.ci.ous.com> keeps track of popular tag info. It is categorized according to the group name which is the tag value. When someone indicates that a web page is about a topic then the link is being put together with the other links having the same topic specification.

Whenever the user wants to reach the links of a specific topic then, all links that are related are brought back. Throughout the development of web site, some statistical methods turn the web site into more efficient situation. There are some basic rules about formatting and describing tags [2]. When users obey, it becomes easier to access and manipulate [3].

1.2 Motivation

Bookmarking can be supported by tagging utilities. Tagging is beneficial for searching bookmarked items. This means, pictures are supported by words/terms for description. Basically, background, action, objects of an image can take part. Stand-alone tagging can be used to some point but support in a collaborative environment for sharing can be useful for finding similar images to an image. One can find or suggest similar alternative images to that image. All tagged information may help users share images between each other. Tags may need to be categorized and tag entries should be unified in a sense that they mean the same meaning to each other while doing similarity check.

1.3 Contribution

This study proposes a collaborative infrastructure that helps users store pictures that they have seen on web pages during surfing. Basically, system consists of two modules: One module is an add-on application that will work on Firefox Mozilla browser and the other one is the web page itself. The add-on application helps user tag and store pictures that will remind of something meaningful to him. Tagging is being used very often in variety of internet applications in order to expedite

access to the resources on the web. Another aim of tagging is using internet as a private agenda. Tagging merits users organizing data with respect to their special needs.

Web 2.0 is an emerging platform that helps fast development of interactivity and interconnections of web data. Our study benefits from web 2.0 and social bookmarking utilities.

In this study, we propose a novel study that employs the use of Wordnet in order to unify the tags described by the users and image similarity is measured by using Jaccard coefficient. Tools and technologies that have been used for our system are described at the next section.

1.4 Technologies and Tools

C# and ASP.Net have been used to build the web site and database has been used at the server side. Firefox has been used as the web browser tool and communication between web services and Firefox has been provided with AJAX (Asynchronous JavaScript and XML) technology. We have created an interface for add-on application to be used via XUL and JavaScript for functionality. We have used VS Studio and MS SQL Server Express, Firefox and Firebug Extension Developer plug-in for Firefox.

2 System overview

As is mentioned, system is formed of two parts. These are Firefox add-on application that works on browser in order to use on the client for tagging and the repository at the server.

2.1 Firefox add-on application

Mozilla Firefox is a web browser that is licensed under MPL/GPL/LGPL. As of today, Firefox is being used around 45% (was %25 in 2006) of users and it becomes one of the most common web browser all over the world. One of the main reasons for the gradual use of the web browser is that it's an open source application and third party developers can easily implement plug-ins onto the browser as an extension.

Firefox extensions basically built over two structures. One of them is XUL(XML User-Interface Language) for interface definition; and the other one is JavaScript that has been used for the client side functionality requirements.

XUL is an XML language that has been designed to define new user interfaces or modify existing ones. It is capable of using XSLT, XPath, and DOM standards. Entire user interface of Firefox has also been defined with the same language that makes it convincing about how powerful language it is [4].

Add-on tool listens to the click events on a web page that is currently open. If a click occurs on a page, it detects whether the click is on an image or not. When a user clicks on a picture, the application will display a pop up window accordingly. Next, the user can check his user information and get the potential tagging recommendations in cooperation with the server and the

user can arrange the tags for the picture by also taking to the given recommendations into account. After arranging the tags, tag information about the image will be sent to the server. Later on, tag information that has been processed can be retrieved back from the server again for confirmation to the user.

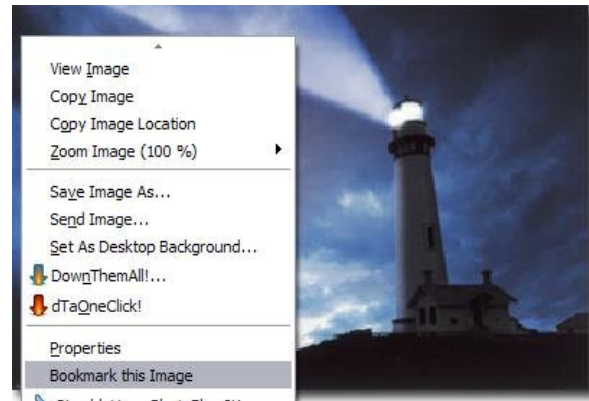


Figure 1: Right Mouse Cursor Click.

During the process, one of the most important points is the arrangement of the obtained tag information. Information can be processed at both client browser and server. However, it will be better to pre-processed at client side by using JavaScript in order to reduce the workload of server. Arrangement of tagging information has two meanings: First one is stripping out the words from whitespace and non-alphanumeric characters; the second is stemming words in order to reduce sparsity and group words under specific categories disregarding the suffixes (e.g. performing, performed, performance and etc.). Popular stemming algorithm, Porter's stemming algorithm has been used [5]. All terms are pre-processed at client browser side and sent to the server.

After cleaning up the words, they are separated with a delimiter. The reason for using delimiter is that it is impossible to send a JavaScript object array directly to a web service. Instead, a string will be sent as a parameter. Another point is to reduce the communication cost as much as possible. In order to accomplish this, recent technology, Ajax has been used to prevent refreshment of entire page at each time a change occurs. This is done by using the basic Ajax object XMLHttpRequest. XMLHttpRequest object provides an asynchronous connection between two parties that is, user is not supposed to wait for the reply of the request sent back and only related portions of the web page is refreshed instead of entire page[8].

An image can be tagged to describe the content of the image for describing the content as well as specific notes taken onto it. Description part of the image content is composed of different parts as <objects>, <actions>, and <background> information fields (Figure 2). Here, a user is free of entering all the information about the picture for tagging at the browser instantly and keep visiting pages during the day.

2.2 Server Application

There are two main tasks at the server side. These are processing and storing tag info and forming the infrastructure for sharing environment.



Figure 2: Tagging an Image.

2.2.1 Communication with browser add-on

The most prominent job of the server related to the add-on tool is the language processing of tag information. In other words, information is checked. If the entries are adjective, verb, adverb and etc. WordNet database has been put in use to accomplish this at the server side.

WordNet was created and has been maintained by the cognitive Science Lab of the Princeton University. It is a semantic lexicon for English. It is used to group words into sets such as synsets. By employing wordNet, it is possible to access the semantic relations between words and their definitions and roots. [6].

After pre-processing, tagged information is sent to the server and server revises them and sends back to the client browser independent from the user for confirmation. User is free of doing modifications on the tags and after confirmation, server stores user, image link, and tag information at the database.

Another aspect of the communication is the user control. This is performed in the following way:

Firefox add-on application at the client first checks if the user whether signed on or not by holding a variable kept in his DOM-like structure.

If signed on then the user is authorized. Otherwise,

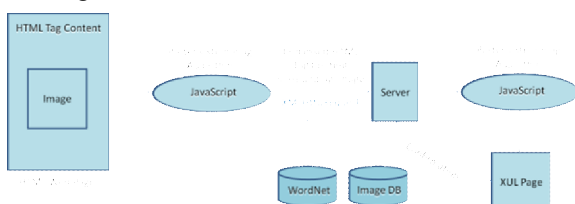


Figure 3: The Flow Diagram of Tagging Process.

user name and password are requested. Ajax based requested user name and password are obtained and sent to a web service with GET method. This service basically fulfils the authentication of the system. Server controls the access with one time login.

The web services are used to support different applications to interwork via machines and exchange messages in XML format [7].

Data flow can be summarized as the following: An image can be accessed via DOM structure. Tag element content surrounding the image. In this study, we extracted the alt and innermost tag element content as the image content information (Figure 4). Non-alphanumeric characters are removed and content is spit into words. These later are put into a string to be able to send to a web server as a parameter. This information is sent to the web service to the server for feedback to the end user. Details of the feedback mechanism will be discussed later but shortly, it includes the interaction between the image db and wordNet in Figure 3.

```

html>
<head>
<title>... </title>
<body>
...

...
</body>
</html>
    
```

Figure 4: HTML Page.

After the tag information is arranged, the feedback is sent to Firefox browser add-on in XML file. Firefox presents it to the end user with an interface that is constructed with XUL.

User is free of accepting all tags specified or they can be changed. After confirmation, final tag information is sent to another web service at the server side. All parameters are resolved and they are used for updating database tables (user id, link address, word/word type).

2.2.2 The Database Model

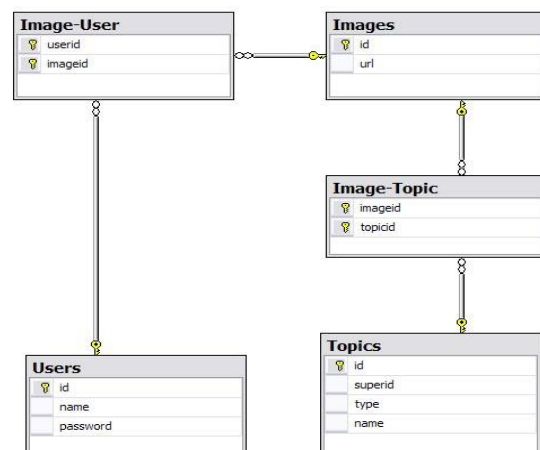


Figure 5: Database Relationship Model.

Database relationship model is given in figure 5. There are five tables they are namely, Users, Image-User, Images, Topics, and Image-Topic tables.

Users table contains the user specific information; Images table contains image specific link information. The most important table is the Topics table. It keeps the tag information: Words, their types (adjective, noun ...) and their hypernym. Topics table is linked to WordNet database.

2.2.3 Wordnet

WordNet has been used for tagging words. While tagging, in order to give idea to the end-user and grouping tag content is provided. It contains approximately 147000 words. WordNet has been defined in a relational database and this formation has been used in our system[10]. Synonym set of a word can be used by doing inquiries on the database. The SQL query that has been used for this purpose has been given figure 6.

```
select lemma from word where wordid in
(select distinct wordid from sense where
synsetid in
(select synsetid from sense where wordid in
(select wordid from word where lemma=@word)))
```

Figure 6 – SQL Query for Finding Synonym Set(Synset)

We have used stored procedures. By using the SQL Query in figure 6, synonyms of a word can be retrieved from the database. While tagging words, synonym words are found. Another function of WordNet is grouping the words in terms of their categories. *categorydef* in wordNet[10] can be used for this purpose. Categories are noun, adjective, adverb and verb. These categories are also subgrouped. For example, nouns can be event, location, and animal, human-being nouns. Each word has a *category_id* and it is related with *synset* table. By using these properties, category of the word can be found(Figure 7).

```
select [name] from categorydef where categoryid in
(select categoryid from synset where synsetid
in
(select synsetid from sense where wordid in
(select wordid from word where
lemma=@word)))
```

Figure 7- SQL Query for Finding Category of Word

Tag information of the image (*alt* property and surrounding innermost element content) are used identify the content of the image.

Whenever potential tag information has been received, each word that has been obtained will be matched with similar words existing at wordNet database and potential tagging content is intended to be increased. Here, type info keeps the type of the word(noun, verb, adverb and etc.) and superid field keeps the broader meaning of word(apple is a fruit).

2.2.4 Web Services

One of the major parts of our system is the web services. There are three web services that was used. These are for:

- User authentication
- Feedback mechanism
- Update of database server with confirmed data.

The most important web service is the feedback mechanism. Our aim is to categorize them according to their types. Object (noun), action (verb) and background. After composing a string message having delimiters in between words, for each word, first we check whether the specified word exists in database or not. If exists, type info is obtained from our database content otherwise, for all similar words and their corresponding types are sought. If found then type of words are the same as their similar words' type. Otherwise, type of that word is found directly from wordNet. At this stage, background, action, noun situation are checked. For example, for the word play there are different type info to be obtained. It has the noun type (as game) or action (as to play). In this situation, it ends up with the type of majority (e.g. Five of them are noun and two of them are verbs then type is accepted as noun).

After all, strings acquired by the service for feedback are sent back to browser XUL window with Ajax for confirmation.

2.2.5 Website

According to the procedure described above, tag information from the end user is obtained. After this, one further is taken a web site as a sharing medium has been constructed. Web site has been implemented by C# and ASP.Net. Users sign in with their user name and password. After logging on, they can view what they tagged for what images at the web site. There is a search interface for their search at their bookmark page. It provides a search environment according to the tag info described. Search can be done either on user's tagged image set or entire system's bookmarked image set.

Search function can be done by entering words for selected fields (object, action, background). Search results do not only depend on the exact words as search criteria, they depend on the words similar to search

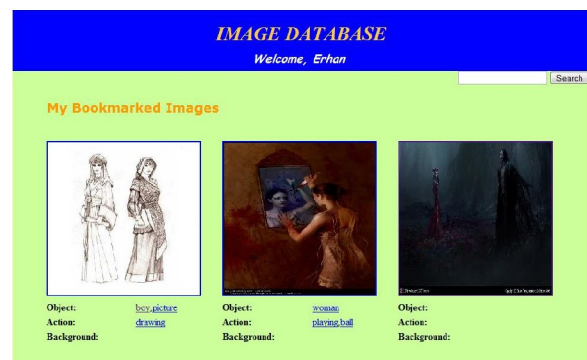


Figure 8: Web Site with Bookmarked Images and Tags.

criteria. Search operation depends on the similarity metric that will be discussed at the next section. Web site has been implemented by using Web user controls [11].

Web site contains the bookmarked pictures and their tag information below. Web site has been supported by the CSS (Cascading Style Sheets) for visual display. Tag information is composed of object (noun), action (verb) and background information of the image.

2.2.6 Image Similarity Search

At the web site user is capable of doing search in the local bookmark list as well as the entire site (across users). Image search aims at finding similar images according to the tag criteria specified at the search entry. In the search process, Jaccard similarity is applied. Given a set of words as tag information, we find the jaccard similarity between two images in terms of set of words as,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Where $A = \{w_{A1}, w_{A2}, w_{A3}, \dots, w_{An}\}$ and $B = \{w_{B1}, w_{B2}, w_{B3}, \dots, w_{Bm}\}$

Jaccard similarity will give back a similarity score between 0 and 1. Our Jaccard similarity is modified. It uses WordNet to find the number of synonym words. As the number of synonym words increase, the contribution of that word to similarity result decreases.

This method is the combination of WordNet and Jaccard similarity. For example given tags of image A and image B,

A = {baby, laughing}

B = {baby, sleeping, bed}

Table 1: Words and their # of synonym word.

Words	# of Synonym Words
baby	13
laughing	8
sleeping	11
bed	36

$A \cap B = \{baby\}$ and
 $A \cup B = \{baby, laughing, sleeping, bed\}$

$$J = \frac{baby}{baby + laughing + sleeping + bed}$$

Baby = 1/13, Laughing = 1/8, Sleeping = 1/11, Bed = 1/36

$$J = \frac{\frac{1}{13}}{\frac{1}{13} + \frac{1}{8} + \frac{1}{11} + \frac{1}{36}} = 0.56$$

Similarity score between images A and B is 0.56. In addition to image similarity, only words entered as search criteria can be used. For example, instead of all tags, baby or laughing could be entered. The picture below gives how the search criteria is entered.

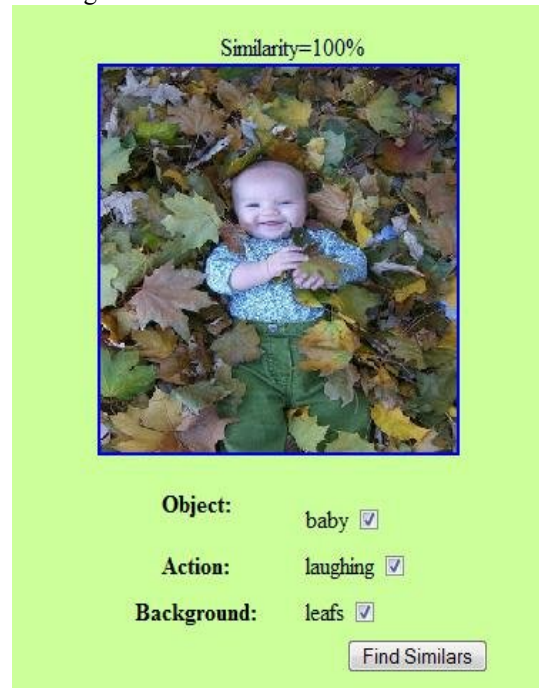


Figure 9: Image Search.

During the search, basically there is an extreme case such as finding the intersection of two sets as 0. For example rose and daisy can be grouped under flowers and hypernym of these words are used instead by taking the β (it is specified as 0.5) to the power of minimum distance they meet up to the tree.

Figure 10 gives the search results for laughing baby criteria entered. It is also possible to search from one of the results. Search results assist user and he can find more results by using one of the images as the query criteria.

3 Experiments

We performed experiments over 20 students and they collected images on specific topics given. Topics are



Figure 10: Search Results.

human beings, vehicles, sports, nature, smart phones, computers, house appliances, clothes, fashion. First, we allowed users find their 50 images per user for each category. They tagged their images and before doing our similarity search, we performed association rule mining [15] algorithm to find the patterns frequently approached. We kept the support value as 0.05, 0.01 and 0.15 and we considered frequent patterns over that value for images that are the same (after link information and fast detection of images of possible likeness). We have considered the union of patterns that are subsets of frequent patterns. Besides, images are checked if they are relevant to categories.

Normally in our system, similarity results less than a user specified threshold are ignored. (Default value is 0.1) but in experiments we adopted to idea of taking top n images from the search result set. We asked users to do five queries for each category and we checked the results if they are related or not.

While doing comparisons, we used f-measure that is dependent on precision and recall values [14].

$$\text{precision} = \frac{\text{no of correctly recommended images}}{\text{no of recommended images}} \quad (2)$$

$$\text{recall} = \frac{\text{no of correctly recommended images}}{\text{no of relevant images}} \quad (3)$$

$$f\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

We asked users to create search corpus as 50 words as search criteria from each user (30 objects, 20 verbs and 10 background). We constructed random search criteria by using our corpus for each category (15 queries for each) and we obtained the f-measure results below.

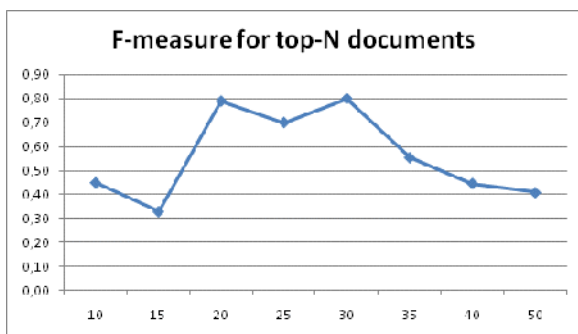


Figure 11: F-measure Value Results for Top-N Documents.

We restricted our queries to top-k queries in order to find the f-measure value. top 20 and top 30 query results in scoring return the highest scores around 80%. We got the lowest in top 15. As the top scoring result increases, f-measure starts to decrease. Our results are promising in terms of top 20 and 30 queries. Here, we also intend to limit the search records for time limitation.

4 Conclusion

In this study, we implemented a system that lets users to bookmark pictures they want to keep with their

tag info according to image content. Later on they may reorganize the tags and they may also do search by using tag info. Rather than image content, we tried to keep the tag information in order to do search and find similar images based on the tag field values entered.

In the future, we will also incorporate different utilities of Wordnet such as antonyms, and etc. Also, usage and rating of images will be engaged with heterogeneous scoring mechanism.

Acknowledgement

I would like to extend my special thanks to anonymous reviewers for their valuable comments and contributions.

References

- [1] John C. Paolillo and Shashikant Penumarthy(2007), The Social Structure of Tagging Internet Video on del.icio.us Proceedings of the 40th Hawaii International Conference on System Sciences.
- [2] Lee Rainie(2007), Interview: Author David Weinberger Describes How Tagging Changes People's Relationship to Information and Each Other Pew Internet and American Life Project January 31.
- [3] <http://del.icio.us/help/tags>
- [4] <http://www.xulplanet.com/tutorials/whyxul.html>
- [5] M. F. Porter(1980), An algorithm for suffix stripping no. 3, pp 130-137, July.
- [6] <http://wordnet.princeton.edu/>
- [7] <http://www.w3.org/TR/ws-arch/#whatis>
- [8] Nilayan Sharma(2005), Snappier Web Applications with XMLHttpRequest, April.
- [9] Lihui Lei; Zhenhua Duan(2007), "Integrating AJAX and Web Services for Cooperative Image Editing," IT Professional , vol.9, no.3, pp.25-29, May-June.
- [10] Che-Yu Yang; Hung, J.C.(2006); Word Sense Determination using WordNet and Sense Co-occurrence Advanced Information Networking and Applications.
- [11] Microsoft Developer Network, [http://msdn2.microsoft.com/enus/library/3457w616\(VS.71\).aspx](http://msdn2.microsoft.com/enus/library/3457w616(VS.71).aspx)
- [12] <http://www.w3.org/Style/CSS/>
- [13] Sudkamp, T.; Cross, V.(1996), Toward fuzzy-valued matching and similarity assessment, Fuzzy Information Processing Society, NAFIPS. Biennial Conference of the North American
- [14] Gerard Salton(1983), McGill, M., 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.
- [15] R. Aggrawal, T. Imielinski, and A. Swami(1993). Mining association rules between sets in large databases. In Proceedings of the ACM-SIGMOD International Conference on Management of Data. Washington, DC.
- [16] List of social bookmarking applications downloaded from: http://en.wikipedia.org/wiki/List_of_social_software#Social_bookmarking on Feb 20, 2009.

Dialog System for Open-Ended Conversation Using Web Documents

Masahiro Shibata

Faculty of Information Science and Electrical Engineering, Kyushu University, Japan
E-mail: shibata@lang.is.kyushu-u.ac.jp

Tomomi Nishiguchi

Toshiba Corporation Digital Media Network Company, Japan
E-mail: tomomi.nishiguchi@toshiba.co.jp

Yoichi Tomiura

Faculty of Information Science and Electrical Engineering, Kyushu University, Japan
E-mail: tom@is.kyushu-u.ac.jp

Keywords: dialog system, web corpus, surface cohesion, semantic coherence

Received: October 12, 2008

We have developed a new type of open-ended dialog system that generates proper responses to users' utterances using the abundant documents available on the World Wide Web as sources. Existing knowledge-based dialog systems provide meaningful information to users, but they are unsuitable for open-ended input. The system Eliza, while it can handle open-ended input, gives no meaningful information. Our system lies between the above two dialog systems; it can converse on various topics and gives meaningful information related to the user's utterances. The system selects an appropriate sentence as a response from documents gathered through the Web on the basis of surface cohesion and shallow semantic coherence. We developed a trial system to converse about movies and experimentally found that the proposed method generated appropriate responses at a rate of 66%.

Povzetek: Razvit je sistem dialoga z uporabo spletnih strani.

1 Introduction

We have developed a new type of open-ended dialog system that generates proper responses to users' utterances using the abundant documents available on the World Wide Web as sources.

Many practical knowledge-based dialog systems, such as telephone weather forecast [10] and online air travel planning systems [11], assume that users can form lingual expressions and make their requests clearly enough to be recognized. Under these conditions, such systems can determine a user's intention using methods like pattern-matching, because the user's aim is definite and the possible input is restricted; therefore, the systems can provide correct answers from prepared databases. This type of dialog system works well for specialized tasks, but it is unsuitable for open-ended input.

Chat systems for open-ended conversations also exist. Eliza [9] is a psychotherapeutic counselling system that converts user's utterances into interrogative sentences or makes non-substantive responses, such as "Really?", "I see", or "What happened?", to avoid answering the user clearly. Consequently, Eliza provides no meaningful information to users.

Previous studies have chiefly examined such dialog systems, that is, those that give users correct information in a narrow domain using structured knowledge and those that give users natural (but valueless) responses in

an unlimited domain. We have developed a new type of dialog system that lies between these two types [7]. Our system converses naturally in a broad domain and gives users information related to their interests. To generate the system's utterances, an appropriate response to a user's utterance is selected from documents gathered through the Web. The appropriateness of each selected response is estimated by its level of surface cohesion and shallow semantic coherence. Surface cohesion, which depends on centering theory, is the naturalness of the dialog on the surface. Semantic coherence, defined according to the degree of predictability of the response with words in the user's utterance and the informational value of the words in that response, is the relativity of the user's and system's utterances.

Most question answering (QA) systems, such as those described in Refs. [1] and [5], also extract the best answer in response to a user's natural language question from a large corpus on the basis of its similarity or relativity to the input question. However, QA systems differ from our system on the following points. First, QA systems mostly deal with questions that have correct answers, while our system generates responses to declarative sentences or interrogative sentences that do not necessarily have specific correct answers. Second, most QA systems depend on a named entity extraction

method [2] and aim to extract the appropriate words or phrases for the answers. In contrast, our system depends on a sentence extraction method that attempts to make appropriate responses to a user's utterances. Therefore, we must consider the naturalness of the dialog as well as its relativity to user's utterances in our system.

Here we report on our trial system for conversing about movies. Our dialog method does not use any domain-specific knowledge, but rather the domain of the corpus. Although the trial system is limited to movie-related topics, our dialog system can converse on various topics by gathering web pages about those topics.

We now describe the method used for selecting a proper response from candidates and experimentally evaluate our trial system.

2 Example dialog

We assume that user's and system's utterances are made alternately. Below is an actual sample dialog between the system (S) and a user (U).

- S_0 : 映画の話をしてしましよう。(Let's talk about movies.)
- U_1 : 亡国のイージスを見ました。(I saw "Aegis.")
- S_1 : へえー。(Ahh.)
- U_2 : 原作は「終戦のローレライ」で知られる福井晴敏さんの小説だそうですね。(I heard the original story of this movie was written by Harutoshi Fukui, who is famous for the novel "Lorelei, the Witch of the Pacific Ocean.")
- S_2 : 日本推理作家協会賞、日本冒険小説協会大賞、大藪春彦賞をトリプル受賞した、58万部を超えるベストセラー小説です。(It is a bestseller, selling over 580,000 copies, which won the Mystery Writers of Japan, Inc. Award, the Japanese Adventure Fiction Association Award, and the Haruhiko Oyabu Award.)
- U_3 : うらかぜがミサイルを迎撃するシーンの緊迫感にドキドキしました。(The tense scene where the Urakaze intercepted a missile was exciting.)
- S_3 : 亡国のイージスは、防衛庁も協力して撮影したようですが、当初は作品の内容から協力するかどうかもめたそうで、あの石破茂の英断で協力することになったようです。("Aegis" was shot with the cooperation of the Japan Defence Agency. At first, thinking about the story, they were hesitant to cooperate, but Shigeru Ishiba made the final decision to cooperate.)
- U_4 : ヨンファ役の中井貴一がはまり役でした。(Kiichi Nakai was a natural for the role of Yeonhwa.)
- S_4 : 揺れない強い意志の仙石を真田広之は好演してましたし、クールなヨンファを中井貴一が良く演じ、役者たちはそれぞれ見せていた。(The main actors played their roles well. Hiroyuki Sanada was good as Sengoku, who had a firm and strong will, and Kiichi Nakai was also

good in the role of the cool Yeonhwa.)

S_0 is the fixed phrase to start the conversation. Our system searches the corpus for the most appropriate sentence for responding to the user's utterance and uses it as the system's utterance. $S_2 - S_4$ are the system's utterances generated using selected sentences from the corpus. If the system fails to select an appropriate sentence, it generates a simple response, such as S_1 , to continue the conversation.

3 Method of proposed system

Our system uses a corpus of sentences gathered from web documents. When a user inputs an utterance, the system searches the corpus for the most appropriate sentence for responding to the user's utterance based on surface cohesion and shallow semantic coherence. Each user's utterance and each candidate sentence in the corpus is only analysed morphologically¹ because we could not prepare enough reliable data, such as case frames, to analyse sentences syntactically in open-ended conversation.

In this section, we describe the construction of the corpus from web documents and explain how surface cohesion and semantic coherence are determined by our system. We then explain how the system selects the most appropriate sentence from the corpus as its response.

3.1 Constructing the utterance candidate corpus

For the source of candidates for the system's utterances, we collected documents about movies to make an utterance candidate corpus for the current trials. We gathered web pages using the web search API developed by Yahoo! Developer Network² with the keywords "α映画(eiga)." α is any expression that is to be the main theme of a dialog. We selected a movie title as α in our trial system. "映画(eiga)" means "movie" in Japanese. This word was used to exclude pages irrelevant to movies from the search. However, the web pages returned by the search engine included documents without α as the main theme. Therefore we then prepared a simple title filter to remove such pages in the following manner. Because the title tag of an html page often expresses the main theme of the page, we kept web pages that included α in the title tag, i.e., "<title>...α...</title>," and removed those that did not. A set of utterance candidates, $\Gamma(\alpha)$, consists of all the sentences extracted from the web pages that passed through the above filter. Furthermore, we added information about the sentence number and the web page URL to each sentence in $\Gamma(\alpha)$. The utterance candidate corpus is $\cup_{\alpha} \Gamma(\alpha)$. In our trial system, this consisted of 2,580,602 sentences extracted from 44,643 documents.

¹ We use ChaSen (<http://chasen.naist.jp/hiki/ChaSen>) for analysing Japanese morphology.

² <http://developer.yahoo.co.jp>

3.2 Surface cohesion

Our system selects utterance candidates that maintain surface cohesion of the dialog. Centering theory [4] deals with the transition of the central concern of the dialog and has been applied to Japanese [8]. On the basis of that application, we regard the centralness of discourse entities in a Japanese sentence to be ranked as follows:

1. zero pronoun,³
2. noun phrase with postposition “は(wa),”
3. noun phrase with postposition “が(ga),” and
4. noun phrase with postposition “を(wo).”

We call these ranks centralness ranks. In Japanese, the case of a noun phrase is determined by the postposition appended to it. “が(ga)” indicates the subject case, “を(wo)” indicates the object case, and “は(wa)” indicates the topic of a sentence. The system requires a case frame dictionary to decide precisely whether an utterance has a zero pronoun or not. Therefore, we apply this simple rule: if a sentence does not have a noun phrase with postposition “は(wa)” or “が(ga),” we assume that a zero pronoun is the subject case. Furthermore, it is difficult to decide the antecedent of a pronoun without a case frame dictionary. We regard the antecedent as the noun phrase (including zero pronouns) that has the highest centralness rank in the previous utterance.

In this paper, we call the noun phrase with the highest centralness rank in a user's utterance (U) the focus. The focus (f) of a user's utterance U is decided on the basis of simplified centering theory:

- if U has a noun phrase NP with postposition “は(wa),” f is NP ,
- if U has a noun phrase NP with postposition “が(ga)” and does not have a noun phrase with postposition “は(wa),” f is NP , and
- if U has neither a noun phrase with postposition “は(wa)” nor a noun phrase with postposition “が(ga),” there is a zero pronoun and f is its antecedent. Centering theory says this antecedent is the entity that has the highest centralness rank in the system's utterance just before U . Therefore, f is the focus of the system's previous utterance.

If the system's utterance S just after U includes a topic, that is the same as the focus of U , then the topic transition between U and S is natural. Moreover, it is also natural in many cases, where S includes a topic that is the same as the main theme of the dialog.

3.3 Semantic coherence

We define the semantic coherence between utterances using content words (nouns, verbs, and adjectives) is defined as:

$$r(w, w') = \log P(w'|w) - \log P_D(w'), \quad (1)$$

where w is a content word in the previous U , and w' is a content word in a candidate for S . $P(w'|w)$ is the probability that a sentence includes w' when its preceding sentence includes w . $P_D(w')$ is defined as $df(w')/|D|$, where $df(w')$ is the number of web documents including w' in the corpus and D is all the web documents in the corpus. $-\log P(w'|w)$ refers to the conditional information of w' the occurrence of a sentence including just after a sentence including w . It represents the predictability from w to w' . When w' can be easily predictable from w , it becomes low. On the other hand, $-\log P_D(w')$ represents the information of w' . When w' only appears in a few documents, it becomes high. Thus, $r(w, w')$ is high if w' is easily predictable from w and if w' only appears in certain documents. In our trial system, $P(w'|w)$ is determined by maximum likelihood estimation using the utterance candidate corpus.

$r(w, w')$ is $-\infty$ when $P(w'|w)$ is zero. We will define the semantic coherence between utterances in Section 3.4.3 so that the existence of w' in a candidate sentence does not affect it when $r(w, w')$ is $-\infty$. $P(w'|w)$ is not confident when the frequency of w in the corpus is low. We also regard $r(w, w')$ as $-\infty$ when the frequency of w is lower than the threshold θ . In the trial system, θ is set to 5. The selection of θ will be investigated in future work.

3.4 Generating system's utterances

Our system generates utterances in the following manner.

3.4.1 Generating system's utterances

We assume that a conversation with our system is relatively short and that the main theme does not change throughout the dialog. Moreover, we assume that the main theme of the dialog has the highest centralness rank in a user's first utterance. Thus, when starting a conversation, our system selects the noun phrase with the highest centralness rank in the user's first utterance⁴ and sets the main dialog theme from this noun phrase. The system's utterance candidates are collected by their main themes, as described in Section 3.1. When α is the main dialog theme, there should be more chances to select appropriate sentences for the system's response from $\Gamma(\alpha)$ than from $\Gamma(\alpha')$; $\alpha \neq \alpha'$. Therefore, we restrict the system's utterance candidates to $\Gamma(\alpha)$.

³ Case elements are often omitted in Japanese. These invisible case elements are called zero pronouns.

⁴ We assume that there are no zero pronouns in this sentence.

3.4.2 Selecting sentences including the focus from $\Gamma(\alpha)$

Let U be a user's utterance, S be a candidate for the system's utterance just after U , and f be the focus of U . As described in Section 3.2, the topic transition between U and S is natural when S includes the same topic as f , and thus system tries to select such candidates.

We assume that the topic of a sentence is a word or a phrase with a high centralness rank. Therefore, our system selects sentences that have a zero pronoun with an f antecedent, or f with postposition “は(wa),” “が(ga),” or “を(wo)” from $\Gamma(\alpha)$. In actuality, considering the accuracy of an anaphoric analysis, the system regards the following type of sentences as sentences having a zero pronoun with an f antecedent: a series of at the most m ($=2$ in the trial system) sentences judged to have a zero pronoun in accordance with the method described in Section 3.2 just after a sentence having f with postposition “は(wa).”

3.4.3 Filtering and ranking by semantic coherence

The candidates that successfully pass through the step outlined in Section 3.4.2 have surface cohesion. Our system selects the sentence that has the highest semantic coherence from among these candidates as the system's final candidate.

The semantic coherence between content words is defined in Section 3.3. Let U be a user's utterance, S a candidate for the system's utterance just after U , $CW(U; f)$ a set of all the content words in U except for the focus f , and $CW(S; f)$ a set of all the content words in S except for f . We define the semantic coherence between U and S basically as the sum of the semantic coherences between the content words in U and S . However, not all content words in S have high semantic coherence with content words in U , even when S is an appropriate response for U . Therefore we restrict the sum to, at the most, K highest values.⁵ In addition, we have to take into account the possibility that $r(w, w')$ is $-\infty$. We, then, define the semantic coherence, $R(U, S; f)$, between U and S with the focus f as

$$R(U, S; f) = \underset{(w, w') \in CW(U; f) \times CW(S; f)}{FSUM(K)} r(w, w'), \quad (2)$$

where $FSUM(K)_{x \in X} g(x)$ is the sum of, at the most, K highest finite values of $g(x)$ ($x \in X$) when there is one or more finite values of $g(x)$ ($x \in X$) and is $-\infty$ when there are no finite values of $g(x)$ ($x \in X$).⁶ In calculating

⁵ In accordance with the results of a preliminary experiment, we set K as 3 in our system.

⁶ For instance, suppose that $g(1) = 5$, $g(2) = -1$, $g(3) = 2$, and $g(4) = g(5) = -\infty$. In this situation,

semantic coherence, we use content words but not the focus f because f has already been used in the filter described in Section 3.4.2.

Before ranking by semantic coherence, the system removes candidates that have fewer content words than K . Such candidates tend to have higher values of R when all candidates have negative values of R because $r(w, w')$ can be a negative finite value. However, such candidates tend to be meaningless as responses.

It is also possible that the semantic coherence between U and every candidate that pass through the filter described in Section 3.4.2 is low. In such case, we use the semantic coherence threshold.⁷ If no candidates that pass through the filter have a higher semantic coherence than the threshold, our trial system makes the judgment that there is no candidate that has sufficient semantic coherence.

3.4.4 Selecting sentences for system's utterances with the main theme as focus

As mentioned in Section 3.2, the topic transition between a user's utterance (U) and the system's utterance (S) just after U is natural if S includes the same topic as the main theme. In a case where no candidate has a semantic coherence (R) higher than the threshold described in Section 3.4.3, the system tries to generate an utterance including the same topic as the main theme α . That is, our system selects sentences that have a zero pronoun with an α antecedent, or α with postposition “は(wa),” “が(ga),” or “を(wo)” from $\Gamma(\alpha)$ and executes the selection described in Section 3.4.3. If no candidate has an R higher than the threshold, our system generates the fixed utterance “へえー (Ahh)” to continue the conversation.

3.4.5 Generating system's utterances

Our system basically outputs a sentence selected by the processes described in Sections 3.4.3 and 3.4.4 without change. However, if the system's utterance candidate was selected on the basis of the focus (f) of the user's utterance and f has no modifier in the candidate sentence, we can remove the noun phrase f in the candidate sentence to make it a zero pronoun sentence because its antecedent can be identified according to centering theory. This removal strengthens surface cohesion.

4 Examination

To evaluate the performance of our dialog system, we investigated the naturalness of the system's utterances given as responses to utterances made by some users who conversed with our trial system.

$$FSUM(2)_{x \in \{1, 2, 3, 4\}} g(x) = g(1) + g(3) = 7,$$

$$FSUM(2)_{x \in \{3, 4, 5\}} g(x) = g(3) = 2,$$

$$FSUM(2)_{x \in \{4, 5\}} g(x) = -\infty, \text{ and } FSUM(2)_{x \in \{1\}} g(x) = -\infty.$$

⁷ In accordance with the results of a preliminary experiment, the threshold was set to -2.5 .

Table 1 : Results of human evaluation of the naturalness of the system’s utterances generated by the proposed method.

User ID	(1)	(2)	(3)	(4)	(5)	(6)	Total (Ratio)
Level-3	14	34	60	42	36	25	211 (29%)
Level-2	48	28	39	59	58	44	276 (38%)
Level-1	64	50	21	20	35	56	246 (34%)

Table 2 : Evaluation results of the naturalness of the system’s utterances for each movie. (JP refers to a Japanese movie and KR refers to a Korean movie in α .)

Movie α	Level-3	Level-2	Level-1	Generation failed	$ \Gamma(\alpha) $
Densha Otoko (JP)	17	37	30	2	28,236
Charlie and the Chocolate Factory	25	45	16	3	21,282
Finding Neverland	24	29	30	5	18,274
Howl’s Moving Castle (JP)	29	33	26	2	16,874
Aegis (JP)	31	35	13	10	13,072
Chicago	25	29	26	5	8,030
Windstruck (KR)	21	18	16	34	7,514
Bridget Jones’s Diary	12	22	35	17	5,635
Giant	6	2	42	39	3,514
Deep Blue	21	26	12	30	2,662

4.1 Experimental methodology

We selected 10 movies, each title (α) having many utterance candidates ($\Gamma(\alpha)$), and asked 6 participants to watch them all.

We then asked them to converse with our dialog system, with the movie titles as the main dialog themes. We call a user’s utterance and the system’s response to it an utterance pair. Every person had three conversations for each α , with one conversation consisting of five utterance pairs, not including S_0 . Generally speaking, when a conversation becomes too long, the main theme may change; therefore, to avoid changes in the main theme, we limited each conversation to five utterance pairs. In this trial of conversations with our system, the 6 participants had 180 conversations, resulting in 900 utterance pairs.

We next asked the participants to grade the naturalness of the system’s utterance in each utterance pair into one of three levels:

- level-3: system’s utterance is natural as a response to the user’s utterance,
- level-2: system’s utterance is acceptable as a response to the user’s utterance,
- level-1: system’s utterance is unnatural as a response to the user’s utterance.

We investigated the performance of our dialog system using these human evaluations.

4.2 Experimental results

Of the 900 user’s utterances, inquiries requiring a specific correct answer were made 20 times. Getting correct answers to such inquiries, however, is not the aim of our system, because such answer can be obtained using other

dialog methods such as knowledge-based systems. Thus we excluded those utterance pairs in which the user’s utterance was such an inquiry, and evaluated the performance of the system using the remaining 880 utterance pairs.

The system’s utterance in 147 utterance pairs was “へえー(Ahh).” That is, our system failed to generate an appropriate response to 147 user’s utterances (17% of the 880 utterance pairs) by the proposed method.

Table 1 shows the results of the human evaluation of the 733 system’s utterances that were selected from the utterance candidate corpus and generated as the system’s responses. Each row corresponds to the naturalness levels and each column corresponds to each user (1) - (6) and the total utterances at each level. Two hundred and eleven system’s utterances (29% of the 733 system’s utterances) were natural; 276 system’s utterances (38%) were acceptable; and 246 system’s utterances (34%) were unnatural. If utterances evaluated as level-3 and level-2 (“natural” and “acceptable”) are regarded as appropriate responses to user’s utterances, then our system succeeded in generating appropriate responses 66% of the time.

Table 2 shows the evaluation results for each movie. We can see that the number of failures in selecting system’s utterances (i.e. the number of generated “へえー(Ahh)” responses) tends to be small when $|\Gamma(\alpha)|$ (the number of sentences in $\Gamma(\alpha)$) is large, as we expected. In contrast, the relation between $|\Gamma(\alpha)|$ and the number of natural or acceptable responses is not clear. For instance, “Densha Otoko” has a large number (28,236) of candidate sentences in the candidate utterance corpus; however, the system generated 30 unnatural utterances. Conversely, “Aegis” has a relatively small number (13,072) of candidate sentences, although the system generated only 13 unnatural utterances.

By investigating the system’s utterances evaluated as level-1, we found that insufficient filtering during

construction of the utterance candidate corpus may have caused the generation of the system's unnatural responses. For example, web pages gathered with the keyword “映画 ジャイアンツ” included not only documents about the movie “Giant,” but also a lot of documents about the Japanese professional baseball team, Yomiuri Giants.⁸ For some movie titles, the simple title filter we prepared was still insufficient for identifying the main theme of all gathered web documents. We must reconsider the filtering function when constructing the utterance candidate corpus.

5 Discussion: semantic coherence

This paper proposes a dialog strategy in which the system selects a sentence appropriate as the response to a user's utterance from the abundant available documents and generates it as the system's utterance. The semantic coherence described in Sections 3.3 and 3.4.3 is only a tentative definition: how to best define semantic coherence remains a matter for debate. In this section, we consider the reformation of semantic coherence in our study.

As Equation 1, we defined the semantic coherence between content words w and w' as the sum of the predictability ($\log P(w'|w)$) and the information of word w' ($-\log P_D(w')$). As Equation 2, we defined the semantic coherence between a user's utterance U and a candidate for the system's utterance S as the sum of the semantic coherences between content words in U and S . However, with regard to the appropriateness for the system's utterance, it is sufficient if at least a part of the content words in S can be predicted easily and a part includes a relatively large amount of information. A content word w' in S does not have to have high predictability and a large amount of information simultaneously. Furthermore, the statistical and information-related theoretic meaning of $R(U, S; f)$, defined as the sum of the K highest $r(w, w')$, is not clear. Therefore, we redefine a new semantic coherence ($R'(U, S; f)$) which is satisfied with the following properties.

- (a) $R'(U, S; f)$ is high when at least a part of the content words in S is associated strongly with U .
- (b) $R'(U, S; f)$ is high when at least a part of the content words in S contains a large amount of information.

We suppose $P(\{w'_1, w'_2, \dots, w'_k\} | U)$ to be the probability of a sentence occurring which contains whole w'_1, w'_2, \dots , and w'_k after U . (This is sufficient even if the sentence contains other words.) When

$\log P(\{w'_1, w'_2, \dots, w'_k\} | U)$ is close to zero, the content words w'_1, w'_2, \dots , and w'_k are associated strongly with U . In contrast, when this value is smaller, w'_1, w'_2, \dots , and w'_k become harder to be associated with U . We assume that w'_i occurs independently of other w'_j . Based on this assumption, $P(\{w'_1, w'_2, \dots, w'_k\} | U)$ can be approximated as follows:

$$P(\{w'_1, w'_2, \dots, w'_k\} | U) \cong \prod_{j=1}^k P(w'_j | U), \quad (3)$$

where $P(w'_j | U)$ is the probability that a sentence containing w'_j occurs after U . When $P(w'_j | U)$ is higher, w'_j is easily predictable from U .

When $P(w'_j | U)$ is zero, w'_j cannot be predicted from U .

We can assume that $P(w'_j | U)$ may depend on the combination of words in U , but huge quantities of training data are required to calculate the reliable estimations of the probabilities based on this assumption. Therefore, we suppose the following approximation:

$$P(w'_j | U) \cong \max_{w \in U} P(w'_j | w), \quad (4)$$

where $P(w'_j | w)$ is the probability that a sentence containing w'_j occurs after a sentence containing w , which is same as the definition given in Equation 1. This equation can be interpreted to mean that the occurrence of w'_j does not depend on the combination of some words, but rather only word w in U .

We substitute Equation 4 into Equation 3 and get the following equation:

$$\log P(\{w'_1, w'_2, \dots, w'_k\} | U) \cong \sum_{j=1}^k \log \max_{w \in U} P(w'_j | w).$$

As Assumption (a), all content words in S are not always predictive easily from U and the number of content words are different among candidate sentences. Therefore, we regard L content words as words contributing to the predictability as the same as $R(U, S; f)$. In addition, considering the case that S has fewer than L content words or the case that S has fewer than L content words which are predictable from U (i.e. $P(w'_j | U) > 0$), we define $R'_1(U, S; f)$, the predictability from U to S , as follows:

$$R'_1(U, S; f) = FSUM(L) \log \max_{w \in U} P(w' | w_i).$$

Unlike the definition of Equation 2 in Section 3.4.3, we consider the predictability from U 's focus f to content word w' in system's utterance candidate S .

The information of w'_j can be evaluated by $-\log P_D(w'_j)$, in the same was as in Equation 1 in Section 3.3. Therefore, considering Assumption (b), we

⁸ Both the movie title “Giant” and the baseball team “Giants” have the same spelling, “ジャイアンツ,” in Japanese.

regard the information that S contains as the sum of, at most, M highest $-\log P_D(w'_j)$, as follows:

$$R'_2(U, S; f) = FSUM(M)(-\log P_D(w')).$$

$w' \in CW(S; f)$

Obviously, future work must consider how to set the most appropriate L and M .

Finally, we can redefine the semantic coherence $R'(U, S; f)$ as the sum of $R'_1(U, S; f)$ and $R'_2(U, S; f)$ as follows:

$$R'(U, S; f) = R'_1(U, S; f) + R'_2(S; f).$$

As mentioned above, some reformations are being considered for semantic coherence, which needs to be more properly defined.

6 Discussion: semantic coherence

In this paper, we selected movies as dialog themes and developed a system for having idle conversations about those movies. However, the use of our system only in idle conversations does not fully show the system's usefulness. For example, one useful area of application of our method is the following. Suppose that a user is interested in something and wants to obtain some information about it; a situation where a user wants to consult on a matter but only has a vague idea about the topic and cannot think of proper keywords for successful searching on the Web. In these situations, our system may be able to provide some clues for searching. In the course of having a conversation with our system, the user has the chance to learn some useful terms or keywords from the system's utterances. Once the user gets these keywords, they can pursue their interest in the theme using information retrieval techniques such as keyword search or the QA method and obtain more detailed information. We may not be able to use our system for information retrieval alone, but it has the potential to attain more flexible dialog for information provision through combined use with other information retrieval techniques.

Other than movies, there are various other dialog domains around which we can develop our system in a similar manner, such as books, food, and baseball. For such domains, we can construct an utterance candidate corpus by the ad-hoc method described in this paper of gathering web documents, and we can relatively effortlessly select the most likely main themes. In contrast, there are also dialog domains for which it is difficult to determine the most appropriate main themes to use to construct the corpus. As for the step of searching for appropriate sentences from the corpus, there are also dialogs in which candidate documents may not be sufficiently narrowed down in terms of only the main theme, such as was shown above by the dialog about "Giant."

Therefore, instead of dividing of corpus sentences into $\Gamma(\alpha)$ by the main theme α , we would estimate the central topics of each document previously gathered by web crawlers and insert them into each document as document keywords. This can be accomplished by existing techniques of automatic keyword extraction [6] [3]. In addition, we can put several keywords into a document in place of a main theme. If a document has several keywords, the content of the document can then be represented by the combination of these keywords, which leads to a solution of synonym problem like that encountered with the word "Giant."

Further, to improve the system generality, we believe that we should abolish the main theme from our dialog strategy. Even in dialogs between humans, the main dialog theme is not often set before starting the conversation. The themes of dialogs are fixed, and also changed, over the course of conversation. The central topics of utterances (focuses) seem to decide the topic of the dialog. Therefore, we should preserve the focuses of some previous utterances and make set F of these focuses and the focus of the current user's utterance. Then, in the selection step of the system's utterance, we match F and keyword set K of each document and narrow down the sentence selection area to those documents matched with F . This will enable a narrowing down of the search area without setting the dialog main theme in advance.

As for matching F and K , we can simply check, for example, whether F and K have more than r common components or not. Of course, we can also suppose other ways of matching. Automatic keyword extraction generally uses a value representing the validity of the keyword (e.g., TF-IDF value). Therefore, each word in K seems to have its evaluated value as its weight. On the other hand, the words in F can be assigned weight depending on either their centralness or weight, reflected in the fact that the topics of utterances are gradually forgotten over time.

7 Conclusion

We explained a method of generating a natural response to a user's utterance in an open-ended conversation by retrieving an appropriate sentence from documents on the Web. Furthermore, we investigated the performance of our trial system using this method by having it actually converse with people. Our system could generate a natural response to a user's utterance 66% of the time. Finally, we discussed the redefinition of semantic coherence and instruction of document keywords as an extension of our method to better apply our system to other dialog domains.

Our trial system is only capable of making a idle conversation about movies. However, our approach of selecting the proper system's utterance from the corpus has potential to be usefulness in a number of engineering applications. For example, by combining other search techniques, our method could be used in an information retrieval system that converses naturally instead of functioning as a conversational Web search engine. Moreover, if the utterance candidate corpus includes

sentences extracted from blog pages about a certain product, we can expect that a person considering whether to purchase the product or not will be able to have a useful conversation with our dialog system.

To realize these applications, we must consider some extensions of our method, such as those discussed in Sections 5 and 6. Furthermore, future work must investigate the usefulness of our method with respect to the practical use of information provision.

References

- [1] T. Akiba, K.Itou, and A. Fujii (2004). Question answering using “common sense” and utility maximization principle. *In Working Notes of 4th NTCIR Workshop*, pp. 297-303.
- [2] A. Bhole. (2007). Extracting named entities and relating them over time based on Wikipedia. *Informatica*, vol. 30, no. 4, pp. 463-468.
- [3] J. Dobša and B. D. Bašić (2007). Approximate representation of textual documents in the concept space. *Informatica*, vol. 31, no. 1, pp. 21-22.
- [4] B. Grosz, A. Joshi, and S. Weinstein (1995). Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, vol. 21, no. 2, pp. 203-226.
- [5] S. Lee and G. G. Lee (2003). A question answering system for Japanese. *In Proceedings of the 3rd NTCIR Workshop*, pp. 31-38.
- [6] S. Sato, H. Hayashi, N. Maki, and M. Inoguchi (2007). Development of automatic accumulated newspaper articles on disasters. *In Proceedings of 2nd International Conference on Urban Disaster Reduction*, (CD-ROM).
- [7] M. Shibata, Y. Tomiura, H. Matsumoto, T. Nishiguchi, K. Yukino, and A. Hino (2006). Dialog system for new idea generation support. *In Proceedings of Computer Processing of Oriental Languages, 21st International Conference, ICCPOL2006*, pp. 490-497.
- [8] M. Walker, M. Iida, and S. Cote (1994). Japanese discourse and the process of centering. *Computational Linguistics*, vol. 20, no. 2, pp. 193-233.
- [9] J. Weizenbaum (1966). Eliza – a computer program for the study of natural language communication between man and machine. *Communication of the ACM*, vol. 9, no. 1, pp. 36-45.
- [10] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hanzen, and L. Hetherington (2000). Jupiter: A telephone-based conversational interface for weather information. *IEEE Trans. SAP*, vol. 8, no. 1, pp. 100-112.
- [11] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill (1994). Pegasus: A Spoken language interface for online air travel planning. *Speech Communication*, vol. 15, pp. 331-340.

Optimization of Actions in Activation Timed Influence Nets

M. Faraz Rafi, Abbas K. Zaidi and Alexander H. Levis
 System Architectures Laboratory, ECE Dept., George Mason University, Fairfax, VA USA
 E-mail: {mrafi1, szaidi2, alevi}@gmu.edu

P. Papantoni-Kazakos
 EE Dept., University of Colorado Denver, Denver, CO USA
 E-mail: Titsa.Papantoni@cudenver.edu

Keywords: influence net, activation timed influence net, Bayesian net

Received: October 12, 2008

A sequential evolution of actions, in conjunction with the preconditions of their environment and their effects, are all depicted by Activation Timed Influence Nets. In this paper, we develop two algorithms for the optimal selections of such actions, given a set of preconditions. A special case for the two algorithms is also considered where the selection of actions is further constrained by the use of dependencies among them. The two algorithms are based on two different optimization criteria: one maximizes the probability of a given set of target effects, while the other maximizes the average worth of the effects' vector.

Povzetek: Predstavljena sta dva algoritma za optimizacijo akcij v časovno odvisnih mrežah.

1 Introduction

We consider the scenario¹ where a sequence of actions needs to be initialized towards the materializing of some desirable effects. As depicted in Figure 1, each action is supported by a set of preconditions and gives rise to a set of effects; the latter become then the preconditions of the following action(s) which, in turn, gives rise to another set of effects. Such sequential evolution of actions is termed Activation Timed Influence Nets (ATINs), where the action performers may be humans. ATINs are an extension of an earlier formalism called Timed Influence Nets (TINs) [6-12, 20-27, 30, 31] that integrate the notions of time and uncertainty in a network model. The TINs are comprised of nodes that represent propositions (i.e., pre-and post-conditions of potential actions as well as assertions of events which may indirectly describe such actions), connected via causal links that represent relationships between the nodes, without any explicit representation of actions. TINs have been experimentally used in the area of Effects Based Operations (EBOs) for evaluating alternate courses of actions and their effectiveness to mission objectives in a variety of domains, e.g., war games [20-22, 25], and coalition peace operations [24, 27], to name a few. A number of analytical tools [6-12, 23, 24, 27, 30] have also been developed over the years for TIN models to help an analyst update conditions/assertions, represented as nodes in a TIN, to map a TIN model to a Time Sliced Bayesian Network for incorporating feedback evidence, to determine best set of pre-conditions for both timed and un-timed versions of Influence Nets, and to assess temporal aspects of the in-

fluences between nodes. A recent work [31] on TINs, underlying constructs and the computational algorithms, provides a comprehensive analytical underpinning of the modeling and analysis approach.

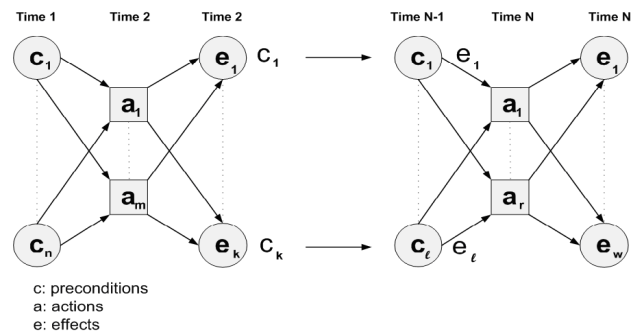


Figure 1: Network Representation of an Activation Timed Influence Net (ATIN)

In contrast to their predecessors (i.e., TINs), ATINs explicitly incorporate as nodes the mechanisms and/or actions that are responsible for changes in the state of a domain; other nodes represent preconditions and effects of actions. A set of preconditions may support a number of different actions, each of which may lead to the same effects, with different probabilities and different costs/awards, however. The objective is to select an optimal set of actions, where optimality is determined via a pre-selected performance criterion. In this paper, we present two algorithms which attain such an objective. We note that an effort to develop an action selection algorithm is also presented in [1].

The organization of the paper is as follows: In Section 2, we present the core formalization of the problem,

¹ This work was supported by the Air Force Office of Scientific Research (AFOSR) under Grants FA9550-05-1-0106 and FA9550-05-1-0388.

including two different optimization criteria. In Section 3, we derive the two algorithms which address the latter criteria. In Section 4, we express the extensions of the two algorithms to the network propagation scenario. In Section 5, we include numerical evaluations while in Section 6, we draw some conclusions.

1.1 Related Work

ATINs include action planning. In the domain of action planning, classical planners assume that the effects of an action are known with certainty and generate a set of actions that will achieve the desired goals [19]. Some planners do monitor for errors as actions are executed, but no action adaptations are incorporated [29]. Other planners assign probabilities to the effects of actions [2, 13, 14, 16, 28], but provide no mechanisms for reacting to changes in the environment. Reactive planners [5, 15, 17, 18] are designed to select and execute actions in response to the current state of the world, but, with a few exceptions [3], [4], they do not use probabilistic information to determine the likelihood of success of the actions. In [1], probabilistic information is used, in an effort to deal with environmental uncertainties, but no optimal action selection strategies are considered and/or proposed.

The ATIN formalism in this paper is similar to an earlier work by Sugato Baghci et al [1] on planning under uncertainty. The similarity, however, stops with the graph representation of preconditions, actions and their effects. Similar parallels can also be drawn with other graph-based planning approaches, e.g. GraphPlan (<http://www.cs.cmu.edu/~avrim/graphplan.html>). The approach in this paper represents a new formalism and is based on well established statistical results.

2 Problem formalization – core

In this section, we consider a modular core problem. We initially isolate a single action with its supporting preconditions and its resulting effects, as depicted in Fig. 2.

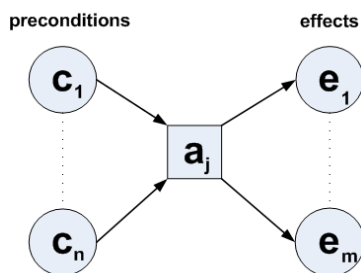


Figure 2: A Single Action ATIN

$X_1^n = [X_1, \dots, X_n]^T$ The status random vector of the preconditions, where $X_i = 1$, if precondition c_i is present and $X_i = 0$ if precondition c_i is absent. x_1^n denotes binary vector value realizations of X_1^n .

$Y_1^m = [Y_1, \dots, Y_m]^T$ The status random vector of the effects, where $Y_i = 1$, if effect e_i is present and $Y_i = 0$ if effect e_i is absent. y_1^m denotes binary vector value realizations of Y_1^m .

$p_j(x_1^n)$ The probability of success for action a_j , given that the value of the precondition status vector is x_1^n ;
 $P(\text{success for action } a_j | x_1^n)$

$q_j(y_1^m)$ The probability that the value of the effects' status vector is y_1^m , given that the action a_j is taken;
 $P(y_1^m | a_j \text{ taken})$

$q_0(y_1^m)$ The probability that the value of the effects' status vector is y_1^m , given that no action is taken;
 $P(y_1^m | \text{no action taken})$

$U_j(y_1^m)$ The utility of the value y_1^m of the effects' status vector, when action a_j is taken.

$U_0(y_1^m)$ The utility of the value y_1^m of the effects' status vector, when no action is taken.

We note that the utility function $U_j(y_1^m)$ measures the net worth of the effects' vector value y_1^m when action a_j is taken; thus, $U_j(y_1^m)$ is computed as the worth of y_1^m minus the cost of deployment for action a_j .

Let us now assume mutually exclusive actions, which are supported by the same preconditions, to lead to the same set of effects (as shown in Fig. 3). Let $\{a_j\}_{1 \leq j \leq k}$ be this set of actions and let X_1^n and Y_1^m denote the common status random vectors of preconditions versus effects, respectively. Let the utility functions for each action in the set $\{a_j\}_{1 \leq j \leq k}$ be nonnegative; let also $U_0(y_1^m)$ be nonnegative.

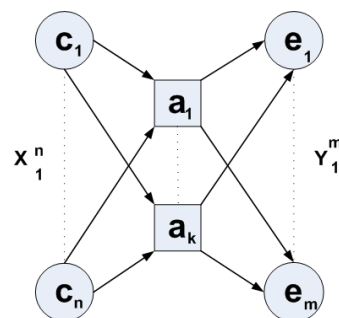


Figure 3: A Single Level ATIN

We now state multiple versions of the core problem, based on two different optimization criteria. Problem 3a and 3b are the constrained versions of the first two problems.

Problem 1 (Optimal Path Problem)

Given a preconditions vector value x_1^n , given an effects vector value y_1^m , find the maximum probability action that connects them. That is, find the action that maximizes the conditional probability $P(y_1^m | x_1^n)$.

Problem 2 (Average Utility Maximization)

Given a preconditions vector value x_1^n , find the action or actions that maximize the effects' average utility.

Problem 3a (Optimal Path Problem with Constrained Actions)

Given a preconditions vector value x_1^n , given an effects vector value y_1^m , and an action dependency matrix, find the maximum probability action that connects them. That is, find the action that maximizes the conditional probability $P(y_1^m | x_1^n)$. In this case, only those action combinations are considered that are allowed by the constraints in the dependency matrix.

Problem 3b (Average Utility Maximization with Constrained Actions)

Given a preconditions vector value x_1^n , find the action or actions that maximize the effects' average utility. As in Problem 3a, only those action combinations are considered that are allowed by the constraints in the dependency matrix.

Action Dependency Matrix (ADM)

An action dependency matrix is a tool which defines dependency among actions in the network. It reduces the number of combinations of actions by considering only those allowed by the dependency matrix. It also reduces significantly the amount of calculations required to obtain the optimal path. The value of the variable a_{ij} reflects the existence or absence of dependency between actions a_i and a_j , where a_{ij} equals 1; for positive dependency and equals 0; for negative dependency, and $1 \leq i, j \leq n$, where 'n' represents the total number of actions in the network. The elements of an ADM are determined as follows:

$$a_{ij} = \begin{cases} 1 & ; \text{ if action } a_i \text{ is selected for execution} \\ & \text{ in level } l, \text{ then } a_j \text{ refers to the action} \\ & \text{ that has to be or has been selected} \\ & \text{ for execution in level } k, \text{ where } l \neq k \\ 0 & ; \text{ if action } a_i \text{ is selected for execution} \\ & \text{ in level } l, \text{ then } a_j \text{ refers to the action} \\ & \text{ that must not be selected for execution} \\ & \text{ in level } k, \text{ where } l \neq k \end{cases}$$

where, level l in an ATIN corresponds to a set of preconditions ($C_1, C_2 \dots C_n$) followed by a set of actions ($a_1, a_2 \dots a_k$) and a set of effects ($e_1, e_2 \dots e_m$) (as shown in Fig. 3). The effects of this level then serve as the preconditions for the next level $l+1$ and so on.

3 Solutions to the core problems

We present the solutions to the two core problems posed in Section 2 in the form of a theorem, whose proof is in the Appendix.

Theorem 1

a. Given x_1^n , given y_1^m , and given a set of actions $\{a_j\}_{1 \leq j \leq k}$, the conditional probability $P(y_1^m | x_1^n)$ is maximized as follows:

by action a_{j^*} ; if

$$q_{j^*}(y_1^m) p_{j^*}(x_1^n) = \max_{1 \leq j \leq k} q_j(y_1^m) p_j(x_1^n) > q_0(y_1^m) \quad (1)$$

where then $\max P(y_1^m | x_1^n) = q_{j^*}(y_1^m) p_{j^*}(x_1^n)$

by no action; if

$$q_0(y_1^m) > \max_{1 \leq j \leq k} q_j(y_1^m) p_j(x_1^n) \quad (2)$$

where then $\max P(y_1^m | x_1^n) = q_0(y_1^m)$

If more than one action satisfy the maximum in (1), then one of these actions may be selected randomly.

b. Given x_1^n , given a set of actions $\{a_j\}_{1 \leq j \leq k}$, and given utility functions $\{U_j(y_1^m)\}_{1 \leq j \leq k}$ and $U_0(y_1^m)$, the average utility

$$\bar{U}(x_1^n) = \sum_{1 \leq j \leq k} \sum_{y_1^m} P(a_j \text{ taken} | x_1^n, y_1^m) \cdot U_j(y_1^m) + \sum_{y_1^m} P(\text{no action taken} | x_1^n) \cdot U_0(y_1^m)$$

is maximized as follows:

by action a_{j^*} ; if

$$A_{j^*}(x_1^n) \stackrel{\Delta}{=} p_{j^*}(x_1^n) \sum_{y_1^m} q_{j^*}(y_1^m) U_{j^*}(y_1^m) = \max_{1 \leq j \leq k} p_j(x_1^n) \sum_{y_1^m} q_j(y_1^m) U_j(y_1^m) > \sum_{y_1^m} q_0(y_1^m) U_0(y_1^m) \quad (3)$$

by no action; if

$$\sum_{y_1^m} q_0(y_1^m) U_0(y_1^m) > \max_{1 \leq j \leq k} p_j(x_1^n) \sum_{y_1^m} q_j(y_1^m) U_j(y_1^m) \quad (4)$$

$A_{j^*}(x_1^n)$ in (3) is the award assigned to action a_{j^*} ; it is also the worth assigned to the precondition vector value x_1^n by the action a_{j^*} .

If more than one action attain the maximum award $A_{j^*}(x_1^n)$ in (3), one of them is selected randomly.

4 Solutions of the network propagation problem

In this section, we generalize the core problem solutions expressed in Theorem 1, Section 3, to the sequence of actions depicted by the ATIN in Fig. 1.

Problem 1 (The Optimal Path Problem)

In the ATIN in Fig. 1, we fix the preconditions vector value $x_1^n(1)$, at time 1, and the effects' vector value $y_1^m(N)$, at time N. We then search for the sequence of actions that maximizes the probability $P(y_1^m(N)|x_1^n(1))$. The solution to this problem follows a dynamic programming approach where $x_1^n(l) = y_1^m(l-1); 2 \leq l \leq N$, in our notation. The proof of the step evolution is included in the Appendix.

Step 1

For each $y_1^m(1) = x_1^n(2)$ value, find

$$r(y_1^m(1)) \stackrel{\Delta}{=} \max [q_0(y_1^m(1)), \max_j q_j(y_1^m(1)) p_j(x_1^n(1))]$$

and the action index $j^*(y_1^m(1))$ that attains $r(y_1^m(1))$.

Step l

The values $r(y_1^m(l-1)) \stackrel{\Delta}{=} \max P(y_1^m(l-1)|x_1^n(l))$, for each $y_1^m(l-1)$ value, are in memory, as well as the actions that attain them. At step l , the values $r(y_1^m(l)) \stackrel{\Delta}{=} \max_{y_1^m(l-1)} r(y_1^m(l-1)) \times \max [q_0(y_1^m(l)), \max_j q_j(y_1^m(l)) p_j(y_1^m(l-1))]$

are maintained, as well as the sequence of actions leading to them.

The complexity of this problem is polynomial with respect to the number of links. Assume that a given ATIN model has 'N' number of levels and each level has 'k' links, then the complexity is given as $O(N \times k)$.

Problem 2 (The Average Utility Maximization)

In the ATIN in Fig. 1, we fix the value of the precondition vector at time 1, denoted $x_1^n(1)$. For each value $y_1^w(N)$ of the effects vector at time N, we assign worth functions $U(y_1^w(N))$. For each action $a_j(l)$, at time l , we assign a deployment cost $c_j(l)$. The utility of the effects' vector value $y_1^w(N)$, when action $a_j(N)$ is taken, is then equal to $U_j(y_1^w(N)) \stackrel{\Delta}{=} U(y_1^w(N)) - c_j(N)$, while the utility of the same value, when no action is taken, equals $U_0(y_1^w(N)) \stackrel{\Delta}{=} U(y_1^w(N))$. We are seeking the sequence of actions which lead to the maximization of the average utility. The evolving algorithm, from part (b) of Theorem 1, back propagates as follows. The proof is in the Appendix.

Step 1

Compute the action awards (including that to no action), with notation of Figure 1, as follows: $0 \leq j \leq r$;

$$A_j(x_1^l(N-1)) \stackrel{\Delta}{=} p_j(x_1^l(N-1)) \sum_{y_1^w(N)} q_j(y_1^w(N)) U_j(y_1^w(N))$$

with $p_0(x_1^l(N-1)) \stackrel{\Delta}{=} 1$

Select $A_{j^*(x_1^l(N-1))}(x_1^l(N-1)) \stackrel{\Delta}{=} \max_{0 \leq j \leq r} A_j(x_1^l(N-1))$;

for each $x_1^l(N-1)$ value.

Take action $a_{j^*(x_1^l(N-1))}(N)$ for preconditions vector value $x_1^l(N-1)$ and simultaneously assign worth $A_{j^*(x_1^l(N-1))}(x_1^l(N-1))$ to $x_1^l(N-1)$. That is, assign: $U(x_1^l(N-1)) = A_{j^*(x_1^l(N-1))}(x_1^l(N-1))$ (5)

Step 2

Back propagate to the preconditions at N-2, as in Step 1, starting with the worth assignments in (5), and subsequent utilizations

$$U_j(x_1^l(N-1)) = \max [A_{j^*(x_1^l(N-1))}(x_1^l(N-1)) - c_j(N-1), 0]$$

Step n

As in Steps 1 and 2 (for subsequent levels) the above described algorithm generates the optimal sequence of actions for given initial preconditions $x_1^n(1)$. The optimal such preconditions can be also found via maximization of the utility $U_j(x_1^k(2))$, with respect to $x_1^n(1)$.

The complexity of this problem is also polynomial with respect to the number of links.

Problems 3a, 3b (Optimization with Constrained Actions)

Problems 3a and 3b impose dependency constraints on the actions in the ATIN network. As explained in Section 2, an ADM defines the dependency of one action on every other one, where positive dependency is depicted by 1 and negative dependency is depicted by 0. The dependency constraints are taken into account, when, at a certain level, an optimal action is finalized. At any given level, only positively related actions are considered in the calculations.

As described in Step 1 of Problem 1 (see Section 4), for the first level, $r(y_1^m(1))$ is calculated the same way for constrained actions also. But for the rest of the levels, it is calculated in a different manner.

Consider,

$$r(y_1^m(l)) \stackrel{\Delta}{=} \max_{y_1^m(l-1)} r(y_1^m(l-1)) \times \max [q_0(y_1^m(l)), \max_j q_j(y_1^m(l)) p_j(y_1^m(l-1))]$$

The parameter $\max_{y_1^m(l-1)} r(y_1^m(l-1))$ corresponds to an ac-

tion selected for execution in level $l-1$. Its dependent actions can be known from the ADM. In this way, those combinations of actions which are not allowed by the ADM are eliminated from the calculation of $r(y_1^m(l))$, hence eliminating all links to and from the actions exhibiting negative dependencies. As a result of which it yields a network with lesser number of links and eases the determination of optimal sequence of actions.

5 Numerical evaluations

In this section, we focus on numerical scenarios. We first state the experimental setup. We then, evaluate and discuss a specific experimental scenario. We only state the experimental setups for Problems 1 and 2, since those of Problems 3a and 3b are straight forward modifications of the former.

5.1 Experimental Setups

Experimental Setup for Problem 1

Assign the probabilities

$\{q_j(x_1^k(l))\}$ and $\{p_j(x_1^k(l))\}$ as in problem 2.

Given these probabilities:

a. Compute first:

$$r(y_1^m(1)) \stackrel{\Delta}{=} \max_{y_1^m(1)} [q_0(y_1^m(1)), \max_j q_j(y_1^m(1)) p_j(x_1^n(1))]$$

and the action $j^*(y_1^m(1))$ that attains $r(y_1^m(1))$.

b. For each $l: 2 \leq l \leq N$, maintain in memory the values

$$r(y_1^m(l-1)) \stackrel{\Delta}{=} \max P(y_1^m(l-1) | x_1^n(l)), \text{ for each } y_1^m(l-1) \text{ value, and the actions that attain them. Then, compute and maintain the values:}$$

$$r(y_1^m(l)) \stackrel{\Delta}{=} \max_{y_1^m(l-1)} r(y_1^m(l-1)) \times \max_j [q_0(y_1^m(l)), \max_j q_j(y_1^m(l)) p_j(y_1^m(l-1))]$$

Also, maintain the actions that attain the values $r(y_1^m(l))$.

Experimental Setup for Problem 2

Considering the network in Fig. 1, assign:

a. Worth function $U(y_1^w(N))$ for all $y_1^w(N)$ values of the effects' status vector, at level N .

b. Probabilities $q_j(x_1^k(l)) \stackrel{\Delta}{=} P(x_1^k(l) \text{ occurring} | \text{action } j \text{ at step } l-1) \text{ at all levels, } 2 \text{ to } N,$

where $q_0(x_1^k(l)) \stackrel{\Delta}{=} P(x_1^k(l) \text{ occurring} | \text{no action } j \text{ at step } l-1) \text{ at all levels, } 2 \text{ to } N,$

c. Probabilities $p_j(x_1^k(l)) \stackrel{\Delta}{=} P(\text{action } j \text{ succeeds} | x_1^k(l) \text{ preconditions}) \text{ at all levels, from } 1 \text{ to } N-1,$

where $p_0(x_1^k(l)) \stackrel{\Delta}{=} 1; \forall l$

d. Implementation/deployment costs $c_j(l)$ for all actions, at all levels 2 to N .

Given the above assignments,

a. Compute first,

$$A_j(x_1^l(N-1)) \stackrel{\Delta}{=} p_j(x_1^l(N-1)) \sum_{y_1^w(N)} q_j(y_1^w(N)) U_j(y_1^w(N))$$

where,

$$p_0(x_1^l(N-1)) \stackrel{\Delta}{=} 1;$$

$$U_j(y_1^w(N)) = \max [U(y_1^w(N)) - c_j(N), 0]$$

$$A_{j^*(x_1^l(N-1))}(x_1^l(N-1)) \stackrel{\Delta}{=} \max_{0 \leq j \leq r} A_j(x_1^l(N-1)) ;$$

for all $x_1^l(N-1)$ values.

b. Take action $a_{j^*(x_1^l(N-1))}$ for each precondition vector value $x_1^l(N-1)$.

Assign worth $A_{j^*(x_1^l(N-1))}(x_1^l(N-1))$ to $x_1^l(N-1)$, as

$$U(x_1^l(N-1)) = A_{j^*(x_1^l(N-1))}(x_1^l(N-1))$$

Repeat steps (a) and (b) for level $N-1$ and back propagate to level $N-2$. Continue back propagation to level 1.

5.2 A Specific Experimental Scenario

In this section, we illustrate the use of Activation Timed Influence Nets with the help of an example ATIN, and present the results of the algorithms included in this paper, when applied to this ATIN. The model used in this section was derived from a Timed Influence Net presented in Wagenhals et al., in 2001 [27] (which was developed with the help of a team of subject matter experts) to address the internal political instabilities in Indonesia in the context of East Timor. For purposes of results illustration, we have selected a part of this network, as shown in Fig. 4.

Example ATIN:

The model provides detailed information about the religious, ethnic, governmental and non-governmental organizations of Indonesia. In this section, the propositions and actions referred are given in *italic* text. According to the model, rebel militia formed by a minority group poses the main concern which has captured a large number of people under its secured territory. Amongst these people in the community, some are against the rebels and considered to be at risk, in case the negotiations with the local government didn't work. For this example,

consider the initial conditions when *the rebels are getting local support*, the *community is in unrest* and the *local administration is losing control*. Based on the data provided, only one action can be executed from a possible set of actions at a given time i.e. either of the *Indonesian press* or *provincial authority* or the *minister of interior would declare resolve to keep peace*. Depending upon this selected action and the data provided for the effects, only a specific set of events can result. For instance, *rebels may or may not start thinking that they are getting publicity*, *GOI (original anti-government of Indonesia) war may or may not expand*, *GOI chances of intervention and international attention may increase or decrease*. Similarly, this specific set of events forms the set of possible pre-conditions for a later time. Depending upon which conditions actually become true, second action can be selected for execution from another set of actions, i.e. *Security Council* and *General Assembly* may or may not pass resolutions or *UN may or may not declare resolve to keep peace*. Depending upon this action and the data provided for the effects, *coalition may or may not form*, *rebels may or may not contemplate talks*, *GOI support may increase or decrease or may not increase at all*, or *GOI may or may not allow coalition into territories*. Ultimately, the *coalition may authorize use of force* which might compel *rebels to negotiate* and the *humanitarian assistance (HA) may start preparing for the worst case*. Depending upon which conditions meet, the *coalition may declare resolve to keep peace* or *may declare war on rebels*. This may affect the *chances of military confrontation*, *rebels' popularity* and *chances of negotiated settlement* which represents the final effects in the network.

Table 1 lists some of the parameters (and their values) required by the network in Fig. 4. The parameters in the table are listed by their abbreviated labels also in addition to the phrases shown inside the network nodes in the figure. For the sake of brevity, we do not list all the values.

Solutions to Problems:

Solution to Problem 1 (Optimal Path Problem):

Consider the example scenario described earlier, we need

to identify an optimal path (i.e., the sequence of actions) resulting into the final effect when, *military confrontation chances are reduced*, while *rebels start losing local support* and *negotiation chances start increasing*. This set of effects (post-conditions) leads to the following output state in the ATIN model:

- *Reduction in the chances of military confrontation* (i.e. $Y_{12} = 0$)
- *Decrease in local support and popularity for Rebels* (i.e. $Y_{13} = 1$)
- *Increase in chances of negotiated settlement* (i.e. $Y_{14} = 1$).

The above defined conditions lead to a post-condition vector $[0, 1, 1]^T$ at level 4, i.e. $y_{12}^{14}(4)$.

After fixing the post-condition vector, we define the initial preconditions, when *rebels have been getting local support*, the *community has been in unrest* and the *local administration has started losing control*. This set of pre conditions given by $x_1^3(1)$ results into a vector value of $[1, 1, 1]^T$, where

- $X_1 = 1$; represents the condition *Rebels are getting Local Support*
- $X_2 = 1$; represents the condition *There is unrest in the Community*
- $X_3 = 1$; represents the condition *Local Administration is losing Local Control*.

We want to find out the sequence of actions which achieves the desired effects $y_{12}^{14}(4)$ given the initial preconditions $x_1^3(1)$. Technically, we want to identify the sequence of actions which maximizes the probability $P(y_{12}^{14}(4) | x_1^3(1))$. Applying the optimal path algorithm (see Section 4) results that if the *provincial authority* and *UN declare resolve to keep peace* and *coalition does not take any action*, instead it *declares resolve to keep peace*, then the desired effects will be achieved which will result into *less chances of military confrontation*, *reduction in local support for rebels* and *more chances of a negotiated settlement*.

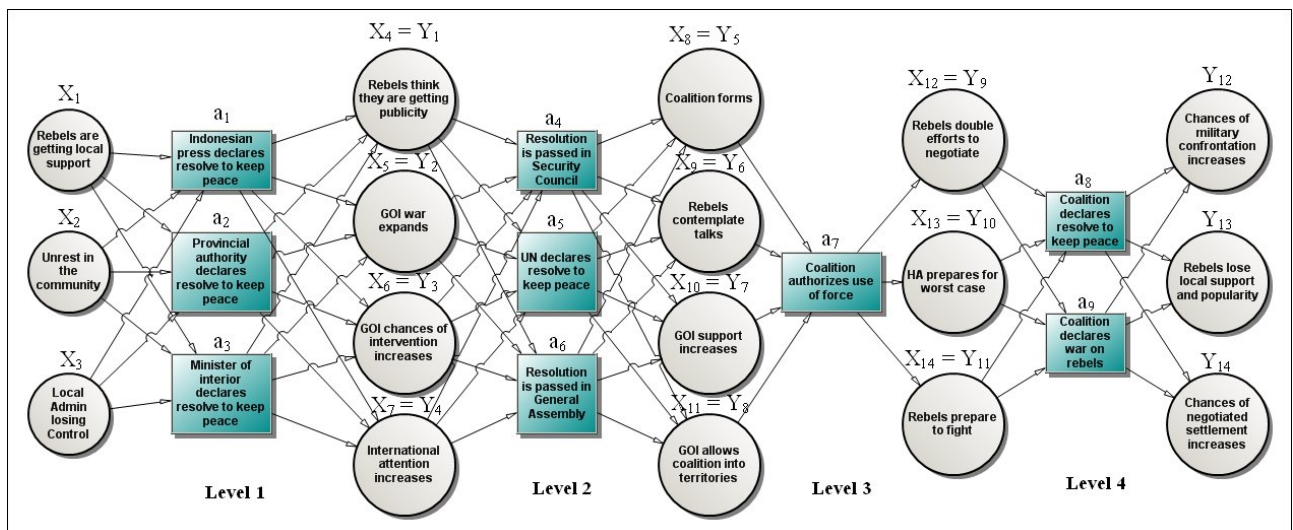


Figure 4: Example ATIN.

Table 1: Parameter values in the Example ATIN

Level 1							
Y_1^4	Action a_j	$q_j(y_1^4)$	$q_0(y_1^4)$	X_1^3	$p_j(x_1^3)$	$p_j(x_1^3) * q_j(y_1^4)$	$r(y_1^4)$
[1,0,1,1] ^T	Indonesian press declares resolve to keep peace (a1)	68.00%	6.45%	[1,1,1] ^T	80.00%	54.40%	63.64%
	Provincial Authority declares resolve to keep peace (a2)	74.00%			86.00%	63.64%	
	Minister of interior declares resolve to keep peace (a3)	56.00%			20.00%	11.20%	
Level 2							
Y_5^8	Action a_j	$q_j(y_5^8)$	$q_0(y_5^8)$	X_4^7	$p_j(x_4^7)$	$p_j(x_4^7) * q_j(y_5^8)$	$r(y_5^8)$
[1,1,1,1] ^T	Resolution is passed in Security council (a4)	14.00%	0.95%	[0,0,0,0] ^T	16.00%	2.24%	34.02%
				
				[1,0,1,1] ^T	91.00%	12.74%	
			
	[1,1,1,1] ^T	18.00%		2.52%			
	UN declares resolve to keep peace (a5)	42.00%		[0,0,0,0] ^T	66.00%	27.72%	
				
				[1,0,1,1] ^T	81.00%	34.02%	
			
	[1,1,1,1] ^T	13.41%		5.63%			
	Resolution is passed in General Assembly (a6)	39.00%		[0,0,0,0] ^T	43.05%	16.79%	
				
[1,0,1,1] ^T			48.20%	18.80%			
...					
[1,1,1,1] ^T	24.87%	9.70%					
Level 3							
Y_9^{11}	Action a_j	$q_j(y_9^{11})$	$q_0(y_9^{11})$	X_8^{11}	$p_j(x_8^{11})$	$p_j(x_8^{11}) * q_j(y_9^{11})$	$r(y_9^{11})$
[1,0,0] ^T	Coalition Authorizes the Use of Force (a7)	43.00%	59.00% (No Action)	[0,0,0,0] ^T	0.00%	0.00%	59.00%
				
				[1,0,1,1] ^T	64.00%	27.52%	
				
[1,1,1,1] ^T	18.00%	7.74%					
Level 4							
Y_{12}^{14}	Action a_j	$q_j(y_{12}^{14})$	$q_0(y_{12}^{14})$	X_{12}^{14}	$p_j(x_{12}^{14})$	$p_j(x_{12}^{14}) * q_j(y_{12}^{14})$	$r(y_{12}^{14})$
[0,1,1] ^T	Coalition declares resolve to keep peace (a8)	21.00%	1.50%	[0,0,0] ^T	16.00%	3.36%	19.11%
				
				[1,0,0] ^T	91.00%	19.11%	
			
	[1,1,1] ^T	38.00%		7.98%			
	Coalition declares war on rebels (a9)	17.00%		[0,0,0] ^T	67.00%	11.39%	
...					
[1,0,0] ^T			18.48%	3.14%			
...					
[1,1,1] ^T	30.88%	5.25%					

The details of this result are given in Table 1. It only contains the values that correspond to the selected actions at their respective levels, while a complete set of probabilities has been used to calculate the actual final sequence. The optimal actions, their corresponding state vectors and the probabilities are underlined in the table. The Optimal Path algorithm is of dynamic programming nature, so it requires two traversals to finalize the sequence of actions. During the forward traversal, $r(y_1^m)$ is calculated for each level for all possible post-condition combinations. At the last level, the post-condition vector $y_{12}^{14}(4)$ is fixed to be the desired effect of the network which is $[0, 1, 1]^T$ as determined earlier. The best action associated with this post-condition vector is identified along with its pre-condition vector $x_{12}^{14}(4)$. Using this pre-condition vector (which is the post-condition vector of the second last level), the network is traversed in reverse direction identifying actions and their corresponding preconditions, from last to the first level. The action at the first level is identified by fixing the pre-condition to the value determined earlier, i.e. $x_1^3(1)$ which is $[1, 1, 1]^T$. Completing both forward and reverse traversals gives the optimal actions which achieve the desired effects when the initial causes are given.

Solution to Problem 2 (Average Utility Maximization):

Consider a scenario where we need to identify the sequence of actions which maximizes the effects' average utility (at level 4) for the same input pre-condition as it was used in the solution of Problem 1, i.e. $[1, 1, 1]^T$. Assume, that the deployment costs for actions a_8 and a_9 are 25 and 30 units, respectively. The worth of each effect in the last level (i.e. level 4) is given by the worth function values $U(y_{12}^{14}(4))$ given in Table 2 and 3. Each effect also has a net utility which is determined by subtracting the deployment cost of the action from the worth of the effect. This net utility $U_j(y_{12}^{14}(4))$ (when action a_j is taken) and the action awards are given in Tables 2 and 3. The action award is calculated for each action corresponding to all of its pre-conditions. Similarly, these calculations are performed for the rest of the actions in ATIN model (after costs are assigned to every action in the model), but for the sake of brevity only the results for actions a_8 and a_9 are shown in Tables 2 and 3, respectively.

As described in Section 4, the action award is calculated for all actions in each level. For instance, starting from the last level, the action awards are calculated for actions a_8 and a_9 . The selected action is the one which maximizes the average utility and its action index 'j' is recorded. As each action award is calculated, it is also assigned as the worth function to the previous level effects vector. The latter worth function is used to calculate the utilities at the previous level, and calculations are repeated similarly. This procedure is back traversed from last to first levels. Table 4 summarizes the action awards of those actions which maximize the effects' average utility at their respective levels.

Table 2: Utility Functions and Action awards for Action a_8

Level 4 – Action a_8						
$X_{12}^{14}(4)$	$p_8(x_{12}^{14}(4))$	$Y_{12}^{14}(4)$	$q_8(y_{12}^{14}(4))$	$U(y_{12}^{14}(4))$	$U_8(y_{12}^{14}(4))$	$A_8(x_{12}^{14}(4))$
$[0,0,0]^T$	16.00%	$[0,0,0]^T$	37.00%	40	15	11.11
$[0,0,1]^T$	24.00%	$[0,0,1]^T$	65.00%	30	5	16.66
$[0,1,0]^T$	75.00%	$[0,1,0]^T$	53.00%	60	35	52.07
$[0,1,1]^T$	85.00%	$[0,1,1]^T$	21.00%	79	54	59.02
$[1,0,0]^T$	91.00%	$[1,0,0]^T$	19.00%	41	16	11.11
$[1,0,1]^T$	72.00%	$[1,0,1]^T$	43.00%	65	40	49.99
$[1,1,0]^T$	16.00%	$[1,1,0]^T$	29.00%	37	12	63.18
$[1,1,1]^T$	38.00%	$[1,1,1]^T$	27.00%	51	26	26.38

Table 3: Utility Functions and Action awards for Action a_9 .

Level 4 – Action a_9						
$X_{12}^{14}(4)$	$p_9(x_{12}^{14}(4))$	$Y_{12}^{14}(4)$	$q_9(y_{12}^{14}(4))$	$U(y_{12}^{14}(4))$	$U_9(y_{12}^{14}(4))$	$A_9(x_{12}^{14}(4))$
$[0,0,0]^T$	67.00%	$[0,0,0]^T$	41.00%	40	10	48.25
$[0,0,1]^T$	97.15%	$[0,0,1]^T$	26.00%	<u>30</u>	<u>0</u>	<u>69.96</u>
$[0,1,0]^T$	58.29%	$[0,1,0]^T$	71.00%	60	30	41.97
$[0,1,1]^T$	13.00%	$[0,1,1]^T$	17.00%	79	49	9.36
$[1,0,0]^T$	18.48%	$[1,0,0]^T$	26.00%	41	11	13.31
$[1,0,1]^T$	39.28%	$[1,0,1]^T$	54.00%	65	35	28.29
$[1,1,0]^T$	38.67%	$[1,1,0]^T$	62.00%	37	7	27.85
$[1,1,1]^T$	30.88%	$[1,1,1]^T$	58.00%	51	21	22.24

From Table 4 it can be seen that the sequence of actions that maximizes the effects' average utility, obtained as a result of applying the algorithm is given by: a_1 (i.e. *Indonesian press declares resolve to keep peace*), a_6 (i.e. *Resolution is passed in General Assembly*), a_7 (i.e. *Coalition authorizes use of Force*), a_9 (i.e., *Coalition declares war on rebels*). The underlined entries in Table 3 correspond to the worth, utility function and action award of action a_9 .

Solution to Problem 3a, 3b (Constrained Actions):

The dependencies among the actions in the example ATIN model are defined in the action dependency matrix given in Figure 5.

Most of the dependencies given in the matrix are quite evident. For instance, *the peace resolution declaration by UN (a5)* ensures that either of *Indonesian press, provincial authority or minister of interior must also have declared the resolution to keep peace* (either of a_1 or a_2 or a_3 must have been executed in the past) which would represent the opinion of the locals in general. Similarly, *resolution passed by the Security Council or General Assembly (a4 or a6)* makes sure that *whether or not the coalition will have to authorize the use of force (a7)*, considering the resolution is in support of use of force. This infers that if *the coalition authorizes the use of force*, it will *declare war on Rebels* otherwise, it will *declare resolve to keep peace*. All of these dependencies can be observed from the ADM (as shown in Fig. 5).

Consider a_{25} in ADM, (as shown in Fig. 5) which

corresponds to a positive dependency between *peace declaration by the provincial authority* (a_2) and *peace declaration by UN* (a_5). The ADM suggests that there exist negative dependencies between action a_2 and actions a_4 , a_6 , a_7 and a_9 which means that if *Provincial authority declares peace resolution*, *Security Council* and *General Assembly won't pass resolution* and the *Coalition will not authorize the use of force* and hence will *declare resolve to keep peace*. This knowledge of dependencies from the ADM certainly reduces an extensive amount of effort in calculating the optimal path. While calculating the optimal path, during the forward traversal, only those paths are considered which satisfy the constraints defined in ADM yielding less number of combinations to consider for calculation and making it easy to back traverse and identify the optimal actions.

The same applies to the solution of the second problem of identifying sequence of actions maximizing the effects' average utility under constraints. The action awards are calculated for those actions only which satisfy constraints defined in ADM, and hence reducing the effort of calculating action awards and assignment of worth function at each level.

Table 4: Action Awards.

Level 1	Level 2	Level 3	Level 4
$A_1(x_1^3(1))$	$A_6(x_4^7(2))$	$A_7(x_8^{11}(3))$	$A_9(x_{12}^{14}(4))$
151.02	85.18	77.92	69.96

6 Conclusion

This paper presented an extension of a Timed Influence Net, termed ATIN (Activation Timed Influence Net). An ATIN utilizes a set of preconditions required for the undertaking of an action and produces a set of effects. These effects become then the preconditions for the next level of action(s), resulting in a sequential evolution of actions. Some other probabilistic planning techniques were also discussed. The paper identified several pre-

selected performance criteria regarding ATINs (i.e., optimal path and average utility maximization with and without constrained actions) and recommended algorithms for their satisfaction. A tool called ADM (Action Dependency Matrix) was introduced, which induces dependencies among the actions. It is represented with the help of a $m \times m$ matrix, where ' m ' represents the total number of actions in the network.

The implementation of the suggested algorithms was illustrated with the help of a real world example. The example demonstrated a politically unstable situation in Indonesia. Sets of actions preceded by preconditions and followed by sets of effects were demonstrated in the form of an ATIN Model (see Figure 4). The experiment was formulated based on a previous Timed Influence Network model for the same scenario. The experimental procedure was applied to the network with a set of probability data. Solutions of both problems were discussed in depth. The optimal path problem required the knowledge of an initial set of causes (preconditions) and the final set of effects (postconditions). With the help of the algorithm, an optimal sequence of actions was identified which maximized the conditional probability of achieving the desired effects, when the initial conditions were given. For the sake of brevity, only significant parts of the probability data used were shown in Table 1. For the same scenario, the second algorithm yielded a sequence of actions, which maximized the effects' average utility. The solution for both problems was comprehended in detail. The experiment was repeated with constrained actions considering only dependent actions as defined in the Action Dependency Matrix (see Figure 5) which produced similar results and required lesser effort to calculate than without ADM.

References

[1] Bagchi S., Biswas G., and Kawamura K., (Nov, 2000), "Task Planning under Uncertainty using a Spreading Activation Network", IEEE Transactions

	a1 Indonesian Press declares resolve to keep peace	a2 Provincial Authority Press declares resolve to keep peace	a3 Minister of Interior Press declares resolve to keep peace	a4 Resolution is passed in Security council	a5 UN Declares resolve to keep peace	a6 Resolution is passed in General Assembly	a7 Coalition authorizes use of Force	a8 Coalition declares resolve to keep peace	a9 Coalition declares war on Rebels
a1 Indonesian Press declares resolve to keep peace	1	0	0	0	1	0	0	1	0
a2 Provincial Authority Press declares resolve to keep peace	0	1	0	0	1	0	0	1	0
a3 Minister of Interior Press declares resolve to keep peace	0	0	1	0	1	0	0	1	0
a4 Resolution is passed in Security council	0	0	0	1	0	0	1	0	1
a5 UN Declares resolve to keep peace	1	1	1	0	1	0	0	1	0
a6 Resolution is passed in General Assembly	0	0	0	0	0	1	1	0	1
a7 Coalition authorizes use of Force	0	0	0	1	0	1	1	0	1
a8 Coalition declares resolve to keep peace	1	1	1	0	1	0	0	1	0
a9 Coalition declares war on Rebels	0	0	0	1	0	1	1	0	1

Figure 5: Action Dependency Matrix

- on Systems, Man and Cybernetics, Vol. 30, No. 6, pp 639-650.
- [2] Dean T. and Wellman (1991) M.P., Planning and Control. San Mateo, CA: Morgan Kaufmann.
 - [3] Dean T., Kaelbling L.P., Kirman J., and Nicholson A., (1993), "Planning with deadlines in stochastic domains," in Proc. 11th Nat. Conf. Artificial Intelligence (AAAI-93). Chicago, IL: AAAI Press, pp. 574–579.
 - [4] Drummond M. and Bresina J., (1990), "Anytime synthetic projection: Maximizing the probability of goal satisfaction," in Proc. 8th Nat. Conf. Artificial Intelligence (AAAI-90). Chicago, IL: AAAI Press, pp.138–144.
 - [5] Firby R.J., (1987), "An investigation into reactive planning in complex domains,"in Proc. 6th Nat. Conf. Artificial Intelligence (AAAI-87). San Mateo, CA: Morgan Kaufmann, vol. 1, pp. 202–206.
 - [6] Haider S., (October, 2003), "On Computing Marginal Probability Intervals in Probabilistic Inference Networks," in Proceedings of IEEE-SMC Conference.
 - [7] Haider S. (August, 2003), "A Hybrid Approach for Learning Parameters of Probabilistic Networks from Incomplete Databases," Design and Application of Hybrid Intelligent Systems, Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03), IOS Press, Amsterdam, The Netherlands, ISBN 158603-394.
 - [8] Haider S., Zaidi A. K., and Levis A. H. (Nov. 2004), "A Heuristic Approach for Best Set of Actions Determination in Influence Nets," in Proc. IEEE International Conference on Information Reuse and Integration, Las Vegas.
 - [9] Haider S. and Levis A. H. (2004), "An Approximation Technique for Belief Revision in Timed Influence Nets," in Proceedings of Command and Control Research and Technology Symposium.
 - [10] Haider S. and Zaidi A. K. (2004), "Transforming Timed Influence Nets into Time Sliced Bayesian Networks," in Proceedings of Command and Control Research and Technology Symposium.
 - [11] Haider S. and Zaidi A. K. (2005), "On Temporal Analysis of Timed Influence Nets using Point Graphs," in the Proc. of the 18th International FLAIRS Conference, FL.
 - [12] Haider S. and Levis A. H. (June 2005), "Dynamic Influence Nets: An Extension of Timed Influence Nets for Modeling Dynamic Uncertain Situations," in Proc. 10th International Command and Control Research and Technology Symposium, Washington DC.
 - [13] Kushmerick N., Hanks S., and Weld D., (1995), "An algorithm for probabilistic planning," Univ. of Washington, Seattle, Artificial Intelligence Journal, pp. 76(1-2):239-86
 - [14] Likhachev M. and Stentz A., (2006), "PPCP: Efficient Probabilistic Planning with Clear References in Partially Known Environments", Twenty-First AAAI Conference on Artificial Intelligence, Boston, Massachusetts, The AAAI Press/Menlo Park, California, pp 860-868.
 - [15] Maes P., (1990), "Situated agents can have goals," Robot. Auton. Syst., vol. 6, pp. 49-70.
 - [16] Mausam and Weld D.S., "Probabilistic Temporal Planning with Uncertain Durations", Twenty-First AAAI Conference on Artificial Intelligence, Boston, Massachusetts, The AAAI Press/Menlo Park, California, pp 880-888.
 - [17] Mitchell T.M., (1990), "Becoming increasingly reactive," in Proc. 8th Nat.Conf. Artificial Intelligence (AAAI-90). Chicago, IL: AAAI Press/The MIT Press, pp. 1051–1058.
 - [18] Nikolova E. and Karger D.R., "Route Planning under Uncertainty: The Canadian Traveler Problem", Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, Illinois, The AAAI Press/Menlo Park, California, pp 969-975.
 - [19] Tate A., Hendler J., and Drummond M., (1990), "A review of AI planning techniques,"in Readings in Planning, J. Allen, J. Hendler, and A. Tate,Eds. San Mateo, CA: Morgan Kaufmann, pp. 26–49.
 - [20] Wagenhals L. W. and Levis A. H. (June 2002), "Modeling Support of Effects-Based Operations in War Games," in Proc. 2002 Command and Control Research and Technology Symposium, Monterey, CA.
 - [21] Wagenhals L.W., Levis A. H., and McCrabb M. (June 2003), "Effects Based Operations: a Historical Perspective for a Way Ahead," in Proc. 8th Int'l Command and Control Research and Technology Symposium, National Defense University, Washington, DC.
 - [22] Wagenhals L. W. and Wentz L. K., (June 2003), "New Effects-Based Operations Models in War Games," in Proceedings of the 2003 International Command and Control Research and Technology Symposium, National Defense University, Washington, DC.
 - [23] Wagenhals L. W. (2000), "Course of Action Development and Evaluation Using Discrete Event System Models of Influence Nets", PhD Dissertation, George Mason University.
 - [24] Wagenhals L. W., Levis A. H. (2000), "Course of Action Development and Evaluation," in Proceedings of the 2000 Command and Control Research and Technology Symposium.
 - [25] Wagenhals L. W. and Levis A. H., "Modeling Effects Based Operations in Support of War Games," in Proc. Of the 15th International Symposium on Aerospace / Defense Sensing, Internal Society for Optical Engineering, Proceedings of SPIE, Vol. # 4367, 2001.
 - [26] Wagenhals L. W., Shin I., and Levis A. H., (1998) "Creating Executable Models of Influence Nets with Coloured Petri Nets," Int. J. STTT, Springer-Verlag, Vol. 1998, No. 2.
 - [27] Wagenhals L. W., Reid T. J., Smillie R. J., & Levis A. H. (June 2001), "Course of Action Analysis for Coalition Operations," Proceedings of 6th Interna-

- tional Command and Control Research and Technology Symposium, Annapolis, Maryland.
- [28] Wellman M. P., (1990), “Formulation of tradeoffs in planning under uncertainty”. San Mateo, CA: Morgan Kaufmann.
- [29] Wilkins D.E., (1988), “Practical Planning: Extending the Classical AI Planning Paradigm”. San Mateo, CA: Morgan Kaufmann.
- [30] Zaidi A. K., Wagenhals L W., and Haider S. (2005), “Assessment of Effects Based Operations Using Temporal Logic,” in Proc. of the 10th International Command and Control Research and Technology Symposium.
- [31] Zaidi A. K., Mansoor F., and Papantoni-Kazakos P. (October 2007), “Modeling with Influence Networks Using Influence Constants: A New Approach,” in Proceedings of IEEE-SMC Conference.

Appendix

Proof of Theorem 1

In the derivations below, the following considerations are incorporated:

1. Effects are fully dictated by the actions taken; thus, when probabilities are conditioned on actions and preconditions, the conditioning on preconditions drops.
2. By probability of action success, we mean the probability that the action may succeed, given the preconditions. The final action is selected among those that have positive probability of success. The probability of action taken, given that the action may succeed is the criterion that dictates the final action selection.

$$(a) \quad P(y_1^m | x_1^n) = \sum_{1 \leq j \leq k} P(y_1^m, a_j \text{ taken} | x_1^n) + P(y_1^m, \text{no action taken} | x_1^n) \quad (1.1)$$

where

$$\begin{aligned} P(y_1^m, a_j \text{ taken} | x_1^n) &= \\ &= P(y_1^m | a_j \text{ taken}, x_1^n) P(a_j \text{ taken} | x_1^n) = \\ &= P(y_1^m | a_j \text{ taken}) P(a_j \text{ taken} | x_1^n) = \\ &= P(a_j \text{ taken} | x_1^n) q_j(y_1^m) \\ &= [P(a_j \text{ taken}, \text{succ for action } a_j | x_1^n) + \\ &+ P(a_j \text{ taken}, \text{no succ for action } a_j | x_1^n)] \times \\ &\times q_j(y_1^m) \\ &= [P(a_j \text{ taken} | \text{succ for action } a_j, x_1^n) \times \\ &\times P(\text{succ for action } a_j | x_1^n) + \\ &+ P(a_j \text{ taken} | \text{no succ for action } a_j, x_1^n) \times \\ &\times P(\text{no succ for action } a_j | x_1^n)] q_j(y_1^m) \end{aligned}$$

$$\begin{aligned} &= [P(a_j \text{ taken} | \text{succ for action } a_j) \times \\ &\times P(\text{succ for action } a_j | x_1^n) + \\ &+ P(a_j \text{ taken} | \text{no succ for action } a_j) \times \\ &\times P(\text{no succ for action } a_j | x_1^n)] q_j(y_1^m) \\ &= P(a_j \text{ taken} | \text{succ for action } a_j) \times \\ &\times p_j(x_1^n) q_j(y_1^m) \\ &\quad \{\text{Using } P(a_j \text{ taken} | \text{no succ for action } a_j) = 0\} \end{aligned}$$

Equating in (1.1)

$$\begin{aligned} P(y_1^m | x_1^n) &= \\ &= \sum_{1 \leq j \leq k} P(a_j \text{ taken} | \text{succ for action } a_j) \times \\ &\times p_j(x_1^n) q_j(y_1^m) + P(y_1^m, \text{no action taken} | x_1^n) \quad (1.2) \end{aligned}$$

where

$$\begin{aligned} P(y_1^m, \text{no action taken} | x_1^n) &= \\ &= P(y_1^m | \text{no action taken}, x_1^n) P(\text{no action taken} | x_1^n) \\ &= P(y_1^m | \text{no action taken}) P(\text{no action taken} | x_1^n) \\ &= P(\text{no action taken} | x_1^n) q_0(y_1^m) \\ &= P(\text{no action taken}, \text{no action succ} | x_1^n) q_0(y_1^m) \\ &= P(\text{no action taken} | \text{no action succ}) q_0(y_1^m) \\ &\quad \{\text{Using } P(\text{no action succ} | x_1^n) = 1\} \end{aligned}$$

Equating in (1.2)

$$\begin{aligned} P(y_1^m | x_1^n) &= \\ &= \sum_{1 \leq j \leq k} P(a_j \text{ taken} | \text{succ for action } a_j) p_j(x_1^n) q_j(y_1^m) + \\ &+ P(\text{no action taken} | \text{no action succ}) q_0(y_1^m) \end{aligned}$$

$\Rightarrow \max P(y_1^m | x_1^n)$ attained
 if $P(a_{j^*} \text{ taken} | \text{succ for action } a_{j^*}) = 1$;
 for $p_{j^*}(x_1^n)q_{j^*}(y_1^m) = \max_{1 \leq j \leq k} p_j(x_1^n)q_j(y_1^m) > q_0(y_1^m)$
 otherwise, $\max P(y_1^m | x_1^n)$ attained
 if $P(\text{no action taken} | \text{no action succ}) = 1$;
 for $\max_{1 \leq j \leq k} p_j(x_1^n)q_j(y_1^m) < q_0(y_1^m)$

(b)

$$\begin{aligned}
 \bar{U}(x_1^n) &= \sum_{1 \leq j \leq k} \sum_{y_1^m} P(a_j \text{ taken}, y_1^m | x_1^n) U_j(y_1^m) + \\
 &+ \sum_{y_1^m} P(\text{no action taken}, y_1^m | x_1^n) U_0(y_1^m) \\
 &= \sum_{1 \leq j \leq k} \sum_{y_1^m} P(y_1^m | a_j \text{ taken}) P(a_j \text{ taken} | x_1^n) U_j(y_1^m) + \\
 &+ \sum_{y_1^m} P(y_1^m | \text{no action taken}) P(\text{no action taken} | x_1^n) \\
 &\times U_0(y_1^m) \\
 &= \sum_{1 \leq j \leq k} P(a_j \text{ taken} | \text{succ for action } a_j) p_j(x_1^n) \times \\
 &\times \sum_{y_1^m} q_j(y_1^m) U_j(y_1^m) + P(\text{no action taken} | \text{no action succ}) \times \\
 &\times \sum_{y_1^m} q_0(y_1^m) U_0(y_1^m)
 \end{aligned}$$

$\Rightarrow \max \bar{U}(x_1^n)$ attained for :
 $P(a_{j^*} \text{ taken} | \text{succ for action } a_{j^*}) = 1$;
 if $p_{j^*}(x_1^n) \sum_{y_1^m} q_{j^*}(y_1^m) U_{j^*}(y_1^m) =$
 $= \max_{1 \leq j \leq k} p_j(x_1^n) \sum_{y_1^m} q_j(y_1^m) U_j(y_1^m) > \sum_{y_1^m} q_0(y_1^m) U_0(y_1^m)$
 $P(\text{no action taken} | \text{no action succ}) = 1$;
 if $\sum_{y_1^m} q_0(y_1^m) U_0(y_1^m) > \max_{1 \leq j \leq k} p_j(x_1^n) \sum_{y_1^m} q_j(y_1^m) U_j(y_1^m)$

Proof of the Network Propagation – Problem 1

Using the notation in Section 4, Problem 1, and via the Theorem of Total Probability and the Bayes Rule, we obtain:

$$\begin{aligned}
 r(y_1^m(N)) &= \max_{\text{sequence of actions}} \sum_{y_1^m(N-1)} P(y_1^m(N), y_1^m(N-1) | x_1^n(1)) = \\
 &= \max_{\text{sequence of actions}} \sum_{y_1^m(N-1)} P(y_1^m(N) | y_1^m(N-1)) \times \\
 &\times P(y_1^m(N-1) | x_1^n(1)) \leq \max_{y_1^m(N-1)} \left[\max_{\text{action}} P(y_1^m(N) | \right.
 \end{aligned}$$

$$\begin{aligned}
 &\left. | y_1^m(N-1) \right] \max_{\text{sequence of actions}} P(y_1^m(N-1) | x_1^n(1)) \Big] \\
 &= \max_{y_1^m(N-1)} \left[\left\{ \max_{\text{action}} P(y_1^m(N) | y_1^m(N-1)) \right\} r(y_1^m(N-1)) \right]
 \end{aligned}$$

where, via Theorem 1 we have:

$$\begin{aligned}
 \max_{\text{action}} P(y_1^m(N) | y_1^m(N-1)) &= \\
 &= \max_j \left[\max_{y_1^m(N)} p_j(y_1^m(N-1)) q_j(y_1^m(N)), q_0(y_1^m(N)) \right]
 \end{aligned}$$

Thus, via substitution principle, we obtain:

$$\begin{aligned}
 r(y_1^m(N)) &\leq \max_{y_1^m(N-1)} \left[r(y_1^m(N-1)) \times \right. \\
 &\times \max \left\{ \max_j p_j(y_1^m(N-1)) \times q_j(y_1^m(N)), q_0(y_1^m(N)) \right\} \Big]
 \end{aligned}$$

with equality iff the $y_1^m(N-1)$ value that attains $r(y_1^m(N-1))$ is selected. The above proves the general step in the network propagation of Problem 1.

Proof of the Network Propagation – Problem 2

Using the notation in Section 4, Problem 2, and via the use of the Theorem of Total Probability and the Bayes Rule, we obtain:

$$\begin{aligned}
 \max_{\text{sequence of actions } y_1^w(N)} \sum_{y_1^w(N)} U(y_1^w(N)) P(y_1^w(N) | x_1^n(1)) &= \\
 &= \max_{\text{sequence of actions } y_1^w(N)} \sum_{y_1^w(N)} U(y_1^w(N)) \sum_{x_1^l(N-1)} P(y_1^w(N), \\
 &x_1^l(N-1) | x_1^n(1)) = \\
 &= \max_{\text{sequence of actions } y_1^w(N)} \sum_{y_1^w(N)} U(y_1^w(N)) \sum_{x_1^l(N-1)} P(y_1^w(N) | \\
 &| x_1^l(N-1)) P(x_1^l(N-1) | x_1^n(1)) = \\
 &= \max_{\text{sequence of actions } x_1^l(N-1)} \sum_{y_1^w(N)} P(x_1^l(N-1) | x_1^n(1)) \times \\
 &\times \sum_{y_1^w(N)} U(y_1^w(N)) P(y_1^w(N) | x_1^l(N-1)) = \\
 &= \max_{\substack{a_j(x_1^l(k)); \\ k=1, \dots, N-2}} \sum_{x_1^l(N-1)} P(x_1^l(N-1) | x_1^n(1)) \times \\
 &\times A_{j^*(x_1^l(N-1))} x_1^l(N-1)
 \end{aligned}$$

The latter expression proves the back propagation property and the steps in the algorithm.

Two-Way Mapping between Object-Oriented Databases and XML

Taher Naser

School of Informatics, Bradford University, Bradford, West Yorkshire, United Kingdom

Reda Alhajj

Department of Computer Science, University of Calgary, Calgary, Alberta, Canada;
Department of Computer Science, Global University, Beirut, Lebanon

Mick J. Ridley

School of Informatics, Bradford University, Bradford, West Yorkshire, United Kingdom

Keywords: algorithms, data migration, data re-engineering, object-oriented databases, XML

Received: October 12, 2008

This paper presents a novel approach for mapping an existing object-oriented database into XML and vice versa. The major motivation to carry out this study is the fact that it is necessary to facilitate platform independent exchange of the content of object oriented databases and the need to store XML in a structured database. There are more common features between the object-oriented model and XML and thus the two-way mapping from object-oriented databases into XML (and vice versa) should be less problematic. To achieve the mapping, what we call the object graph is derived based on characteristics of the schema to be mapped. For object-oriented schema, the object graph simply summarizes and includes all nesting and inheritance links, which are the basics of the object-oriented model. Then, the inheritance is simulated in terms of nesting to get a simulated object graph. This way, everything in a simulated object graph is directly representable in XML format. Finally, we handle the mapping of the actual data from the object-oriented database into corresponding XML document(s). On the other hand, the common features between the object-oriented model and XML make it is more attractive to map from XML into object-oriented database; such mapping preserves database specifics. To achieve the mapping, the object graph is derived based on characteristics of the XML schema; it simply summarizes and includes all complex and simple elements and the links, which are the basics of the XML schema. Then, the links are simulated in terms of nesting to get a simulated object graph. This way, everything in a simulated object graph is directly representable in object-oriented database. Finally, we handle the mapping of the actual data from XML document(s) into the corresponding object-oriented database.

Povzetek: Prispevek predstavlja izvorno dvostransko preslikavo med objektnimi podatkovnimi bazami in XML.

1 Introduction

XML is emerging as the standard format for data exchange between different partners. Since most of the data nowadays reside in structured databases including relational and object-oriented databases, it is important to automate the process of generating XML documents containing information from existing databases. Of course, one would like to preserve as much information as possible during the transformation process. The object-oriented database [12, 13] to XML conversion involves mapping the classes and attributes' names into XML elements and attributes' names, creating XML hierarchies, and processing values in an application specific manner. This paper addresses the mapping of the contents of an existing object-oriented database into XML; the reverse process is also supported to allow storing XML data in object-oriented database. The ma-

ior motivation to carry out this study is the fact that there is a need for platform independent format for exchanging data; XML is accepted as one standard for achieving such task. We initiated this study based on our previous research related to object-oriented databases and database re-engineering as illustrated, respectively, in [3, 4, 5, 6, 7, 8] and [9, 10].

The mapping from object-oriented data into XML has not received considerable attention. On the other hand, there exist several tools that enable the composition of XML documents from relational data, such as IBM DB2 XML Extender, SilkRoute, and XPERANTO. XML Extender [15] serves as a repository for XML documents as well as their Document type declarations (DTDs), and also generate XML documents from existing data stored in relational database; it is used to define the mapping of DTD to relational tables and columns. XSLT and Xpath syntax are

used to specify the transformation and the location path. SilkRoute [17] is described as a general, dynamic, and efficient tool for viewing and querying relational data in XML. XPERANTO [14] is a middleware solution for publishing XML; object-relational data can be published as XML documents. It can be used by developers who prefer to work in a “pure XML” environment. However, the mapping from the relational schema to the XML schema is specified by human experts. Therefore, when a large relational schema and corresponding data need to be translated into XML documents, a significant investment of human effort is required to initially design the target schema. Finally, the work described in [20] requires knowing the catalog contents in order to extract the relational database schema. The conversion of Relational-to-ER-to-XML has been proposed in [18]. This reconstructs the semantic model, in the form of ER model, from the logical schema capturing user’s knowledge, and then converts it to the XML document. However, many-to-many (M:N) and nary relationships are not considered properly. Finally, DB2XML [26] is a tool for transforming data from relational databases into XML documents; DTDs are generated by describing the characteristics of the data for making the documents self contained and usable as a data exchange format.

The conversion of Relational-to-ER-to-XML is described in [18]. VXE-R [21] is an engine for transforming a relational schema into equivalent XML schema. As the mapping from XML to object-oriented databases is concerned, the work described in [16] generates an object-oriented database schema from DTDs, stores it into the object-oriented database and processes XML queries; it mainly concentrate on representing the semi-structural part of XML data by inheritance. However, in this paper we differentiate between inheritance and nesting, which is a more natural approach for handling object-oriented databases. The work presented in [19] focuses on the ability to wrap an XML schema definition in an object-oriented virtual database mediator system to help solving the integration problems between XML documents and other applications that are not using XML. Toth and Valenta [25] investigated possibilities of reusing already known techniques from object and object-oriented processing in XML-native database systems.

This paper addresses the two-way mapping of the contents of an existing XML and object-oriented database. The major motivation to carry out this study is the fact that there are more common features between XML and object-oriented databases; thus it is more attractive to store XML schema and Data, and more data is preserved. This is actually the backward engineering [24]; the forward engineering part extracts XML from object-oriented database [23]. The forward engineering process takes a given object-oriented database as input and produces a corresponding XML schema and XML document(s). The first step in the process is to derive a summary of the object-oriented schema. This has been realized as object graph which includes inheritance and nesting links present in the object-

oriented schema. Then, the object graph is transformed into XML schema and the object-oriented data is mapped into corresponding XML document(s). The process is capable of producing both nested and flat XML schemas. However, as the transformation is from object-oriented databases, producing the nested schema is preferred and more emphasized. The backward engineering process, on the other hand, takes a given XML schema as input and produces a corresponding object-oriented schema. The first step in the process is to derive a summary of the XML schema. This has been realized as object graph, which includes inheritance and nesting links derived from the XML schema. Then, the object graph is mapped into object-oriented schema. The process is capable of taking as input both nested and flat XML schemas. However, as the mapping is into object-oriented schema, nested XML schema is preferred and more emphasized.

The rest of the paper is organized as follows. Described in Section 2 is the information related to the object-oriented schema and the XML schema; the object graph is also defined. Section 3 presents the algorithm that derives the XML schema from the object graph. Section 4 describes the backward engineering process. Section 5 includes a summary and the conclusions.

2 The necessary background and terminology

2.1 Object-Oriented Database Characteristics

In this section, we investigate characteristics of the given object-oriented database and as a result derive the object graph. We start by presenting the basic terminology and definitions required to understand the analysis.

2.1.1 The Basic Terminology and Definitions

We are mainly interested in class characteristics as present in Definition 2.1 and illustrated in Example 2.1, given next.

Definition 2.1 (Class).

A class is defined to be a tuple, $(C_p(c), C_b(c), L_{attributes}(c), L_{behavior}(c), L_{instances}(c), OIDG)$, where c is class identifier, $C_p(c)$ is a list¹ of direct superclasses of class c , $C_b(c)$ is a set² of direct subclasses of class c , $L_{attributes}(c)$ is the set of additional attributes locally defined in class c , $L_{behavior}(c)$ is the set of additional methods added to the definition of class c , $L_{instances}(c)$ is the set of object identifiers of objects added locally to class c , and $OIDG$ is object identifier generator that holds

¹A list notation is used for the superclasses because their order is important for conflict resolution due to polymorphism and overriding. Conflicts are resolved according to certain predefined rules discussed in [4].

²Conflict resolution is not applicable here because only objects are concerned, hence the set notation is utilized.

the identifier to be granted to the next object to be added to $L_{instances}(c)$.

Every attribute in a class has a domain. Inheritance makes it possible for a class to utilize the attributes and methods defined for its superclasses, without violating polymorphism and overriding rules discussed in [4]. The set of objects of a class includes objects in its subclasses. All of this is formalized in the following definitions.

Definition 2.2 (Domain). Let c_1, c_2, \dots , and c_n be primitive and user defined classes, where primitive classes include reals, integers, strings, etc. The following are possible domains,

1. $(a_1:c_1, a_2:c_2, \dots, a_n:c_n)$ is a tuple domain; a possible value is a tuple with the constituting values being object identifiers selected from classes c_1, c_2, \dots , and c_n , respectively.
2. $c_i, 1 \leq i \leq n$, is a domain; a possible value is an object identifier from class c_i .
3. $\{d\}$ is a domain, where d may be any of the two domains defined in 1 or 2; a possible value is a set of values from domain d .
4. $[d]$ is a domain, where d may be any of the two domains defined in 1 or 2; a possible value is a list of values from domain d .

Definition 2.3 (Attributes). Given a class c ; the set of attributes that determine the state of each object in $L_{instances}(c)$ is denoted by $W_{attributes}(c)$ and defined recursively in terms of the attributes defined for objects of the classes in $C_p(c)$.

$$W_{attributes}(c) = L_{attributes}(c) \cup_{i=1}^n W_{attributes}(c_{p_i}).$$

Definition 2.4 (Behavior). Given a class c and let $C_p(c) = [c_{p_1}, c_{p_2}, \dots, c_{p_n}]$ be the list of its direct superclasses. The whole behavior for class c , denoted by $W_{behavior}(c)$, is recursively defined to include the whole behavior of the classes in $C_p(c)$.

$$W_{behavior}(c) = L_{behavior}(c) \cup_{i=1}^n W_{behavior}(c_{p_i}).$$

Definition 2.5 (Extent). Given a class c and let $C_b(c) = \{c_{b_1}, c_{b_2}, \dots, c_{b_m}\}$ be the set of its direct subclasses. All objects that understand at least the behavior in $W_{behavior}(c)$, constitute the extent of class c , denoted by $W_{instances}(c)$. This set is recursively defined in terms of the extents of the classes in $C_b(c)$.

$$W_{instances}(c) = L_{instances}(c) \cup_{i=1}^m W_{instances}(c_{b_i}).$$

Example 2.1 (Classes).

Next are characteristics of the classes in the object-oriented schema:

Person:

$$\begin{aligned} C_p(Person) &= [C_b(Person) = \{Student, Staff, Secretary\}] \\ L_{attributes}(Person) &= \{SSN:integer, name:string, age:integer, \\ &sex:character, spouse:Person, nation:Country\} \\ L_{behavior}(Person) &= \{SSN(), SSN(i), name(), name(t), age(), \\ &age(i), sex(), sex(t), spouse(), spouse(p), nation(), nation(c)\} \end{aligned}$$

Country:

$$\begin{aligned} C_p(Country) &= [] \quad C_b(Country) = \{\} \\ L_{attributes}(Country) &= \{Name:string, area:integer, popula- \\ &tion:integer\} \\ L_{behavior}(Country) &= \{Name(), Name(t), area(), area(i), popu- \\ &lation(), population(i)\} \end{aligned}$$

Student:

$$\begin{aligned} C_p(Student) &= [Person] \\ C_b(Student) &= \{ResearchAssistant\} \\ L_{attributes}(Student) &= \{StudentID:integer, gpa:real, stu- \\ &dent_in:Department, Takes:\{(course:Course, grade:string)\}\} \\ L_{behavior}(Student) &= \{StudentID(), StudentID(i), gpa(), gpa(i), \\ &student_in(), student_in(d), Takes(), Takes(t)\} \end{aligned}$$

Staff:

$$\begin{aligned} C_p(Staff) &= [Person] \\ C_b(Staff) &= \{ResearchAssistant\} \\ L_{attributes}(Staff) &= \{StaffID:integer, salary:integer, \\ &works_in:Department\} \\ L_{behavior}(Staff) &= \{StaffID(), StaffID(i), salary(), salary(i), \\ &works_in(), works_in(d)\} \end{aligned}$$

ResearchAssistant:

$$\begin{aligned} C_p(ResearchAssistant) &= [Student, Staff] \\ C_b(ResearchAssistant) &= \{\} \\ L_{attributes}(ResearchAssistant) &= \{\} \\ L_{behavior}(ResearchAssistant) &= \{\} \end{aligned}$$

Course:

$$\begin{aligned} C_p(Course) &= [] \quad C_b(Course) = \{\} \\ L_{attributes}(Course) &= \{Code:integer, title:string, credits:integer, \\ &Prerequisite:\{Course\}\} \\ L_{behavior}(Course) &= \{Code(), Code(i), title(), title(t), credits(), \\ &credits(i), Prerequisite(), Prerequisite(c)\} \end{aligned}$$

Department:

$$\begin{aligned} C_p(Department) &= [] \quad C_b(Department) = \{\} \\ L_{attributes}(Department) &= \{Name:string, head:Staff\} \\ L_{behavior}(Department) &= \{Name(), Name(t), head(), head(t)\} \end{aligned}$$

Secretary:

$$\begin{aligned} C_p(Secretary) &= [Person] \quad C_b(Secretary) = \{\} \\ L_{attributes}(Secretary) &= \{words/minute:integer, \\ &works_in:Department\} \\ L_{behavior}(Secretary) &= \{words/minute(), words/mimute(i), \\ &works_in(), works_in(d)\} \end{aligned}$$

It is clear from Example 2.1 that the behavior of a class contains two methods for every attribute. These methods are automatically generated by the system when the attribute is defined. For instance, the two methods $SSN()$, $SSN(i)$ are included in $L_{behavior}(Person)$ because attribute SSN belongs to $L_{attributes}(Person)$. While the first method retrieves the value of attribute SSN from the receiving object, the second method $SSN(i)$ sets the value of attribute SSN within the receiving object to the value of the argument i .

2.1.2 The Necessary Object-Oriented Schema Information

Related to the object-oriented schema, the analysis is based on the domain information summarized in the following table.

ObjectAttributes(class name, attribute name, domain)

Class Name	Attribute Name	Domain
Person	ssn	integer
Person	name	string
Person	age	integer
Person	sex	integer
Country	name	string
Country	area	integer
Country	population	integer
Student	studentID	integer
Student	gpa	real
Staff	staffID	integer
Staff	salary	integer
Course	code	integer
Course	title	string
Course	credits	integer
Department	name	string
Secretary	word_minute	integer
T1	grade	string

(a)

Class Name	Attribute Name	Domain
Person	spouse	Person
Person	nation	Country
Student	student_in	Department
Student	Takes	T1
Staff	work_in	Department
Course	prerequisite	Course
Department	head	Staff
Secretary	work_in	Department
T1	course	Course

(b)

Table 1: *ObjectAttributes*: (a) a list of all attributes with primitive domains (b) a list of all attributes with non-primitive domains

The *ObjectAttributes* table includes information about all attributes in the object-oriented schema. For each attribute, it is required to know its name, class and domain. Attributes with primitive domains and attributes with non-primitive domains are placed in separate occurrences of the *ObjectAttributes* table, namely *ObjectAttributes(a)* and *ObjectAttributes(b)*, respectively. Table 1 includes the *ObjectAttributes* information related to the object-oriented schema introduced in Example 2.1. Each domain of the tuple type is assigned a short name that consists of the letter 'T' suffixed with a consecutive non-decreasing number, starting with 1. For instance, as shown in the fourth row in Table 1(b), the short name T_1 has been assigned to the domain of the attribute *Takes* from $L_{instances}(Student)$. This way, it becomes trivial to identify attributes that appear within a tuple domain as illustrated in the last row of each of *ObjectAttributes(a)* and *ObjectAttributes(b)* in Table 1.

2.2 XML Schema Characteristics

XML schema is a language for describing the structure and constraining the content of XML documents. So, it can be described as a set of rules to which an XML document must conform in order to be considered *well-formed* and *valid* document.

In our work, we will use XML Schema complex type elements, primitive type elements and sequence indicator. XML Schema allows us to define the cardinality of an element (*i.e.*, the number of its possible occurrences) with some precision. This cardinality constraint can be explicated by associating the two XML built-in attributes *minOccurs* and *maxOccurs*, with subelements under the “complexType” element. We can specify both *minOccurs* (the minimum number of occurrences) and *maxOccurs* (the maximum number of occurrences). It is possible to set *maxOccurs* to unbounded, which means that there can be

as many occurrences of the character element as the author wishes. Both attributes have a default value of one. If both *minOccurs* and *maxOccurs* are omitted, the subelement must appear exactly once.

Definition 2.6 (ComplexType). A complextype element is defined as a tuple, $(C_{complextypes}(ct), C_{primitivetypes}(ct), C_{keys}(ct), C_{keyrefs}(ct))$, where *ct* is the complextype identifier, $C_{complextypes}(ct)$ is the complextype elements (subelements) of complextype *ct*, $C_{primitivetypes}(ct)$ is the set of primitive type elements of *ct*, $C_{keys}(ct)$ is the set of keys defined for *ct*, $C_{keyrefs}(ct)$ is the set of key references defined for *ct*.

To demonstrate the complex type concept introduced in Definition 2.6, consider Example 2.2 which starts by a description of some related entities followed by the corresponding XML schema definition.

Example 2.2 (XML Schema). Consider the following set of related entities. Each entity has a set of attributes and a primary key; the domain of each attribute is also specified. Drawing the corresponding entity-relationship diagram is straightforward by considering the given summary.

Person Complex Type: Key = SSN
 attributes = {ssn :integer; name:string; age :integer; sex :character; spouse :Person; nation :Country}

Country Complex Type: KEY= Name
 attributes = {name:string; area :integer; population :integer}

Student Complex Type: Key=StudentID
 attributes = {StudentID:integer; gpa :real; student_in :Department; Takes:{{course :Course; grade :string}}}

Staff Complex Type: Key=StaffID
 attributes = {StaffID:integer; salary :integer; works_in :Department}

ResearchAssistant Complex Type: links to both student and staff; hence gets the key of either one.

attributes = {student:Student,staff:Staff}

Course Complex Type: *Key=code*

attributes = {Code :integer; title :string; credits :integer; Prerequisite :{Course}}

Department Complex Type: *Key=name*

attributes = {name:string; head :Staff}

Secretary Complex Type: *links to person; hence gets the primary key of person.*

attributes = {person:Person, wordsperminute :integer; works in :Department}

The next XML Schema segment depicts part of the nested XML Schema. It describes the PERSON and COUNTRY complextype elements. It shows the COUNTRY element as a child element of the PERSON complex type element. PERSON complextype includes an empty PERSON_Object element, where PERSON_Object element is defined as a complex type that includes all attributes of PERSON element. The "sequence" indicator is used as sequential semantic for the set of subelements. It is a flat XML Schema, where the subelement COUNTRY in the complex type element PERSON is defined as a string element type. The two parts of XML schema are connected by "key" and "keyref" constraints.

```
<xsd:complexType name=" PERSON">
  <xsd:sequence>
    <xsd:element name=" PERSON_Object"
      type="PERSON_Tuple" maxOccurs="unbounded"/>
  </xsd:sequence>
</xsd:complexType>
<xsd:complexType name=" PERSON_Object">
  <xsd:sequence>
    <xsd:element name="SSN" type="xsd:int" />
    <xsd:element name="NAME" type="xsd:string"/>
    <xsd:element name="AGE" type="xsd:int" />
    <xsd:element name="SEX" type="xsd:string"/>
    <xsd:element name="SPOUSE" type="PERSON"/>
    <xsd:element name="NATION" type="xsd:String"/>
  </xsd:sequence>
</xsd:complexType> <xsd:complexType
  name=" COUNTRY_Class">
  <xsd:sequence>
    <xsd:element name=" COUNTRY_Tuple"
      type=" COUNTRY_Tuple"
      maxOccurs="unbounded"/>
  </xsd:sequence>
</xsd:complexType> <xsd:complexType
  name="COUNTRY_Object">
  <xsd:sequence>
    <xsd:element name="NAME" type="xsd:string" />
    <xsd:element name="AREA" type="xsd:int" />
    <xsd:element name="POPULATION" type="xsd:int"/>
  </xsd:sequence>
</xsd:complexType>
<!-- Define Primary Keys and Keyrefs -- >
<xs:key name="PERSON_PrimaryKey">
  <xs:selector xpath="db:PERSON/db:PERSON_Object"/>
  <xs:field xpath="db:SSN"/>
</xs:key>

<xsd:key name="COUNTRY_PrimaryKey">
  <xsd:selector
    path="db:COUNTRY_Class/db:COUNTRY_Object"/>
  <xs:field xpath="db:Name"/>
</xs:key>
```

```
<xs:keyref name="PERSON.nation"
  refer="db:COUNTRY_PrimaryKey">
  <xs:selector
    xpath="db:PERSON/db:PERSON_Object"/>
  <xs:field xpath="nation"/>
</xs:keyref>
```

A corresponding nested XML schema may be constructed by the same way; but "key" and "keyref" are replaced by the actual nested instead. In other words, a Person element will have a subelement "nation", which includes details of the Country element representing the nationality of the Person element. Here it is worth mentioning that although a nested XML schema reflects better the natural structure and linkage between elements, a corresponding XML document would occupy much more space and will be more difficult to handle in order to maintain database consistency in case of a dynamic database with frequent updates.

2.3 The Object Graph

In this section, we use the information present in the *ObjectAttributes(b)* table and the inheritance information as defined in Section 2.1.1, to draw what we call the *Object Graph (OG)* that includes all possible relationships between the classes present in the given object-oriented schema. Nodes in *OG* are classes and representatives of tuple type domains. Two nodes are connected by a link to show the inheritance or a nesting relationship between them. Nodes and links are represented by small rectangles and directed arrows, respectively. Inheritance and nesting links are assigned the scores 0 and 1, respectively. A link is assigned the score 2 if it is connecting a node that represents a tuple domain and the class in which it is referenced. To illustrate this, refer to attribute *Takes* in *Lattributes(Student)* in Example 2.1 and to the corresponding link connecting the two nodes T_1 and *Student* in Figure 1. More formal details related to *OG* are included in Definition 2.7, given next.

Definition 2.7 (Object Graph).

Every object-oriented schema has a corresponding OG graph (V, E) such that,

1. *for every class c in the object-oriented schema there is a corresponding node c in V ,*
2. *for all classes c_1 and c_2 , such that $c_2 \in C_p(c_1)$, an edge $(c_1, c_2, 0)$ is added to E*
3. *for every class c*
for every attribute $a \in Lattributes(c)$, such that a has a non-primitive domain,
if domain of a involves a class, say c' , then an edge $(c, c', 1)$ is added to E
else if domain of a involves a tuple T_i , ($i \geq 1$)
then
a node T_i is added to V and an edge $(T_i, c, 2)$

is added to E
 for every class c'' that appears as a domain in
 tuple T_i , an edge $(T_i, c'', 1)$
 is added to E

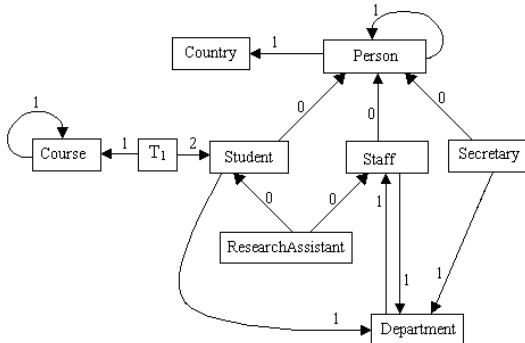


Figure 1: The Object Graph of the object-oriented schema in Example 2.1

As definition 2.7 is concerned, every node T_i in V , such that there exists an edge $(T_i, c'', 2)$ in E , corresponds in the relational graph to a node that represents a relationship that involves more than two relations, or a relationship with attributes. This will become more clear later in Section ??, when the equivalence of the two graphs RG and OG is investigated. Shown in Figure 1 is the OG graph derived from the information present in Table 1(b) and the inheritance information in the C_p lists in Example 2.1.

3 Forward engineering: transforming object graph into XML schema

In this section, we first present the algorithm for transforming the object graph to flat XML schema (OG2FXML); then we present the algorithm for transforming the object graph to nested XML schema (OG2NXML).

3.1 Object Graph to Flat XML Schema Transformation

The OG2FXML in pseudo-code is depicted in Algorithm 3.1.

Algorithm 3.1 OG2FXML (object graph to Flat XML Conversion)

Input: The Object Graph

Output: The corresponding flat XML schema

Step:

1. Transform each node in the object graph (we call it class hereafter) into a “complexType” in the XML schema.

2. Map each attribute in a class transformed in Step (1) into a subelement within the corresponding “complexType”.
3. Create a root element as the object-oriented database schema name and insert each class identified in Step (1) as a subelement with the corresponding “complexType”.
4. Define the primary key for each class identified in Step (1) by using “key” element.
5. Map in the object graph each link between classes identified in Step (1) by using “keyref” element.

EndAlgorithm 3.1

To understand the steps of Algorithm 3.1, we present more details with supporting examples.

- Each class E in the object graph is translated into an XML “complexType” of the same name E in the XML schema. In each “complexType” E , there is only one empty element, which includes several subelements. For example, COUNTRY is translated into a “complexType” named COUNTRY_Class. The empty element is called COUNTRY_Object.

```
<xs:complexType name="COUNTRY_Class">
  <xs:sequence>
    <xs:element name="COUNTRY_Object"
      type="db:COUNTRY_Object" maxOccurs="unbounded"/>
  </xs:sequence>
</xs:complexType>
<xs:complexType name="COUNTRY_Object">
  <xs:sequence>
    . . . . .
  </xs:sequence>
</xs:complexType>
```

The cardinality constraint can be explicated by associating two XML built-in attributes (also called indicators), namely “minOccurs” and “maxOccurs”, with subelements under the “complexType” element. The default value for both “maxOccurs” and “minOccurs” is 1. If specified, the value for “minOccurs” should be either 0 or 1 and the value for “maxOccurs” should be greater than or equal to 1. If both “minOccurs” and “maxOccurs” are omitted, the subelement must appear exactly once.

- Each attribute A_i in E is mapped into a subelement of the corresponding “complexType” E . For example, COUNTRY is mapped into a “complexType” named COUNTRY_Object, inside which there are several subelements such as Name, Area and Population. They are attributes of the COUNTRY class. The XML schema for COUNTRY is:

```
<xs:complexType name="COUNTRY_Object">
  <xs:sequence>
    <xs:element name="Name" type="xs:string" />
    <xs:element name="Area" type="xs:int" />
    <xs:element name="Population" type="xs:int" />
  </xs:sequence>
</xs:complexType>
```

The “sequence” specification in the XML schema captures the sequential semantics of a set of subelements. For instance, in the “sequence” given above, the subelements appear in the order: Name, Area and Population. They must appear in instance documents in the same order as they are declared here. The XML schema also provides another constructor called “all”, which allows elements to appear in any order, and each element must appear once or not at all.

- Each class in the object graph is mapped into the XML schema. We first need to create a root element that represents the entire given object-oriented database. We create the root element as a “complexType” in XML schema and give it the same name as the object-oriented database schema. It then inserts each class as a subelement of the root element. An example which contains the eight classes PERSON, COUNTRY, STUDENT, STAFF, RESEARCH_ASSISTANT, COURSE, DEPARTMENT, and SECRETARY is now presented. We give the root element the name UNIVERSITY:

```
<xs:element name="UNIVERSITY">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="PERSON_Class"
        type="db:PERSON_Class" />
      <xs:element name="COUNTRY_Class"
        type="db:COUNTRY_Class" />
      <xs:element name="STUDENT_Class"
        type="db:TUDENT_Class" />
      <xs:element name="STAFF_Class"
        type="db:STAFF_Class" />
      <xs:element name="RESEARCH_ASSISTANT_Class"
        type="db:RESEARCH_ASSISTANT_Class" />
      <xs:element name="COURSE_Class"
        type="db:COURSE_Class" />
      <xs:element name="DEPARTMENT_Class"
        type="db:DEPARTMENT_Class" />
      <xs:element name="SECRETARY_Class"
        type="db:SECRETARY_Class" />
    </xs:sequence>
  </xs:complexType>
  <!-- definition of keys and keyrefs -->
  . . . . .
</xs:element>
```

- The elements “key” and “keyref” are used to enforce the uniqueness and referential constraints. They are among the key features introduced in the XML schema. Further, we can use “key” and “keyref” to specify the uniqueness scope and multiple attributes in creating composite keys. Consider this example:

```
<xs:key name="COUNTRY_PrimaryKey">
  <xs:selector
    xpath="db:COUNTRY_Class/db:COUNTRY_Object"/>
  <xs:field xpath="db:Name" />
</xs:key>
<xs:key name="PERSON_PrimaryKey">
  <xs:selector
    xpath="db:PERSON_Class/db:PERSON_Object"/>
  <xs:field xpath="db:nation" />
</xs:key>
<xs:keyref name="PERSON.nation"
  refer="db:COUNTRY_PrimaryKey">
  <xs:selector
    xpath="db:PERSON_Class/db:PERSON_Object"/>
  <xs:field xpath="nation" />
</xs:keyref>
```

In this example, nation is like a foreign key in PERSON, so we use “keyref” to specify the foreign key relationship between COUNTRY and PERSON. Compared to DTD, the XML schema provides a more flexible and powerful mechanism through “key” and “keyref”, which share the same syntax as “unique” also make referential constraints possible in XML documents.

In general, OG2FXML is a straightforward and effective transformation algorithm, but it is only applicable when generating a flat XML structure from an object graph of an object-oriented database. As the name implies, OG2FXML cannot handle the nested features provided by XML. We remedy this problem in the OG2NXML algorithm which will be presented in the following section.

3.2 Object Graph to Nested XML Schema Transformation

In XML schema, we can use nested complex type elements to define the relationship between two elements. One advantage of the nested XML structure is to store all related information in one fragment of an XML document. This reduces the time for data retrieval when users query on the XML document. Algorithm 3.2 (OG2NXML) does the transformation from the object graph to a nested XML structure.

The OG2NXML depends on the nesting sequence specified in the object graph and generates an output of nested XML schema. The OG2NXML in pseudo-code is depicted in Algorithm 3.2.

Algorithm 3.2 OG2NXML (object graph to Nested XML Conversion)

Input: The object graph

Output: The corresponding nested XML schema

Step:

For classes connected by a link labeled with 1 in the object graph, we nest the element that correspond to the class at the head of the arrow inside the element that correspond to the class at the tail of the arrow.

For classes connected by a link labeled with 0 do

Extend the element that correspond to the subclass to include the content of the element that correspond to the superclass.

EndAlgorithm 3.2

To illustrate the nesting process, consider the UNIVERSITY database; it is taken as input by OG2NXML which generates as output the XML schema in a nested structure. The element of COUNTRY is nested under the element of PERSON. The nested element then included separately inside the elements of STAFF and STUDENT because PERSON is a superclass of each of the two latter classes. This way, inheritance is resolved by extending the content of

the subclass to include the attributes defined in the super-classes. Handling the inheritance relationship in this way is more natural because it is not supported in XML.

3.2.1 Generating XML Documents

After the XML schema is obtained, the next step is to generate XML document(s) from the considered object-oriented database. Algorithm 3.3 (GenXMLDoc) checks top-down through the list of selected objects and generates an element for each object.

Algorithm 3.3 GenXMLDoc (Generating XML Document)

Input: XML schema and object-oriented database

Output: The corresponding XML Document

Step:

Create XML document and set its namespace declaration

Create a root element of the XML document with the same name as the root name of the XML schema

For each class R in the object-oriented database do

If R is selected and does not contain any nested classes

Create R_Class element for R

Let queryString = “select * from R ”

ResultSet = execute(queryString)

For each object T in ResultSet do

Create R_Object element for object T

Create an element for each attribute in R and insert it into R_Object element

else if R is selected and contains a nested class R_c then

Create R_Class element for R and R_c_Class for R_c

Let queryString = “select selectedAttrs from R , R_c ”

ResultSet = execute(queryString)

For each object T in ResultSet do

Create R_Object element for the tuple of R , and R_c_Object element for the object of R_c

Create an element for each selected attribute in R and insert it to R_Object element, and do same for R_c

EndAlgorithm 3.3

Algorithm 3.3 can generate flat XML documents as well as nested XML documents, depending on the processed XML schema. In Algorithm 3.3, a query is executed to obtain all objects that satisfy the constraints so one element is created to store data of each object in the result set.

4 Backward engineering: from XML schema to object-oriented database

4.1 XML Schema Information

Related to the nested XML schema, the analysis is based on the domain information summarized in the following table:

XMLAttributesNE(complextype, element name, domain)

To understand better the content and purpose of this table, *XMLAttributesNE* shown in Table 2 includes information about all elements and attributes in the XML schema given in Example 2.2. For each element, it is required to know its complextype name, element name, and the domain. Elements with primitive domains and elements with non-primitive domains are placed in separate occurrences of the *XMLAttributes* table, namely *XMLElementsNE(a)* and *XMLElementsNE(b)*, respectively.

Concerning the information in Table 2(b), user involvement is required to suggest which element is inherited - representing a superclass in the object-oriented Database -, the element that can represent the nested non primitive domain attributes and the element that can represent a tuple type domain attribute. In Table 2(b), an “Inheritance Flag” is assigned to each element. The score 0 is given for the candidate superclass elements (inherited element), score 1 is given for the nested non-primitive domain elements and score 2 is assigned for the tuple domain elements. Subclasses are not included in the attributes of the class because they could be inspired by considering the superclasses list of subclasses. For instance, the “nation” element is given the value 1 for the “Inheritance Flag” because it is a nested non-primitive domain. The inheritance flag in row 3 is given the value 1 because student_in of type DEPARTMENT is a nested type, while it is given the value 0 in row 5 because PERSON is candidate superclass for STUDENT (inheritance). Also, row seven is given the value 0 as ResearchAssistant is a subclass of STUDENT and STAFF; that means STUDENT and STAFF are superclasses for ResearchAssistant. This way, it becomes trivial to identify superclasses, subclasses and non-primitive domain attributes using *XMLElementsNE(a)* and *XMLElementsNE(b)* as given in Table 2.

Related to the flat XML schema, the analysis is based on the domain information summarized in the following table: *XMLAttributesFL (complextype, element name, domain, expected domain, inheritance flag, keys information, key ref information)* The need for the information depicted in *XMLAttributesFL* is better understood by considering the XML schema in Example 2.2; the corresponding *XMLAttributesFL* shown in Table 3 includes information about all elements and attributes in the flat XML schema. This information is described in 3 tables named “a”, “b”, and “c”. For each element, it is required to know its complextype name, element name, domain, expected domain name and the inheritance status (inherited or not). Also,

Complex Type Name	Element Name	Domain
Person	ssn	integer
Person	name	string
Person	age	integer
Person	sex	integer
Country	name	string
Country	area	integer
Country	population	integer
Student	studentID	integer
Student	gpa	real
Staff	safffID	integer
Staff	salary	integer
Course	code	integer
Course	title	string
Course	credits	integer
Department	name	string
Secretary	word_minute	integer
T1	grade	string

(a)

Complex Type Name	Element Name	Domain	Inheritance Flag
Person	spouse	Person	1
Person	nation	Country	1
Student	student_in	Department	1
Student	Takes	T1	2
Student	person	Person	0
Staff	work_in	Department	1
ResearchAssistant	student	Student	0
ResearchAssistant	staff	Staff	0
Course	prerequisite	Course	1
Secretary	work_in	Department	1
T1	course	course	1

(b)

Table 2: *XMLAttributesNE*: (a) a list of all elements attributes with primitive domain (b) a list of all elements attributes with non-primitive domains

Complex Type Name	Element Name	Domain	Expected Domain	Inheritance Flag
Person	ssn	integer	integer	9
Person	name	string	string	9
Person	age	integer	integer	9
Person	sex	integer	integer	9
Person	spouse	string	Person	1
Person	nation	string	Country	1
Country	name	string	string	9
Country	area	integer	integer	9
Country	population	integer	integer	9
Student	studentID	integer	integer	9
Student	gpa	real	real	9
Student	student_in	string	Department	1
Student	Takes	T1	T1	2
Student	person	integer	Person	0
Staff	safffID	integer	integer	9
Staff	salary	integer	integer	9
Staff	work_in	integer	Department	1
ResearchAssistant	student	integer	Student	0
ResearchAssistant	staff	integer	Staff	0
Course	code	integer	integer	9
Course	title	string	string	9
Course	credits	integer	integer	9
Course	prerequisite	string	Course	1
Department	name	string	string	9
Secretary	word_minute	integer	integer	9
Secretary	work_in	string	Department	1
T1	Course	string	Course	1
T1	grade	string	string	9

(a)

Key Name	Complex Type Name	Type	Element Name
Person_pk	Person		ssn
Country_pk	Country		name
Student_pk	Student		studentid
Staff_pk	Staff		staffid
Course_pk	Course		code
Department_pk	Department		name

(b)

Key Reference Name	Ref. Complex type	Ref. element	Refer to Element
Person.nation	Person	nation	country_pk
student.Student_in	Student	student_in	Department_pk
Secretary.work_in	Secretary	work_in	Department_pk

(c)

Table 3: *XMLAttributesFL*: (a) a list of all elements attributes with primitive and non primitive domains (b) a list of all keys for the complex type elements (c) a list of key references of the complex type elements

it is required to know the complex type elements "keys" and "keyrefs". Information about the elements is placed in *XMLElementsFL* Table 3(a), Keys information is placed in *XMLKeys* Table 3(b), while key reference information is replaced in *XMLKeyRef* Table 3(c).

In Table 3(a), user involvement is required to suggest which element is inherited (a candidate superclass in the object-oriented database), the element that can represent the nested non-primitive domain attributes and the element that can represent the tuple type domain attributes. A value is assigned to "Inheritance Flag" for each element. The score 0 is given for the inherited element (candidate superclass elements), score 1 is given for the nested non-primitive domain elements, score 2 is assigned for the tuple domain elements, and score 9 is assigned to the primitive domain elements. User involvement and the information available in *XMLKeys* Table 3(b) and in *XMLKeyRef* Table 3(c) can define the expected non-primitive domain for flat XML primitive domain elements. For instance, the "nation" element is given the value 1 for the "Inheritance Flag" and an expected domain COUNTRY. This is because by analyzing and connecting the information in *XMLKeys* Table 3(b) and in *XMLKeyRef* Table 3(c), it can be shown that there is a reference link between the nation element and the country_pk in COUNTRY complex type. Row 1 is given the score 9 because it is a primitive domain element. In STUDENT complex type element, "person" element of primitive type element is given the score 0 because the user involvement can decide that this is an inherited element, and thus it could be mapped as a superclass for STUDENT. As a result, it becomes easy to construct Table 2 from information in Table 2. Explicitly, primitive domains can be mapped into *XMLElementsNE* Table 2(a), and expected non-primitive domains can be mapped into *XMLElementsNE* Table 1(b). This way, it becomes trivial to identify superclasses, subclasses and non-primitive domain attributes using *XMLElementsNE* (a) and (b) in Table 2.

To sum up, the information needed to map into the object-oriented database is summarized in table *XMLElementsNE*. This table is derived directly derivable from a nested XML schema. However, for flat XML schema, the process involves like a preprocessing step to derive the information in table *XMLElementsFL*, which is used to construct *XMLElementsNE*. This opens the door for a new relational to object-oriented database conversion by converting a relational database directly into a flat XML schema and then map the latter into object-oriented schema.

As Example 2.2 is concerned, shown in Figure 2.1 is the object graph derived from the information present in Table 2(b) and the inheritance information provided by an expert based on the content of Table 2.

4.2 Transferring Object Graph into Object-Oriented Schema

In this section, we present an algorithm for transforming the object graph to object-oriented Schema (OG2OODB).

Algorithm 4.1 OG2OODB (OG to object-oriented Schema conversion)

Input: The Object Graph

Output: The corresponding object-oriented Schema

1. Transfer each node in the object graph (we call it complex type hereafter) into a class in the object-oriented schema. Exclude nodes like T_i , ($i \geq 1$).
2. Map each subelement of primitive type in table *XMLElementsNE*(a) into a primitive attribute in the corresponding class. Exclude subelements like T_i , ($i \geq 1$).
3. Map each subelement of non primitive domain with score 1 defined in table *XMLElementsNE*(b) into the non primitive attributes in the corresponding class. Exclude subelements like T_i , ($i \geq 1$).
4. Map each subelement of non-primitive domain with score 2 defined in table *XMLElementsNE*(b) as a tuple non-primitive attributes in the corresponding class. Add to this tuple non-primitive attributes all elements of complex type name equivalent to its domain.
5. Add each subelement of non-primitive domain with score 0 defined in table *XMLElementsNE*(b) into the superclasses list of the corresponding class.

EndAlgorithm 4.1

To understand the steps of Algorithm 4.1, we present more details with supporting examples.

Each complex type E in the object graph is translated into a class of the same name E in the object-oriented schema. In each "complexType" E , there is only one empty element, which includes several subelements. Those primitive and non-primitive subelements and their domains are mapped into class attributes with the same domains. Also, superclasses of the class are added to its superclasses list. Information related to three example classes is given next; only attributes, superclasses and subclasses are shown; functions are excluded for space limitation and because they are trivial. Each attribute satisfies encapsulation by having two corresponding functions, one to set its value and one to return its value.

PERSON class can be depicted as

$PERSON_{attributes} = \{ssn : integer; name : string; age : integer; sex : character; spouse : Person; nation : Country\}$

$PERSON_{superclasses} = []$

$PERSON_{subclasses} = \{Student, Staff, Secretary\}$

COUNTRY class can be depicted as

$COUNTRY_{attributes} = \{name : string; area : integer; population : integer\}$

$COUNTRY_{superclasses} = []$

$COUNTRY_{subclasses} = []$

STUDENT class can be depicted as

$STUDENT_{attributes} = \{StudentID : integer; gpa : real; student : Department; Takes : \{(course : Course; grade$


```
:string)}}
STUDENTsuperclasses = [PERSON ]
STUDENTsubclasses =ResearchAssistant
```

5 Summary and conclusions

In this paper, we considered the mapping between object-oriented database and XML. This turns into forward and backward mappings. For the forward mapping from object-oriented into XML, we first analyze the object-oriented database to construct the object graph, which is equivalent to the class hierarchy with all inheritance and nesting links indicated. Different scores are assigned to links in both graphs in order to differentiate inheritance from nesting links. Then we developed two algorithms to produce for the object graph a corresponding flat or nested XML schema. Here, it is worth noting that the inheritance is handled differently by the two algorithms. While the former resolves the inheritance using key and keyref, the latter expands the subclass element to include the content of the superclass element; the latter is a more natural way, but the former is easier to deal with if we need to transform the XML into the relational model, which does not support nesting. Finally, we handle the mapping of the object-oriented data into XML document(s). For the backward mapping from XML into object-oriented database, we first analyze flat and nested XML Schema to construct the object graph, which is equivalent to the class hierarchy with all inheritance and nesting links indicated. Nested XML schema complex types are directly mapped to the proposed candidate classes. User involvement is required to differentiate between the nested complex types and the inheritance ones that will be mapped as superclasses. Flat XML schema depends on key and keyref data to resolve the inheritance and the nesting. User involvement is required to decide for those complex types that do not have enough information in key and keyref data. Flat XML data is mapped into the same tables used for the nested XML data, so one algorithm was sufficient to handle both types of nested and flat XML schemas.

References

- [1] U. Ahmad, et al., "An Integrated Approach for Extraction of Objects from XML and Transformation to Heterogeneous Object Oriented Databases," *Proc. of ICEIS*, pp.445-449, 2003.
- [2] R. Alhajj, F. Polat and C. Yilmaz, "Views as First-Class Citizens in Object-Oriented Databases," *VLDB Journal*, Vol.14, No.2, pp.155-169, 2005.
- [3] R. Alhajj and A. Elnagar, "Incremental Materialization of Object-Oriented Views," *Data & Knowledge Engineering*, Vol.29, No.2, pp.121-145, Nov. 1998.
- [4] R. Alhajj and F. Polat, "Reusability and Schema Evolution in an Object-Oriented Query Model," *Proc. of the ASME European Conference on Systems Design and Applications*, France, pp.21-29, Jul. 1996.
- [5] R. Alhajj and F. Polat, "Proper Handling of Query Results Towards Maximizing Reusability in Object-Oriented Databases," *Information Sciences: An International Journal*, 107/1-4, pp.247-272, Jun. 1998.
- [6] R. Alhajj and M.E. Arkun, "A Formal Data Model and Object Algebra for Object-Oriented Databases," *Applied Mathematics and Computer Science*, Vol. 2, No. 1, pp. 49-63, 1992.
- [7] R. Alhajj and F. Polat, "Closure Maintenance in an Object-Oriented Query Model," *Proc. of the ACM International Conference on Information and Knowledge Management*, Maryland, pp.72-79, Nov. 1994.
- [8] R. Alhajj and M.E. Arkun, "A Query Model for Object-Oriented Database Systems," *Proc. of the 9th IEEE International Conference on Data Engineering*, Vienna, pp.163-172, Apr. 1993.
- [9] R. Alhajj and F. Polat, "Database Reverse Engineering," *Proc. of the 14th International Symposium on Computer and Information Sciences*, Kusadasi, Oct. 1999.
- [10] R. Alhajj, "Documenting Legacy Relational Databases," *Proc. of the International Workshop on Reverse Engineering of Information Systems, in conjunction with the International Conference on Conceptual Modeling*, Lecture Notes in Computer Science, Springer-Verlag, Paris, pp.161-172, Nov. 1999.
- [11] M. Andersson, "Extracting an Entity-Relationship Schema from a Relational Database through Reverse Engineering," *Proc. of the 13th International Conference on Entity-Relationship Approach*, Manchester, pp.403-419, Dec. 1994.
- [12] R.G.G. Cattell, et al., *The Object Database Standard: ODMG-93*, Morgan Kaufmann, 1994.
- [13] R.G.G. Cattell, *Object Data Management: Object-Oriented and Extended Relational Database Systems*, Addison-Wesley, 1994.
- [14] M. Carey, et al, "XPERATO: Publishing Object-Relational Data as XML," *Proc. of the International Workshop on Web and Databases*, May 2000.
- [15] J. Cheng and J. Xu, *IBM DB2 XML Extender*, IBM Silcom Valley, February, 2000.
- [16] T.-S. Chung, S. Park, S.-Y. Han, and H.-J. Kim. "Extracting Object-Oriented Database Schemas from XML DTDs Using Inheritance," *Proc. of the International Conference on Electronic Commerce and Web Technologies*, pp.49-59, 2001.

- [17] M.F. Fernandez, W.C. Tan, and D. Suci, “SilkRoute: Trading between Relational and XML,” *Proc. of the International Conference on World Wide Web*, May 2000.
- [18] J. Fong, F. Pang, and C. Bloor, “Converting Relational Database into XML Document,” *Proc. of the International Workshop on Electronic Business Hubs*, pp61-65, Sep. 2001.
- [19] T. Johansson and R. Heggbredda. “Importing XML Schema into an Object-Oriented Database Mediator System,” In Uppsala Master’s Theses in Computing Science no. 260 Examensarbete DV3 20 p, 2004-01-12, ISSN 1100-1836, 2003.
- [20] D. Lee, et al, “Nesting based Relational-to-XML Schema Translation,” *Proc. of the International Workshop on Web and Databases*, May 2001.
- [21] C. Liu, M. W. Vincent, J. Liu, and M. Guo, *A Virtual XML Database Engine for Relational Databases*, Springer-Verlag, 2003.
- [22] A. Lo, R. Alhajj and K. Barker, “VIREX: Visual Relational to XML Conversion Tool,” *Visual Languages and Computing*, Vol.17, No.1, pp.25-45, 2006.
- [23] T. Naser, K. Kianmehr, R. Alhajj and M. J. Ridley, “Transforming Object-Oriented Database into XML,” *Proc. of IEEE IRI*, pp.600-605, Aug. 2007.
- [24] T. Naser, R. Alhajj and M. J. Ridley, “Reengineering XML into Object-Oriented Database,” *Proc. of IEEE IRI*, Jul. 2008.
- [25] D. Toth and M. Valenta, “Using Object and Object-Oriented Technologies for XML-native Database Systems,” *Proc. of the DATESO Annual International Workshop on Databases, Texts, Specifications and Objects*, 2006.
- [26] V. Turau, “Making Legacy Data Accessible for XML applications,” 1999, <http://www.informatik.fh-wiesbaden.de/~turau/ps/legacy.pdf>.
- [27] C. Wang, A. Lo, R. Alhajj, and K. Barker, “Converting Legacy Relational Database into XML Database through Reverse Engineering,” *Proc. of ICEIS*, 2004.
- [28] “Extensible Markup Language (XML) 1.0 (Fourth Edition).” W3C Recommendation 16 August 2006, edited in place 29 Sept 2006

Rate Distortion Manifolds as Model Spaces for Cognitive Information

James F. Glazebrook
 Department of Mathematics and Computer Science
 Eastern Illinois University
 600 Lincoln Avenue
 Charleston IL 61920–3099, USA
 E-mail: jfglazebrook@eiu.edu

Rodrick Wallace
 Division of Epidemiology
 The New York State Psychiatric Institute
 Box 47, 1051 Riverside Drive
 New York NY 10032, USA
 E-mail: wallace@pi.cpmc.columbia.edu

Keywords: global workspace, entropy, cognition, rate distortion function, giant component, groupoid, stochastic resonance.

Received: December 18, 2007

The rate distortion manifold is considered as a carrier for elements of the theory of information proposed by C. E. Shannon combined with the semantic precepts of F. Dretske's theory of communication. This type of information space was suggested by R. Wallace as a possible geometric–topological descriptive model for incorporating a dynamic information based treatment of the Global Workspace theory of B. Baars. We outline a more formal mathematical description for this class of information space and further clarify its structural content and overall interpretation within prospectively a broad range of cognitive situations that apply to individuals, human institutions, distributed cognition and massively parallel intelligent machine design.

Povzetek: Predstavljena je formalna definicija prostora za opisovanje kognitivnih procesov.

1 Introduction

The concept of an *information space* seems to have various definitions and interpretations within the mathematical and life sciences literature. The quest for suitable models of cognitive processes in the large scale is likely to involve an eclectic array of techniques drawn from topology, geometry and category theory. For it appears sensible, if not absolutely necessary, to introduce structures that are comparatively weaker than the conventional 'smooth manifold' approach as it is often employed in mainstream dynamical systems, statistical inference, automata, and pattern/speech/facial recognition. Whereas in such situations we do not dispute the usefulness of Finsler or Riemannian structures (e.g. the well-known Fisher information metric about which much has been written), their comparative rigidity does not allow the flexibility of topological modeling which is necessary for the more organismic-like nature of cells of information as they function in the *local-to-global processing* of cognitive mechanisms. The quest then is to provide a descriptive framework of cognitive-interactive modules which are 'organisms' in their own right, possessing their own internal dynamics, semantic language, channels of communication and their own 'im-

immune' systems. This affords them further richness of content compared to the schemata of classical neural networks, or to the over-simplified flow boxes of cybernetic processes and other stock-in-trade tools of traditional AI.

Wallace [125] has brought together the fundamental ideas of F. Dretske's semantic communication theory [39] with the Global Workspace theory of B. Baars [12] within a framework governed by the acclaimed theory of information of C. E. Shannon [32, 112] originating with necessary conditions gauging the reliability of a source entropy rate relative to a channel capacity. Subsequent ideas are blended in with the mathematical models for statistical mechanics/chemical thermodynamics as introduced by L. Onsager along with K. Wilson's theory of renormalization. In this way the overall treatment incorporates several notable examples of 20th century mathematical-physical creation. Motivated by the main results of Shannon's theory, Wallace [125] introduced the concept of a *rate distortion manifold* as a descriptive model for analyzing a schemata of information-based cognition in a range of contexts (such as e.g. psycho-social epidemics, sleep patterns, obesity, stress related illnesses, inattentive blindness and the language of the immune system [126, 127, 128, 129]). A

prime ingredient used here is the concept of the *rate distortion function* $R(D)$ that characterizes the minimum rate of information needed to reproduce a source message within a specified fidelity D .

Although the basis for having a manifold atlas topology had been suggested in these works, there remains the task for a comprehensive description of a rate distortion manifold in more formal geometric and topological terms as a means of understanding and better representing the intrinsic dynamics underlying a wide range of cognitive processes. These may involve direct comparisons between individual/institutional cognition on the one hand, and distributed cognition/massively parallel computation on the other. In this way we may gain some insight into how failures in one category induce corresponding causes of failures in another.

Here we propose such a formal and conceptual treatment of the rate distortion manifold within a stretch of mathematical ideas, and at the same time provide a discussion of how these ideas are applicable to a variety of cognitive situations. The ‘weaker’ structures, incumbent to an extent on the abstract principles of the topology of manifolds and the theory of categories, are preferred over the less flexible structures which appear in the examples mentioned previously. More specifically, we consider certain topological spaces which are intrinsically path spaces over an alphabetical–coding system in the Dretske semantics of communication and adhering to the Shannon theory [125]. But viewed as particular (rate distortion) metric spaces with length structure, they are nevertheless conducive to admitting manifold/atlas topologies in a general sense, and moreover, may enjoy a sufficiently weaker notion of ‘differentiability’ beyond the framework of classical differential calculus. Putting it another way, we propose a rate distortion manifold to be a general ‘cognitive information space’ adapted to, or designed towards, analyzing a particular cognitive situation.. Such a space admits a manifold/atlas topology to encompass its framework, and which serves as a *descriptive ‘cell’ or ‘organism’ of cognitive information* gauged as such by certain principles of information/entropy and statistical physics.

By ‘cognitive’, we mean pertaining to cognition at large; that is, in relationship to the neurosciences, cross–cultural psycho–sociological and epidemiological/immunological phenomena (both *distributed* and *institutional*). In the context of the Global Neuronal Workspace [12], such an ‘organism’ is a constituent module of a ‘broadcasting system’ that cooperates or competes within a specified hierarchy of contexts. Symmetry-breaking within the underlying information network engenders a phase transition, and thus the complexity of information increases. The cultural models of [125, 135] apply not only to individual cognition, but also to the large–scale cognition of socio–political–environmental mechanisms which inevitably shape the vital drive of the former.

Underlying our proposed information spaces are cellular and combinatorial structures as, for instance, provided

by the general topological notion of a *CW–complex* or of a *simplex* (obtained via a simplicial decomposition of a given topological space) which provides a route into the graph–theoretic concepts of mainstream network analysis. There is yet another categorical twist. We propose that the structures of our information spaces be richer and deeper than abstract categories, but it is even more meaningful if the actual objects of a given category are the information spaces themselves. The other slant is that these spaces may be reasonably viewed in terms of ‘small’ categories with invertible morphisms, that is, they can be realized via path equivalence as *groupoids* and therefore can be treated quite appropriately within the framework of the algebraic topology of groupoids, their actions and atlases of such (see e.g. [16, 21, 143] and Appendix I of this paper). Applicable to a broad range of cognitive models, this is a central idea already advanced in [125](cf [60]) and can be compared with recent approaches to biological and neural network systems as adopted by a number of other authors (see e.g. [15, 61, 66, 67, 118] and the many references therein). The cognitive ‘cells’ or ‘modules’ of information founded on rate distortion manifolds are also relevant in the modeling of *autopoietic* and related systems [93](cf [34, 60, 70]) which will be discussed in a later section. Future directions (which we will address later) are likely to involve the more ambitious tasks of dealing with informational processes and rate distortion theory within the context of such general categorical concepts as *groupoid atlases*, *stacks in groupoids* and the general concept of a *topos*.

2 Information spaces for general cognition

2.1 Characteristics of a cognitive information space

Let us mention that Chalmers [25](Chapt. 8), for instance, has suggested a possible framework for linking the processes of the physical and phenomenal worlds in terms of a conceptually–based information science cast within the mathematical methods of geometry and topology. This plan has a significant overlap with the development in [125] and the more formal treatment described herein. Such an approach seems imperative for advancing the Workspace structure within the Theater of Consciousness [12] by means of semantic information and dynamical processes. Moreover, the proximity to our formal description of a rate distortion manifold as an information space, will eventually become relevant. We rephrase the overall requirements within the scope of our proposed development:

1. The model assists in addressing the phenomenological aspect (states of context) in relationship to physical reality.
2. An information space could be viewed as in part rep-

representing a ‘causal pathway’ embedded within some culture, but included are semantic, dynamic principles seeking to incorporate states of experience, properties often lacking in traditional cognitive theories.

3. We consider a model based upon the structure of contexts and language of thought within fluctuating paths of information, not entirely in the discrete sense, but possibly within a suitable notion of a ‘continuum limit’, a ‘manifold–atlas topology’, or a ‘dynamic categorical process’, as examples of possible working environments.
4. The manifold–atlas topology can be coupled with an information based, (weakly) stochastic structure where essential distinctions can be represented in terms of such properties as homotopy and diffeomorphism type, homology, curvature, etc. Thus we turn to state of the art techniques of differential geometry/topology, category theory, on the one hand, and on the other, to combine these techniques with those of information theory, stochastic processes and statistical mechanics. This spread of ideas reflects upon the eclectic framework proposed for extending the traditionally acceptable descriptive methods for studying cognitive processes and their emergence through orders of complexity.
5. By introducing simplicial methods to analyze the underlying combinatorial structure of the manifold, we may recover graph-theoretic models as suited to the navigation through various types of information highways [2], systems of coding, symbolic dynamics [89] and complexity [88].
6. The approach aims at constructing a geometrical/topological carrier for the Shannon information theorems about which the rate distortion manifold is formulated, thus leading to a framework for interacting cognitive modules upon which the prevailing cultural environment inevitably writes its image. Essential here is that the manifold accommodates a homology between the corresponding informational laws of asymptotic probability and certain thermodynamic limits of statistical mechanics. In this way, altered states of cognitive processes can be seen to be caused by phase transitions analogous to how the latter can induce sharp transformations between one thermodynamic state to another.

A wide range of descriptive possibilities are likely. Of these we could view the structure of awareness and experience as represented within the structure of an information space with phenomenal states. Conversely, such a predictive representation may feed its way back into the cognitive system with the enhanced prospects of obtaining an improved model which may lead to the eventual solution of a given problem.

A limitation of classical information theory is that it was not preoccupied with semantics. The theory was destined originally for the testing ground of noisy telephone exchanges – some time before ‘The Brain is a Noisy Processor’ became a standard assumption. Information in Shannon’s theory evolved essentially within a combinatorial/probabilistic framework for representing how states are manifest within an information space. Its main constituents include:

- a) Application of the asymptotic limit theorems of probability theory.
- b) Mutual information.
- c) The Shannon Coding Theorem (fixing signal and oscillator) assumes an optimal coding scheme involving noise so that the rate of error–free output of the signal will attain some positive value.

2.2 The Global Workspace

A principal aim is to apply rate distortion manifolds as descriptive features of the Global Workspace theory of [12, 14]. The general dictum goes as follows [14]:

- (1) The brain can be viewed as a collection of distributed specialized networks (processors).
- (2) Consciousness is associated with a Global Workspace in the brain – a fleeting memory capacity whose focal contents are widely distributed (broadcast) to many unconscious specialized networks.
- (3) Conversely, a Global Workspace can serve to integrate many competing and cooperating input networks.
- (4) Some unconscious networks called ‘contexts’, shape conscious contents. For example, unconscious parietal maps modulate visual feature cells that underlie the perception of color in the ventral stream.
- (5) Such contents work together jointly to constrain conscious events.
- (6) Motives and emotions can be viewed as goal contexts.
- (7) Executive functions work as hierarchies of goal contexts.

Recent research [37, 38] has enhanced the validity of this model. As pointed out in [126], the special properties of representing embedding and interpenetrating contexts provide a framework for understanding the synergism of consciousness and mental disorders in humans within a socio–cultural context. This framework bears startling analogies with the institutional cognition of epidemics versus the public health sector as a phenomenon of disorder of information [129, 137, 140]. Indeed, institutions such as the latter may themselves function within their respective cultural environments as ‘distributed’ cognitive systems having their own sovereign mechanisms, implicitly

different from that of humans but nevertheless influencing the degree of effectiveness of human involvement [70], a situation closely in tune with the organisms of environmental autopoietic systems [93] (see §9.4). But these large-scale cognitive mechanisms, although not constrained so much by biological evolution, are certainly prone to analogous cognitive disorders such as environmental (psycho-social) stress, inattentive blindness, (social) network failure and many other ailments that plague human society.

As incorporating fundamental geometric techniques, rate distortion manifolds each possessing characteristic topological and geometric properties, along with their own internal dynamics, are thus proposed as descriptive cells within this blueprint for cognition. As abstract topological manifolds they are designed to model the shape and flow of information that can be adapted to analyze a broad range of cognitive situations. The next stage is to outline a more specific mathematical description of their structure.

3 Towards a rate distortion manifold

3.1 Manifold-atlas topology

The standard concept of a ‘differentiable manifold’ as to be found in e.g. [1, 87], is of long-standing importance in geometry and physics. However, for the sake of the more flexible structures as sought after, we need to have a handle on an even more general concept, namely that of an *atlas-manifold topology* (such as to be found in e.g. [21]).

The idea is to start by defining a weaker notion of ‘function’ valid for set-valued mappings. Let A and B be sets and consider a triple (A, B, F) where $F \subset A \times B$, with the property that if $(a, b), (a, b') \in F$, then $b = b'$. Such a triple is called a *partial function* between A and B , denoted $f : A \rightarrow B$, and written $f(a) = b$. The domain of f is the set of $a \in A$ such that $f(a)$ is defined. The concepts of composition, continuity, etc. apply in accordance their usual topological definitions. The domain of f can be any subset of A , and if B is a some scalar field such as \mathbb{R} or \mathbb{C} , then the definition reduces to the standard one for that of a function.

Consider a set A and a family $\{A_\lambda\}_{\lambda \in \Lambda}$ of topological spaces, together with a partial function $f_\lambda : A \rightarrow A_\lambda$, for each $\lambda \in \Lambda$. A topology \mathcal{T} on A is said to be *initial* with respect to $\{f_\lambda\}$ if for any topological space B , a partial function $k : B \rightarrow A_{\mathcal{T}}$ is continuous if and only if the composition $f_\lambda \circ k : B \rightarrow A_\lambda$ is continuous. Such a topology on A is the coarsest of topologies such that each $f_\lambda : A_{\mathcal{T}} \rightarrow A_\lambda$, is continuous.

Let E be a topological space and let M be a set. An E -chart on M is an injective partial function $\varphi : M \rightarrow E$ whose image is open in E . For some indexing set \mathcal{I} , an E -atlas for M consists of a family $\mathcal{A} = \{\varphi_\alpha\}_{\alpha \in \mathcal{I}}$ of E -charts for M such that if $\varphi_\alpha, \varphi_\beta : M \rightarrow E$ are charts in \mathcal{A} , then the composition $\varphi_\beta \circ \varphi_\alpha^{-1} : E \rightarrow E$, is continuous.

Suppose then we are given such an E -atlas \mathcal{A} , and let M have the initial topology with respect to all E -charts in

\mathcal{A} . Then $\varphi_\alpha^{-1} : E \rightarrow M$ is continuous, since $\varphi_\beta \circ \varphi_\alpha^{-1} : E \rightarrow E$, is continuous for all $\beta \in \mathcal{A}$. Therefore, φ_α maps its domain homeomorphically to its image. We may call E the *model space* of the atlas in accordance with the terminology in the case where M is a manifold in the more concrete sense, and when E is some suitable vector space (which could be infinite dimensional).

In order to realize a suitable cognitive information space, we probe beyond some of the typical manifold structures of information geometry and so the above abstraction has potential value. Information geometry can involve using parametric and non-parametric probability densities in order to construct appropriate statistical manifolds for inference. In the parametric case, befitting a Fisher metric structure say, the manifold can be treated from the point of view of a Euclidean topology, whereas in the non-parametric case (useful for e.g. perception/recognition as in [122]), other topologies have to be considered leading to infinite dimensional manifolds modeled, say, on spaces of operators (such as projections) and which include special Banach spaces with differing topologies (e.g. exponential convergence as introduced in [101] and applied in [145]).

In this respect the model space E might be taken as a carrier space for operator-valued probability densities, suitable say, for dealing with ‘sharp’ or ‘fuzzy’ stochastic processes (as in e.g. [64]). Manifolds for statistical inference, stochastic processes and those serving as a descriptive mechanism for modeling the various information highways are seemingly too rigid in structure for effectively describing cognitive cells. In the latter case, we consider these as significantly influenced by linguistic and cultural factors. Additionally, there are potentially useful structures weaker than the standard manifold/atlas, such as that of an *orbifold atlas* which would accommodate certain types of singularities [96] (see Appendix I §11.2). More generally, there is the notion of a *groupoid atlas* [16] (see §7.4) which incorporates groupoid actions and thus may be viewed as an abstract dynamical system in its own way (see §7.4).

3.2 The information space (X, s_X)

More specifically, suppose $E = E^\Gamma$ is a high dimensional state space modeled on some ‘alphabetical/coding/syntactical’ structure denoted Γ . This is instrumental for a semantic base-model following the dictum of F. Dretske [39, 125]. In mathematical terms, we grant the possibility that E as a kind of *state space* may be formally structured as a vector space which may possess certain properties such as local convexity, etc. For instance, E could be taken to be the underlying vector space of a general *events algebra* in the sense of [64]. We also leave open the possibility that E is endowed with some norm denoted $\|\cdot\|$, although we may not always insist on this property.

We proceed to consider a set $X \subset E$, where points $x \in X$ correspond to paths of convoluted signals; typically, $x = (a_0, a_1, \dots, a_n, \dots)$ where a_k represents some functional composition of internal and external signals. In this respect

X could be deemed to be a ‘weak path space’ over E . It is quite possible that X could be considered as having a filtered or cellular structure (as will be described below). The path space (X, s_X) with model (or atlas) space E , is considered as a metric space with metric $s_X \equiv d(x, \hat{x})$ induced by a distortion measure d (see §3.4). This measure then leads to defining a rate distortion function complying with Shannon’s theorem (see §3.4 and §4.3 later).

Following [125], suppose we take a decision oscillator generating an output as given by a set valued (partial) function $h : X \rightarrow B$. For instance, we set $B = B_0 \cup B_1$ where

$$h(x) \in B_0 \equiv b_0, \dots, b_k, \tag{3.1}$$

if the pattern is not recognized, and

$$h(x) \in B_1 \equiv b_k, \dots, b_m, \tag{3.2}$$

if the pattern is recognized.

The set B is prospectively one that is highly extensible and could be viewed as the underlying set of a suitably constructed algebra of responses or events. Also, the fact that higher order cognitive decisions and several options of response along a given path are likely to be necessary, suggests further intrinsic properties needed for sets of the type B_0 and B_1 .

Remark 3.1. Note that patterns may well undergo a filtering in stages of recognition. Thus a generalization is to suppose that B_0 and B_1 admit countable filtrations of the sort:

$$\begin{aligned} B_0 &= B_0^0 \subseteq B_0^1 \subseteq B_0^2 \subseteq \dots \\ B_1 &= B_1^0 \subseteq B_1^1 \subseteq B_1^2 \subseteq \dots \end{aligned}$$

where at level j we have set $B_0^j \equiv b_0^j, \dots, b_k^j$, and $B_1^j \equiv b_{k+1}^j, \dots, b_m^j$.

3.3 The Shannon entropy

Shannon conceived of *entropy* as a measure H of the capacity of a communications system to transmit information. The idea was to directly tie a given response rate $r(t)$ to a function of the probability of achieving $r(t)$. In a more specific way we will recall below some of the basic results of the theory in terms of *meaningful paths*.

For each $n \in \mathbb{N}$, let $N(n)$ denote the number of paths of length n beginning with a particular a_0 with $h(a_0) \in B_0$, and leading to the condition that $h(x) \in B_1$. We call such paths *meaningful* and, for cognitive reasons, regard $N(n)$ to be much less than the number of all paths of length n . Further, we assume that the limit

$$H \equiv \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n}, \tag{3.3}$$

exists, is finite, and is independent of the path x . Such a cognitive process is then said to be *ergodic*. The non-ergodic case (more pertinent to cognition) will be discussed later.

Relative to the path space (X, s_X) , we define a corresponding ergodic information source \mathbf{X} with stochastic variables \mathbf{X}_j having joint and conditional probabilities $P(a_0, \dots, a_n)$ and $P(a_n | a_0, \dots, a_{n-1})$ respectively, so that the joint and conditional Shannon probabilities may be defined and satisfy the relations [32]:

$$\begin{aligned} H[\mathbf{X}] &= \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n} \\ &= \lim_{n \rightarrow \infty} H(\mathbf{X}_n | \mathbf{X}_0, \dots, \mathbf{X}_{n-1}) \\ &= \lim_{n \rightarrow \infty} \frac{H(\mathbf{X}_0, \dots, \mathbf{X}_n)}{n+1}. \end{aligned} \tag{3.4}$$

Such an information source is considered to be *dual* to the ergodic process.

Remark 3.2. Technically, \mathbf{X} is taken to be an *adiabatically, piecewise stationary, ergodic* (APSE) information source (see e.g. [6, 32] and explanations relative to cognitive modules in [125, 135]).

The Shannon–McMillan theorem provides ‘a law of large numbers’ and permits the definition of uncertainties in terms of cross sectional sums of the form

$$H = - \sum_k P_k \log[P_k], \tag{3.5}$$

where the P_k are derived from a probability distribution and satisfy $\sum_k P_k = 1$. Different languages can be defined by different divisions of the total universe of possible responses into various pairs of sets B_0, B_1 above, or by insisting upon more one than response in B_1 along a path. Allocating the full set of possible responses into B_0, B_1 may necessitate engaging higher order cognitive decisions.

3.4 The Rate Distortion Theorem

Following [125], suppose we have an (ergodic) information source \mathbf{Y} with output from a particular alphabet generating sequences of the form

$$y^n = y_1, \dots, y_n \tag{3.6}$$

‘digitalized’ in some sense, and inducing a chain of ‘digitalized’ values

$$b^n = b_1, \dots, b_n \tag{3.7}$$

where the b -alphabet is considered more restricted than the y -alphabet. In this way, b^n is *deterministically retranslated* into a reproduction of the signal y^n . That is, each b^n is mapped onto a unique n -length, y -sequence in the alphabet of \mathbf{Y} :

$$b^n \rightarrow \hat{y}^n = \hat{y}_1, \dots, \hat{y}_n. \tag{3.8}$$

We remark that many y^n sequences may be mapped onto the same retranslation sequence \hat{y}^n , the set of which is denoted $\hat{\mathbf{Y}}$; this may be interpreted as a loss of information.

A distortion measure $d : \mathbf{Y} \times \widehat{\mathbf{Y}} \rightarrow \mathbb{R}^+$, between paths y^n and \hat{y}^n is defined as

$$d(y^n, \hat{y}^n) = \frac{1}{n} \sum_{j=1}^n d(y_j, \hat{y}_j), \quad (3.9)$$

for some suitable distance function d (such as the Hamming distance). Suppose that with each path $y^n \in \mathbf{Y}$ and each b^n -path retranslation $\hat{y}^n \in \widehat{\mathbf{Y}}$ into the y -language, we consider the associated individual, joint, and conditional probability distributions

$$p(y^n), p(\hat{y}^n), p(y^n|\hat{y}^n). \quad (3.10)$$

The average distortion is then defined to be

$$D = \sum_{y^n} p(y^n) d(y^n, \hat{y}^n). \quad (3.11)$$

For the corresponding strings \mathbf{Y} (incoming), $\widehat{\mathbf{Y}}$ (outgoing), the Shannon uncertainty rule is

$$\begin{aligned} I(\mathbf{Y}, \widehat{\mathbf{Y}}) &\equiv H(\mathbf{Y}) - H(\mathbf{Y}|\widehat{\mathbf{Y}}) \\ &= H(\mathbf{Y}) + H(\widehat{\mathbf{Y}}) - H(\mathbf{Y}, \widehat{\mathbf{Y}}). \end{aligned} \quad (3.12)$$

Definition 3.1. The information rate distortion function $R(D)$ for a source sequence \mathbf{Y} , a retranslated sequence $\widehat{\mathbf{Y}}$, along with a distortion measure $d : \mathbf{Y} \times \widehat{\mathbf{Y}} \rightarrow \mathbb{R}^+$, is defined as follows.

Let $\Upsilon = \sum_{(y, \hat{y})} p(y) p(y|\hat{y}) d(y, \hat{y})$. Then

$$R(D) = \sum_{p(y, \hat{y}) : \Upsilon \leq D} I(\mathbf{Y}, \widehat{\mathbf{Y}}). \quad (3.13)$$

To explain this notation, the minimization is over all conditional distributions $p(y|\hat{y})$, for which the joint distribution $p(y, \hat{y}) = p(y) p(y|\hat{y})$ satisfies average distortion less than or equal to D .

The Rate Distortion Theorem (see e.g. [32, 36]) states that $R(D)$ is the maximum achievable rate of information which does not exceed the distortion D .

These are some of the basic ingredients for considering the optimal rate of precise information transfer in relationship to channel capacity and to which extent noise is hazardous to the system. The path-space modeled rate distortion manifolds in question, are assumed to comply with the above theorem implicitly. Moreover, as far as cognition is concerned, [125] postulates a fundamental homology with thermodynamic processes, quite similar to how distortion and fidelity in network information can be studied involving techniques of statistical physics such as the Ising lattice and spin-glass networks in conjunction with the usual industry of error correcting and coding (for related work in this direction see e.g. [98, 89, 116]).

3.5 Channel capacity

The channel capacity is defined to be

$$C \equiv \max_{P(\mathbf{X})} I(\mathbf{X}|\mathbf{Y}), \quad (3.14)$$

subject to the subsidiary condition that $\sum P(\mathbf{X}) = 1$. This is a measure of the maximum transmission rate of information across a channel with the likelihood of error tending to zero. Effectively, the critical trick of the Shannon Coding Theorem for sending a message with arbitrarily small error along the channel \mathbf{Y} at any rate $R < C$, is to encode it in longer and longer ‘typical’ sequences of the stochastic variable \mathbf{X} ; that is, those sequences whose distribution of symbols approximates the probability distribution $P(\mathbf{X})$ above which maximizes C .

Thus for an information source X , the Shannon entropy $H(\mathbf{X})$ as given in (3.3), can be seen to satisfy for a given channel capacity C , the inequality

$$H(\mathbf{X}) \leq C. \quad (3.15)$$

If $S(n)$ is the number of such ‘typical’ sequences of length n , then $\log[S(n)] \approx nH(\mathbf{X})$. Some consideration shows that $S(n)$ is much less than the total number of possible messages of length n . Thus, as $n \rightarrow \infty$, only a vanishingly small fraction of all possible messages is meaningful in this sense. This observation, after some considerable development, is a principle that allows the Shannon Coding Theorem to work so well. In sum, the prescription is to encode messages in typical sequences, which are sent at very nearly the capacity of the channel. As the encoded messages become longer and longer, their maximum possible rate of transmission without error approaches channel capacity as a limit (for details see [32, 79, 112]).

Rate distortion manifolds may be characterized by a type of inversion of this procedure. Examples of noisy channels are telephone lines, optical wave guides and interplanetary plasmas around which a message is to be structured so as to attain an optimal error-free transmission rate. These examples are, relatively speaking, fixed on the timescale of most messages, as are most sociogeographic networks. Indeed, the capacity of a channel, is defined by varying the probability distribution of the ‘message’ process \mathbf{X} so as to maximize $I(\mathbf{X}|\mathbf{Y})$. For instance, suppose there is some message \mathbf{X} so critical that its probability distribution must remain fixed. The trick is to fix the distribution $P(\mathbf{X})$ but to modify the channel; that is, to tune it so as to maximize $I(\mathbf{X}|\mathbf{Y})$. The dual channel capacity C^* is then defined as

$$C^* \equiv \max_{P(\mathbf{Y}), P(\mathbf{Y}|\mathbf{X})} I(\mathbf{X}|\mathbf{Y}). \quad (3.16)$$

But

$$C^* = \max_{P(\mathbf{Y}), P(\mathbf{Y}|\mathbf{X})} I(\mathbf{Y}|\mathbf{X}), \quad (3.17)$$

since we have

$$\begin{aligned} I(\mathbf{X}|\mathbf{Y}) &= H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}) \\ &= I(\mathbf{Y}|\mathbf{X}). \end{aligned} \quad (3.18)$$

Thus, in a purely formal mathematical sense, the message transmits the channel, and there will indeed be, according to the Shannon Coding Theorem, a channel distribution $P(\mathbf{Y})$ which maximizes C^* . Variations on this theme are realized in [134](see also [135]).

3.6 Noise in the system

In many sensory and cognitive systems not all noise corrupts the processing of information. Indeed, adding noise under the right circumstances may actually amplify and enhance the transmission, and may even reduce randomness in the system by the presence of *stochastic resonance*. In biological circumstances, the effect is often detected in large ion channels in the presence of stochastic processes, such as in the Hodgkin–Huxley model for instance [57]. Noise itself is not without its own peculiar ‘linguistics’ and semantic coding. Standard martingale analysis coupled with stochastic resonance reveals in [129] that the noise of a socio–economic structure (in the form, say, of misguided or regressive social policies), is most likely than not a major catalyst for the spreading of endemic illnesses, psycho–social disorders, therapeutic failure, inadequate public health services, and the deterioration of urban residential districts, as much as these factors influence each other [137, 138, 139].

3.7 A canonical model (M, s_M)

In some cases it will be necessary to project to a manifold model for the path space (X, s_X) based on this local description, but to one that is inherently less complicated and more conducive to standard topological/geometrical techniques. So let us proceed to define a topological space M which can be associated to X via a suitable map. Initially, we can grant M the structure of a metric space with a distortion measure s_M as induced by s_X in a sense to be made precise. We may assume that M admits an E -atlas manifold topology with a system of E -charts $\{(V, \varphi_V)\}$ while thinking of E as a suitable state space as above. Several possible ways to proceed are discussed below. To an extent (M, s_M) could be viewed as a more structured, simplified information space serving as a *canonical model* for the path space (X, s_X) (but as pointed out earlier, we do not insist on the speciality of Finsler or Riemannian spaces). Thus we consider a procedure similar to a dimensional reduction. We wish then to specify a projection map

$$\Pi : X \longrightarrow M, \tag{3.19}$$

with suitable properties, such as surjective, Lipschitz, etc., which will be outlined below. To an extent this will reflect the nature of information sources in (X, s_X) , be they ergodic, or non–ergodic.

Remark 3.3. The use of the the term ‘canonical’ is similar in spirit to how intricate and chaotic systems of neural networks can be transformed into certain blueprints (canonical models) representing the dynamics of reduced systems of differential equations, and which can thus be studied with the standard techniques of differentiable dynamical systems theory (as in e.g. [72]).

3.8 Length space structures

Hypothesizing a class of ‘admissible paths’ in (M, s_M) leads to giving the latter a *length space* in the sense of [24]. In our case, the admissible paths are to be considered as ‘meaningful’ in the sense introduced below (and for which M is considered as a locally path connected space).

To see how a length structure induced by the metric s_M arises, let $\gamma : [a, b] \longrightarrow M$ be a (continuous) path in M and choose a partition \mathcal{J} of the interval $[a, b]$, that is, a finite collection of points $\mathcal{J} = \{y_0, \dots, y_N\}$ such that

$$a = y_0 \leq y_1 \leq y_2 \leq \dots \leq y_N = b. \tag{3.20}$$

We can define the length of γ with respect to the metric s_M as the supremum of the sums over all partitions \mathcal{J} :

$$L(\gamma) = L_{s_M}(\gamma) := \sup_{\mathcal{J}} \sum_{i=1}^N s_M(\gamma(y_{i-1}), \gamma(y_i)). \tag{3.21}$$

The length structure induced by s_M can then be specified in terms of: a) all continuous paths parametrized by closed intervals are admissible, and b) the length is given by the function L in (3.21).

Consequently, we can draw upon generalizations of some traditional (but elementary) differential–geometric concepts in terms of a length structure. For instance, a curve

$$\gamma : [a, b] \longrightarrow M, \tag{3.22}$$

is said to be *rectifiable* if its length is finite, and a *shortest path* if its length is minimal among curves with the same endpoint, that is, $L(\gamma_1) \geq L(\gamma)$ for any curve γ_1 connecting $\gamma(a)$ and $\gamma(b)$. In particular, a curve $\gamma : I \longrightarrow M$ is said to be a *geodesic* if for every $t \in I$, there exists an interval J containing a neighborhood of t in I , such that $\gamma|_J$ is a shortest path. We remark that the concept of ‘geodesic’ can also be formulated in the context of graphs and networks (to be discussed later).

The metric space (M, s_M) is said to be *complete* if there exists a shortest path between two languages A, A' , and said to be (locally) *homogeneous* if for every A, A' , there exists a (local) isometry $\mathfrak{J} : M \longrightarrow M$, such that $\mathfrak{J}(A) = A'$. Other possible length space structures could be considered thus allowing the flexibility of going beyond the traditional Finsler and Riemannian structures which are common frameworks for inference and stochastic processes.

Remark 3.4. Suppose in M we have an admissible class \mathcal{A} of curves $\{\gamma(t)\}$. For $V \subset M$ an open set, suppose there is defined a nonnegative homogeneous function $\tilde{F} = \tilde{F}(x, v)$, where $x, v \in V$, that can be integrated over curves

$$\gamma = \gamma(t) : [a, b] \longrightarrow M, \tag{3.23}$$

in \mathcal{A} invariant under reparametrization. The homogeneity condition implies the relationship $\tilde{F}(x, kv) = |k|\tilde{F}(x, v)$. For $\gamma \in \mathcal{A}$, the length between $A = \gamma(a)$ and $\hat{A} = \gamma(b)$, may then be defined by

$$\ell_\gamma(A, \hat{A}) = \ell(\gamma, a, b) = \int_a^b \tilde{F}(\gamma(t), \gamma'(t)) dt. \tag{3.24}$$

Observe that we have not required $\tilde{F}(\cdot, \cdot)$ to be a norm, or even that \tilde{F} is convex or symmetric, so the length space structure of M is weaker than that of a Finslerian structure (see e.g. [24]). A metric on M can be defined in the usual way by

$$s_M(A, \hat{A}) = \inf\{\ell_\gamma(A, \hat{A}) : \gamma(a) = A, \gamma(b) = \hat{A}\}. \tag{3.25}$$

Given a local isometry $\Pi : X \rightarrow M$, this metric may then be assumed to agree locally with s_X on pulling-back under Π .

Once M is endowed with an atlas-manifold topology we can then postulate M as a CW-complex, that is, a space constructed from a collection of points via the successive attachment of cells. The topology then is *weak*, meaning that a set in M is closed if and only if its intersection with every cell is closed. Viewed as a CW-complex, M then has the same homotopy type of a simplicial complex which affords further considerations particularly when reducing matters to a skeletal-like, graph-theoretic analysis.

Remark 3.5. The (path) space X may indeed be very complex in its structure of informational data, whereas the canonical model M (possibly via dimensional reduction of the former) by dint of its manifold structure, is expected to be more conducive to geometric analysis as may be the case for nonlinear optimization. Such principles apply in work concerning imaging-recognition data as in [121] where multi-imaging ‘noisy’ data in a carrier akin to X is analyzed (such as with regard to pixel intensity) and is projected to a local metric structure via maps such as Π . The convergence of subsequent ‘data manifolds’ in reduced dimensions may then yield the exact model for a solution space. In [121] examples include intricate ‘Swiss Roll’ data spaces which are dimensionally reduced to some convex region in \mathbb{R}^n . Here geodesic distances between data points are calculated, and typically one wants to minimize a cost function based on an operational norm of data differences. Such examples (re. pixel intensity) do not generally fit into the context of rate distortion theory and semantic communication, and for our purposes further considerations are clearly necessary. These we will proceed to discuss in the following.

3.9 Stationary ergodic information sources

For a rate distortion manifold, an alternative procedure is to consider stationary ergodic information sources, although not all cognitive processes are expected to be of this type. Here the Khinchin’s E-property [79](p. 74) is evoked. Under the ergodicity assumption, the path space (X, s_X) can be partitioned into high and low probability subsets, $X = X_h \cup X_\ell$.

The projection map $\Pi : X \rightarrow M$ can be specified as follows. Each equivalence class of paths in the appropriate space is identified with its associated language-of-thought characterized by a stationary ergodic information source

having a source uncertainty $H(A)$, where A is taken as the language having a set of paths Ax . Thus for $x \in X$, we define the projection Π via $x \mapsto \Pi(x) = A$ (the language having a set of paths Ax). There are other possible variations on this theme.

Note also that we have restrictions

$$\begin{aligned} \Pi|_{X_h} : X_h &\rightarrow M \\ \Pi|_{X_\ell} : X_\ell &\rightarrow M. \end{aligned} \tag{3.26}$$

For each $A \in M$, let $U \subset M$ be an open set consisting of approximately similar languages near A . For all pairs of languages $A, \hat{A} \in U$, let us suppose that we have available a suitable metric $s_M(A, \hat{A})$ induced by the path space metric s_X . As an alternative to the metric in (3.25), we may choose a metric of the form [134]

$$s_M(A, \hat{A}) \equiv \left| \int_{A, \hat{A}} d(Ax, \hat{A}x) - \int_{A, A} d(Ax, Ax) \right|, \tag{3.27}$$

where Ax and $\hat{A}x$ are paths in the languages A, \hat{A} respectively, d is the distortion measure, and the second term is a ‘self-distance’ for the language A , such that $s_M(A, A) = 0, s_M(A, \hat{A}) > 0, A \neq \hat{A}$.

Since choosing stationary ergodic sources presents a different scenario to the non-ergodic sources, we may rethink the appropriate properties assumed by the projection Π . Possibilities might include: the projection

$$\Pi : (X, s_X) \rightarrow (M, s_M), \tag{3.28}$$

is a local isometry, or, Π is *Lipschitz*, meaning there exists a constant $C > 0$, such that

$$s_M(\Pi(x), \Pi(\hat{x})) \leq C s_X(x, \hat{x}). \tag{3.29}$$

Next, for each $A \in U$, we consider a source uncertainty $H(A)$, such that the *information source derivative* $\nabla_s H(A)$ is defined to be

$$\nabla_s H(A) \equiv \lim_{s \rightarrow 0} \frac{H(\hat{A}) - H(A)}{s(\hat{A}, A)}, \tag{3.30}$$

when this limit exists and is finite. A number of concepts follow from the basic principles of calculus, such as the *logarithmic derivative*

$$\nabla_s(\log H(A)) = \frac{\nabla_s H(A)}{H(A)}, \quad (H(A) \neq 0) \tag{3.31}$$

a measure of the relative rate of change of the source uncertainty through language of thought.

Prior to admitting an atlas-manifold topology on M , we remark that the metric s_M defined in (3.25) appears to be *intrinsic* in the sense that it arises from a supposed length structure on M induced by the above integration over languages, and not by the restriction of a metric on some ambient space.

3.10 Non-ergodic information sources

Non-ergodic information sources are likely to be favorable options for the purpose of understanding more complex cognitive processes. Suppose here the path space X consists of length n high probability paths $x_n \rightarrow x$ (as $n \rightarrow \infty$) that correspond to *non-ergodic* information sources. Let $N(n)$ be the number of high probability ‘grammatical’ and ‘syntactical’ paths of length n having $h(a_0) \in B_0$, and leading to $h(x) \in B_1$. Such paths are called *meaningful* where once more the limit

$$H \equiv \lim_{n \rightarrow \infty} \frac{\log N(n)}{n}, \tag{3.32}$$

exists, but is now generally taken to be path *dependent*.

We have a partial function $h : X \rightarrow B$, where B denotes a set of pattern responses for which given $x_n \rightarrow x$, we have

$$\lim_{n \rightarrow \infty} h(x_n) = h(x). \tag{3.33}$$

For all $x \in X$, we take an open set $U \subset X$ such that for all such $x \in U$, the following conditions hold [125]:

- (1) For all paths $\hat{x}_n \rightarrow \hat{x} \in U$, a distortion measure $s_n \equiv d_U(x_n, \hat{x}_n)$ exists.
- (2) For each path $x_n \rightarrow x$ in U , there exists a pathwise invariant function $h(x_n) \rightarrow h(x)$ [79].
- (3) A function $F_U(s_n, n) \equiv f_n \rightarrow f$ exists, such as for example

$$f_n = s_n, \frac{\log[s_n]}{n}, \text{ or } \frac{s_n}{n}. \tag{3.34}$$

- (4) The limit

$$\lim_{n \rightarrow \infty} \frac{h(x_n) - h(\hat{x}_n)}{f_n} \equiv \nabla_F h|_x, \tag{3.35}$$

exists and is finite.

In a similar way to stationary ergodic sources, we may consider introducing (M, s_M) as a more finely structured information space corresponding to a path space (X, s_X) , using a dimensional reduction procedure via a projection map

$$\Pi : X \rightarrow M, \tag{3.36}$$

with properties such as e.g. a surjective, local isometry, or Lipschitz, etc. that suitably projects the informational architecture on X to that on M .

Again, one might consider several options for the space M which would allow a weaker than Finsler structure. With regards to Remark 3.4, let us consider defined on each open set $V \subset M$, a nonnegative homogeneous function of two variables

$$\begin{aligned} \tilde{F} : V \times V &\rightarrow \mathbb{R}^+, \\ (u, v) &\mapsto \tilde{F}(u, v). \end{aligned} \tag{3.37}$$

Let A, \hat{A} be points in $V \subset M$ corresponding respectively to paths of length n in X , denoted $x_n \rightarrow x$ and $\hat{x}_n \rightarrow \hat{x}$,

under the map Π (so that we have $A = \Pi(x)$ and $\hat{A} = \Pi(\hat{x})$).

For the first variable u , set $u = s_M$ where $s_M \equiv d_M(A, \hat{A})$ denotes a choice of a suitable metric on M , and the second variable $v = v(n)$, so that $F(u, v) = F(s_M, v(n))$. Consider the pulled back function $\Pi^* \tilde{F} = \tilde{F} \circ \Pi$, with the property that on $U \subset M$, we have for some sufficiently small $\tilde{\epsilon} > 0$, the inequality

$$|F_U(s_n, n) - \Pi^* \tilde{F}(s_M, v(n))| < \tilde{\epsilon}. \tag{3.38}$$

Of course, such considerations may be suitable for stationary ergodic sources as well. We leave open the possibility that a suitable metric space structure on M will be conducive to introducing methods from dynamical systems such as homoclinic points, hyperbolic sets, stable manifolds and related ideas (see e.g. [77]).

Remark 3.6. The constructions proposed above may be compared with that for an optimal manifold representation of (information) data in [28] for instance. Suppose $X \subset E$ is defined by a density function $\rho(x)$ and E is a vector space with norm $\|\cdot\|$. There is a stochastic map $\Pi : X \rightarrow M$ which is seen as a projection to a lower dimensional manifold $M \subset E$. On M , a distortion measure $D(M, \Pi, \rho)$ is defined by

$$\int_{x \in E} \int_{m \in M} \rho(x) \Pi(x) \|x - m\|^2 dx dm. \tag{3.39}$$

The map Π along with the density ρ determines a joint probability function $P(M, X)$ that allows calculation of the mutual information $I(X, M)$ between X (higher dimensional) and M its lower dimensional manifold representative M , as given by

$$\int_{x \in X} \int_{m \in M} P(x, m) \log \left[\frac{P(x, m)}{\rho(x) \Pi(x)} \right] dx dm. \tag{3.40}$$

3.11 Semimartingale processes and noise

Given the possible stochastic nature of bio-cognitive behavioral and response mechanisms, we expect the underlying processes to be driven to an extent by stochastic and noise-driven diffusion processes conducive to the creation of new information. This necessitates introducing a noise mechanism into the system as an agent towards self-organization and complexity (cf [9]), just as open systems far from equilibrium require some sort of internal ‘amplification’ in order to attain to a macroscopic dynamical structure (cf [108]).

We have already mentioned how martingale analysis is instrumental in describing stochastic resonance within cognitive and epidemiological systems. In order to build this feature into a rate distortion manifold we need to consider a more general approach. A *submartingale* on the real line \mathbb{R} consists of those stochastic processes of the form $\Lambda + A$, where $\Lambda : C^0(\mathbb{R}) \times \mathbb{R}^+ \rightarrow \mathbb{R}$ represents Brownian motion on \mathbb{R} commencing at $0 \in \mathbb{R}$, possibly with a random time

change, and A denotes a continuous increasing process on \mathbb{R} .

In order to translate these concepts so as to work on a rate distortion manifold, we will need a suitable measure such as that of (3.9). A suggested definition is provided in [45]: given some domain Ω , let us say that a stochastic process

$$\chi : \Omega \times \mathbb{R}^+ \longrightarrow (M, s_M), \tag{3.41}$$

is a *martingale on* (M, s_M) , if for any convex function $g : U \longrightarrow \mathbb{R}$ defined on an open set $U \subset M$, the composition $g \circ \chi|_U$ is a submartingale on \mathbb{R} . An example is Brownian motion

$$\Lambda^M : C_a^0(M) \times \mathbb{R}^+ \longrightarrow (M, s_M), \tag{3.42}$$

starting at $a \in M$.

A sum of a continuous local martingale and a process of finite variation gives rise to the notion of a *semimartingale* [49]. Such local martingales with respect to Brownian motion say, admit certain integral representations. If $\alpha \in \Omega^1(M, \mathbb{R})$ is a 1-form and ζ a semimartingale on M , then the real semimartingale $\int \langle d\alpha, d\zeta \rangle$ (where $\langle \cdot, \cdot \rangle$ denotes the dual pairing, and d denotes derivation on tangent vectors) is called *the Stratonovich integral of α along ζ* . We refer to [49] §7 for further properties where this integral is denoted by $\int \langle \alpha, \delta\zeta \rangle$.

If M, N are (smooth) manifolds, one can define *the Stratonovich operator*

$$\mathfrak{E} = \{e(x, y)\}, \quad x \in M, y \in N, \tag{3.43}$$

to be a family of linear maps where the map $e(x, y) : T_x M \longrightarrow T_y N$ is a linear map that depends on (x, y) (to some degree of differentiability) thus defining a map $e : TM \times N \longrightarrow TN$. The latter also has a corresponding adjoint mapping $e^*(x, y) : T_y^* N \longrightarrow T_x^* M$. Given an M -valued semimartingale ζ , an N -valued semimartingale η is said to be a solution of the Stratonovich differential equation $\delta\eta = e(\zeta, \eta) \delta\zeta$, if for every 1-form α on N , there is the equality of Stratonovich integrals

$$\int \langle \alpha, \delta\eta \rangle = \int \langle e^*(\zeta, \eta)\alpha, \delta\zeta \rangle. \tag{3.44}$$

In terms of Markovian game theory the above concepts may be seen more concretely as follows. Let $\{X_n\}$ be a sequence of stochastic variables defining a game (possessing noise), with conditional expectations given by

$$E(X_{n+1} | X_1, X_2, \dots, X_n) \equiv E(X_{n+1}). \tag{3.45}$$

The definition of terms and interpretations are then:

- (i) $E(X_{n+1} | n) \geq X_n$ –*submartingale* (favorable to player).
- (ii) $E(X_{n+1} | n) = X_n$ –*martingale* (completely fair game).
- (iii) $E(X_{n+1} | n) \leq X_n$ –*supermartingale* (favorable to the house).

As in [135](§5.1), this is exemplified for an epidemiological model whereby the ‘player’ is an infectious agent, X_n is the number of people infected at stage n (the player’s fortune), and the ‘house’ is some socioeconomic system. A submartingale then represents a spreading infection, and a supermartingale represents a declining infection. The convolution of the community structure (the ‘signal’) with the opportunity structure (the ‘noise’) then leads to the simple epidemic model as a *generalized stochastic resonance*. Such situations have been studied by similar means in the case of childhood illnesses where, besides the internal transmission of an infection within a community, there is also an external effect due to individual migration between communities, and weak ‘seasonality’ together with low transmission levels are seen to induce stochastic effects with amplified noise that may generate resonance [3]. Our perspective is that models such as these could be treated in terms of *groupoids* and convoluted path space, as will be discussed in §7.1.

4 Equivalence relations and tuning

4.1 Equivalence relations

For the purpose of describing cognitive modules we can also append the structure(s) with equivalence relations $\mathcal{R}^X, \mathcal{R}^M, \mathcal{R}^E$ defined on X, M, E respectively, and a sequence of maps

$$(X, \mathcal{R}^X) \xrightarrow{\Pi} (M, \mathcal{R}^M) \xrightarrow{\varphi} (E, \mathcal{R}^E). \tag{4.1}$$

Suggestive of the orbit equivalence theorem (relative to a more abstract setting of e.g. [59, 100]), we will suppose the equivalence relations are tied by

$$\begin{aligned} \mathcal{R}^M &= \Pi \times \Pi(\mathcal{R}^X), \\ \mathcal{R}^E &= \varphi \times \varphi(\mathcal{R}^M). \end{aligned} \tag{4.2}$$

4.2 Tangent spaces

As we have suggested, a rate distortion manifold need not necessarily be a differentiable manifold in the conventional sense, but may admit an abstract differentiable space structure (such as that described in Appendix III). In some instances, however, we may have to address the question of tangency at a point $m \in M$ and thus assume that the tangent space $T_m M$ is defined accordingly. We recall how $T_m M$ can be defined in terms of an equivalence classes of curves. Consider the equivalence relation $c_1(R_m^M)c_2$ on curves c_1, c_2 as meaning: c_1 and c_2 are tangent at $m \in M$, if and only if $(\varphi \circ c_1)(R_{\varphi(m)}^E)(\varphi \circ c_2)$ means they are tangent at the point $\varphi(m)$ in E . In which case, the equivalence class $[c]_m$ at $m \in M$, is defined to be the tangent space at m , and this is usually denoted by $T_m M$. This way of viewing tangency in terms of equivalence classes globalizes to the construction of the tangent bundle $TM \longrightarrow M$ as described in e.g. [1, 87].

From an information–theoretic point of view, this description of tangency is useful for characterizing ‘tuning’. We give a more general interpretation. Suppose p_1, p_2 are paths (or ‘sequences’) in X that are projected by Π down to the manifold M ; we keep p_1, p_2 to denote their projected images. For $m \in M$, let (U, φ) be a local chart with $m \in U$ and $\varphi(U) \subset E$. With regards to the equivalence relations $\mathcal{R}^M, \mathcal{R}^E$, we want to consider paths (sequences) p_1, p_2 as being *equivalently tuned* at $m \in M$, denoted $p_1(R_m^M)p_2$, if at $\varphi(m)$ in the atlas space E , the equivalence denoted $(\varphi \circ p_1)(R_{\varphi(m)}^E)(\varphi \circ p_2)$, holds. The equivalence class of paths $[p]_m$ at $m \in M$ may then be thought of as *the tuning space at the point m in the manifold M* , in a sense corresponding to a focal point of attentive processing.

4.3 Higher and lower dimensional information sources

Let \mathbf{X} and \mathbf{Y} be information sources whereby $\dim \mathbf{X} \geq \dim \mathbf{Y}$. The ‘higher’ source \mathbf{X} is one that may be considered ‘fast’ and the ‘lower’ source \mathbf{Y} is considered ‘slow’. Associated to \mathbf{X} and \mathbf{Y} , are their respective path spaces with respective distortion metrics (X, s_X) , and (Y, s_Y) . We consider a projection map

$$\Phi : (X, s_X) \longrightarrow (Y, s_Y), \tag{4.3}$$

as complying with the following version of the Rate Distortion Theorem: *for any chosen maximum average distortion such that*

$$d(x, \Phi(x)) < \epsilon, \tag{4.4}$$

there is, in relationship to Φ , a maximum possible transmission rate δ , such that the average distortion will be less than ϵ [125].

The above condition can be represented in terms of a subset of the graph of Φ . Recalling that the graph Γ_Φ of Φ is given by

$$\Gamma_\Phi := \{(x, y) \in X \times Y : y = \Phi(x)\}, \tag{4.5}$$

we define a ‘rate distortion’ subset $\Gamma_{rd} \subseteq \Gamma_\Phi$ by

$$\Gamma_{rd} := \{(x, \Phi(x)) \in \Gamma_\Phi : d(x, \Phi(x)) < \epsilon\}. \tag{4.6}$$

Recall that each of (X, s_X) and (Y, s_Y) have associated lower dimensional and structured language of thought spaces M and N (with possible manifold/atlas topologies) and induced distortion metrics s_M and s_N , respectively. For each of these we have projections onto the lower dimensional source, Π_1 and Π_2 respectively (with whatever properties are assumed), and for which the diagram below commutes

$$\begin{array}{ccc} (X, s_X) & \xrightarrow{\Pi_1} & (M, s_M) \\ \Phi \downarrow & & \downarrow \Psi \\ (Y, s_Y) & \xrightarrow{\Pi_2} & (N, s_N) \end{array} \tag{4.7}$$

that is, $\Psi \circ \Pi_1 = \Pi_2 \circ \Phi$. In this way we see that induced on Ψ is the constraint of the Rate Distortion Theorem as it holds on Φ , such that the corresponding languages of thought adhere accordingly.

4.4 Tunable states

The genesis of a rate distortion manifold lies in the concept of a *generalized tunable retina model* as introduced in [125]. Specifically, let us suppose that threshold behavior for individual, distributed or institutional (cognitive) reaction requires some elaborate system of nonlinear relationships defining a set of renormalization parameters

$$\Omega_k \equiv \omega_1^k, \dots, \omega_m^k. \tag{4.8}$$

The critical assumption is that there is a tunable zero order state, and any changes about that state are, in first order, relatively small, although their effects on a punctuated process may not be at all small. Thus, given an initial m -dimensional vector Ω_k , the parameter vector at time $k + 1$, Ω_{k+1} , can, in first order, be written as

$$\Omega_{k+1} \approx \mathbf{R}_{k+1}\Omega_k, \tag{4.9}$$

where \mathbf{R}_{k+1} is an $m \times m$ matrix, having m^2 components. If the initial parameter vector at time $k = 0$ is Ω_0 , then at time k

$$\Omega_k = \mathbf{R}_k \mathbf{R}_{k-1} \dots \mathbf{R}_1 \Omega_0. \tag{4.10}$$

The interesting correlates of individual, institutional or machine consciousness are, in this development, *now represented by an information-theoretic path defined by the sequence of operators \mathbf{R}_k* , each member having m^2 components, for some m . The grammar/syntax of the path defined by these operators is associated with a dual information source, in the usual manner.

The effect of an information source of external signals \mathbf{Y} , is now seen in terms of more complex joint paths in Y and the \mathbf{R} -space (of operators) whose behavior is, again, governed by a mutual information splitting criterion according to the Joint Asymptotic Equipartition Theorem (a variant of the Shannon–McMillan Theorem). The complex sequence in m^2 -dimensional \mathbf{R} -space has, by this construction, been projected down onto a parallel path, the smaller set of m -dimensional ω -parameter vectors $\Omega_0, \dots, \Omega_k$.

If the punctuated tuning of institutional or machine attention is now characterized by a ‘higher’ dual information source – an embedding generalized language – so that the paths of the operators \mathbf{R}_k are autocorrelated, then the autocorrelated paths in Ω_k represent output of a parallel information source which is, given rate distortion limitations, apparently a grossly simplified, and hence highly distorted, picture of the ‘higher’ conscious process represented by the \mathbf{R} -operators, having m as opposed to $m \times m$ components. High levels of distortion may not necessarily be the case for such a structure, *provided it is properly tuned to the incoming signal*. If it is inappropriately tuned, however, then distortion may be extraordinary.

Let us examine a single iteration in more detail, assuming now there is a (tunable) zero reference state, \mathbf{R}_0 , for the sequence of operators \mathbf{R}_k , and that

$$\Omega_{k+1} = (\mathbf{R}_0 + \delta \mathbf{R}_{k+1})\Omega_k, \tag{4.11}$$

where $\delta\mathbf{R}_k$ is ‘small’ in some sense compared to \mathbf{R}_0 . Note that in this analysis the operators \mathbf{R}_k are implicitly, determined by linear regression. We thus can invoke a quasi-diagonalization in terms of \mathbf{R}_0 . Let \mathbf{Q} be the matrix of eigenvectors which Jordan–block–diagonalizes \mathbf{R}_0 . Then

$$\mathbf{Q}\Omega_{k+1} = (\mathbf{Q}\mathbf{R}_0\mathbf{Q}^{-1} + \mathbf{Q}\delta\mathbf{R}_{k+1}\mathbf{Q}^{-1})\mathbf{Q}\Omega_k. \quad (4.12)$$

If $\mathbf{Q}\Omega_k$ is an eigenvector of \mathbf{R}_0 , say Y_j with eigenvalue λ_j , it is possible to rewrite this equation as a generalized spectral expansion

$$\begin{aligned} Y_{k+1} &= (\mathbf{J} + \delta\mathbf{J}_{k+1})Y_j \equiv \lambda_j Y_j + \delta Y_{k+1} \\ &= \lambda_j Y_j + \sum_{i=1}^n a_i Y_i, \end{aligned} \quad (4.13)$$

where \mathbf{J} is a block-diagonal matrix

$$\delta\mathbf{J}_{k+1} \equiv \mathbf{Q}\mathbf{R}_{k+1}\mathbf{Q}^{-1}, \quad (4.14)$$

and δY_{k+1} has been expanded in terms of a spectrum of the eigenvectors of \mathbf{R}_0 , with

$$|a_i| \ll |\lambda_j|, \quad |a_{i+1}| \ll |a_i|. \quad (4.15)$$

The point is that, provided \mathbf{R}_0 has been tuned so that this condition is true, the first few terms in the spectrum of this iteration of the eigenstate will contain most of the essential information about $\delta\mathbf{R}_{k+1}$. This appears quite similar to the detection of color in the retina, where three overlapping non–orthogonal eigenmodes of response are sufficient to characterize a huge plethora of color sensation. Here, if such a tuned spectral expansion is possible, a very small number of observed eigenmodes would suffice to permit identification of a vast range of changes, so that the rate distortion constraints become quite modest. That is, there will not be much distortion in the reduction from paths in \mathbf{R} –space to paths in Ω –space. Inappropriate tuning, however, can produce very marked distortion as in inattentive blindness (individual, institutional or machine–oriented), in spite of multitasking. We remark that higher order rate distortion manifolds are likely to give better approximations than lower ones, in the same sense that second order tangent structures give better, if more complicated, approximations in conventional differentiable manifolds. The formal mathematical background to this idea can be found in [102]; we will be more specific about this observation in §6.4.

Remark 4.1. A possible and more general geometric way of viewing such constructions is to consider a vector bundle $V \rightarrow M$ and some operator $\mathbf{R} : \Gamma(V) \rightarrow \Gamma(V)$ on sections of V . The above local description may serve to describe the (time) evolution of \mathbf{R} and its spectral properties. In certain cases this reveals an associated ‘spectral’ set (or submanifold, which could be M itself). Alternatively, the tangent space TM (pointwise) may split into a particular ‘eigenmode’ decomposition, a feature often found in the field of differentiable dynamics (see e.g. [1, 77]).

4.5 Description of the rate distortion manifold as an information space

At this stage we can summarize some of the essential properties required for a rate distortion manifold (M, s_M) to serve as a cognitive information space within the Global Workspace setting.

- (1) M with its distortion metric s_M and atlas–manifold topology serves as a canonical model of the path space (X, s_X) , where $X = \mathcal{P}(E^\Gamma)$. The corresponding distortion measure leads to defining a rate distortion function on M complying with the Rate Distortion Theorem.
- (2) In terms of the metric s_M , M admits a length space structure in the general sense as described (*a priori*, weaker than a Finsler structure say).
- (3) M may possibly admit a ‘weak’ differentiable space structure in some suitable sense (for instance, in terms of the abstract calculus of manifolds such as [27, 55] described briefly in Appendix III).
- (4) Both ergodic and non–ergodic processes may be considered. We may also require that M admits certain stochastic properties suited to representing e.g. stochastic resonance in the informational context. On the other hand, there may be situations where M carries a flow engendered by solutions to some wave equation or the time–evolution of an operator on sections of a vector bundle over M . Spectral eigenmode decompositions of TM may be expected as §4.4 suggests.
- (5) For the purposes of a ‘directed’ theory of information, M may admit a partial ordering ‘ \leq ’ and thus may admit an underlying ‘partially ordered space’ structure (see §6.6).

5 Thermodynamic limit and the Onsager relations

Feynman [53] (following in part the work of C. Bennett) considered the problem of extracting useful work from a transmitted message. The essential argument is that computing of any form, requires work. Consequently, on recalling from (3.3), the asymptotic limit

$$H \equiv \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n}, \quad (5.1)$$

is postulated as formally homologous to the thermodynamic limit in the definition of the free energy density of a physical system as given by

$$F(K) = \lim_{V \rightarrow \infty} \frac{\log[Z(K)]}{V}, \quad (5.2)$$

where F is the free energy density, K the inverse temperature, V the system volume and $Z(K)$ the partition function defined by the Hamiltonian of the system.

In [125] it is shown at some length how this homology permits the natural transfer of renormalization methods from statistical mechanics to information theory, producing phase transitions and analogs to evolutionary punctuation in systems characterized by piecewise, adiabatically stationary, ergodic information sources. Crosstalk, as a particular characteristic, may then serve as an ‘inverse temperature parameter’.

This homology is essential for understanding the type of model spaces as described here. The point being, that the more intricate a cognitive process, measured by information source uncertainty, the greater its energy consumption. Biological phase changes appear to be ubiquitous in natural systems and can be expected to dominate information machine behaviors as well, particularly those which seek to emulate biological paradigms. In [133] these arguments are used to explore the differences and similarities between evolutionary punctuation in genetic and learning plateaus in neuronal systems.

As much as thermodynamic laws influence most kinds of vital phenomena, certain types of epidemiological and cognitive processes may be represented in terms of a thermodynamic limit on the processing capacity (as for instance in the case of inattentional blindness [128] or sleep patterns [127]).

In order to see how suitable models may be designed accordingly, consider the dual source uncertainty of a cognitive process as parametrized by a vector $\mathbf{K} \equiv (K_1, \dots, K_n)$. In analogy with nonequilibrium thermodynamics we define the *disorder* as a function S given by

$$S \equiv H(K) - \sum_{j=1}^m K_j \frac{\partial H}{\partial K_j}. \tag{5.3}$$

Further expanding the homology, leads to defining the generalized Onsager relations of temporal dynamics

$$\frac{dK_j}{dt} = \sum_i L_{ji} \frac{\partial S}{\partial K_i}, \tag{5.4}$$

where $L = [L_{ij}]$ is a matrix of constants associated to the underlying cognitive phenomena. From the symmetric matrix

$$U = [U_{ij}] = \left[\frac{\partial^2 S}{\partial K_i \partial K_j} \right] = [U_{ji}], \tag{5.5}$$

one can define associated metric coefficients as follows:

$$g_{ij} = \frac{L^2}{2} \sum_k U_{ik} U_{kj}. \tag{5.6}$$

Next, consider the source uncertainty

$$\mathbf{K}(t) \equiv (K_1(t), \dots, K_n(t)), \tag{5.7}$$

as time dependent and defining a (smooth) curve $\mathbf{K} : \mathbb{R}^+ \rightarrow M \subset \mathbb{R}^n$, in a rate distortion manifold M . Use

of standard procedures (see e.g. [1, 24]) leads to defining a suitable length space structure on M via a distance function s_M between languages A, \hat{A} , suitably represented by points along some dynamic path in M . Here s_M is given by

$$s_M(A, \hat{A}) = \int_A^{\hat{A}} \left[\sum_{i,j} g_{ij} \frac{dK_i}{dt} \frac{dK_j}{dt} \right]^{\frac{1}{2}} dt. \tag{5.8}$$

Accordingly, the curve $\mathbf{K}(t)$ with respect to the above metric structure is a geodesic in M precisely when the second order equation

$$\frac{d^2 K_i}{dt^2} + \sum_{j,m} \Gamma_{jm}^i \frac{dK_j}{dt} \frac{dK_m}{dt} = 0, \tag{5.9}$$

is satisfied, where the Γ_{jm}^i denote the associated Christoffel symbols (see e.g [1, 24]).

Remark 5.1. One may hypothesize that under the right circumstances, geodesics sufficiently near to a reference state A_0 are bound by some estimate, and external physiological forcing must be imposed to effect a transition to a different condition. This, as is pointed out in [128], may be specified in terms of regions of fatal attraction and strong repulsion akin to Black or White hole phenomena which can either trap or deflect the path of consciousness.

5.1 The torus–sphere example: differing homologies on languages of cognition

Most textbooks on algebraic topology contain the relevant definitions and concepts pertaining to singular homology of the space M in terms of the constituent homology groups $H_k(M, \mathbb{Z})$ with integer coefficients (see e.g. [20, 87, 117]):

$$H_*(M) = \sum_{k=0}^{\dim M} H_k(M, \mathbb{Z}). \tag{5.10}$$

Loosely speaking, the $H_k(M, \mathbb{Z})$ are ‘groups of cycles of differing dimensions’ which contribute to an overall characteristic of the space, namely its *homological structure*. In particular, if $M = U \cup V$ is the union of two open sets, then a finer analysis can be made in terms of the Mayer–Vietoris homology sequence

$$\begin{aligned} \dots \rightarrow H_q(U \cap V) \xrightarrow{f} H_q(U) \oplus H_q(V) \xrightarrow{g} H_q(M) \\ \rightarrow H_{q-1}(U \cap V) \rightarrow \dots \end{aligned} \tag{5.11}$$

where f denotes the map induced by a signed inclusion $a \mapsto (-a, a)$ and g is that of the sum $(a, b) \mapsto a + b$.

Rate distortion manifolds pertaining to distinct homology types are then expected to represent distinct homologies in relationship to information, intrinsic languages of cognition and culture, and to which such homological techniques can be applied. A topological example, as presented in [125] (Chapter 5) concerns the case of the two–torus T^2

versus the two–sphere S^2 , both of which are surfaces embedded in 3–space, and both possess two–dimensional tangent spaces (pointwise). Thus sitting in a small local coordinate patch, an observer cannot really notice much difference. However, their homotopy type and (singular) homologies are fundamentally different, since at level $k = 1$ we have $H_1(S^2) = 0$, whereas $H_1(T^2) = \mathbb{Z} \oplus \mathbb{Z}$.

In general, *the manifold itself forms an envelope of the entirety of its tangent planes and this envelope will in turn describe a homology type which reflects a particular topological structure.* This straightforward topological observation underscores the point of how rate distortion manifolds of distinct (fundamental) homology, homotopy and diffeomorphism types, provide an informational blueprint for the significant differences in bio–cultural/psycho–cognitive choices in Workspace informational processing. Such fundamental differences might be realized in a variety of cognitive situations where culture, as always, plays a significant role. Consider for instance the case in [99] which compares the main differences between the Asian mode of perception, on the one hand, widely framed and holistic, and the Western mode, on the other hand, more analytically and logically centered. In other words, culture bears an influence upon the ‘topologies’ of the respective information processing which partly explains the difference between the two perceptual characteristics.

6 Embeddings of contexts

This section takes up some of the development in the previous sections with a tentative outline of how we might formulate a finer geometric description of the Global Workspace in terms of a rate distortion manifold M :

- (1) The Workspace and access to it may be considered in terms of nested sequence of rate distortion manifolds

$$M_1 \subset M_2 \subset \dots \subset M_n \subseteq M.$$

This represents a hierarchy of cognitive processes based on nonlinear dynamical principles.

- (2) Cooperating and competing contexts as unconsciousness networks to be integrated within the Workspace, participate within a higher dimensional dominant context describable by embeddings as in (1).
- (3) Within each context there is a cooperating group of specialized processors where access to the Workspace can be represented through such a chain of inclusions/embeddings.

Let us start by taking an E –chart (U, ϕ) for a rate distortion manifold M relative to the model space E . Here (U, ϕ) is taken to be a chart representative of a ‘context’. We consider (U, ϕ) to be sufficiently ‘large’ so as to admit a sequence of ‘embeddings’ of charts through the above sequence. Initially we have a hierarchial context given as

before by a projection $\Pi : X \rightarrow M$ from some high dimensional information source space X (typically, a ‘culture’ with rate–distortion features) to a lower dimensional information carrier, a rate distortion manifold M (typically, a ‘structured’ or ‘canonical’ cognitive system which could be a canonical model for X).

6.1 Cooperating contexts

Cooperating contexts (cf specialized unconscious processors) within a hierarchial structure are represented as a nested sequence of rate distortion manifolds $\{M_k\}$ given by

$$(M_1; (V_1, \psi_1)) \subset (M_2; (V_2, \psi_2)) \cdots \subset (M_m; (V_m, \psi_m)) \subseteq (M; (U, \phi)), \tag{6.1}$$

each with their respective chart/atlas system (V_k, ψ_k) , and where each M_k represents a processing stage within some cognitive (sub)system corresponding to a level of information k . These come complete with inclusions through their respective chart/atlas systems

$$V_1 \xrightarrow{\lambda_1} V_2 \xrightarrow{\lambda_2} \dots \xrightarrow{\lambda_{m-1}} V_m \xrightarrow{\lambda_m} U, \tag{6.2}$$

satisfying, for $1 \leq i \leq m$, the composition

$$\psi_i = \phi \circ \lambda_m \circ \lambda_{m-1} \cdots \circ \lambda_i. \tag{6.3}$$

In this sequence, we may include at each level k retraction mappings (or projections when defined) $p_k : M \rightarrow M_k$, for $k \leq m$, such that we have a commuting diagram

$$\begin{array}{ccc} X & \xrightarrow{\Pi} & M \\ & \searrow & \downarrow p_k \\ & & M_k \end{array} \tag{6.4}$$

suggesting the influence of the higher (or faster) dimensional information source at subsystem level k .

Thus the nested sequence of embeddings could be interpreted as levels of information representing a broad dominant context hierarchy through executive functions and levels of cooperation such as (e.g. in [12] §4):

Perceptual Contexts \implies Conceptual Contexts \implies Goal Contexts

specified at each level k by M_k . The above sequence can also be appended with a string of (sub)–equivalence relations “ \sim_i ”, when specified:

$$(M_1, \sim_1) \subset (M_2, \sim_2) \cdots \subset (M_m, \sim_m) \subseteq (M, \sim). \tag{6.5}$$

The sequence is thus seen to represent a parallel series of specialized processors (cf e.g. [12, 38]). Also, the adoption of the equivalence relations affords a useful interpretation (cf the notion of ‘frames’ in the cognitive sense). Each (M_i, \sim_i) then leads in a straightforward way to a *groupoid* structure upon which we will elaborate below.

6.2 Filtration by Morse functions

The proof that an n -dimensional manifold M has the homotopy type of a CW-complex of dimension $\leq n$, relies on the use of Morse functions $f : M \rightarrow \mathbb{R}$ which are ordered in a suitable sense (see e.g. [95, 113]). Specifically, we choose a sequence of numbers c_1, c_2, \dots, c_{n-1} for which the following is satisfied: if a_i^λ (resp. $b_j^{\lambda+1}$) are critical points of index λ (resp. $\lambda + 1$), we have $f(a_i^\lambda) < c_\lambda < f(b_j^{\lambda+1})$. Then the manifolds $M_\lambda = f^{-1}[0, c_\lambda]$ define a filtration

$$M_* : M_0 \subseteq M_1 \subseteq \dots M_\lambda \subseteq \dots \subseteq M_n = M. \quad (6.6)$$

6.3 Competing contexts

Next we consider another such nested sequence $\{N_\ell\}$ of rate distortion manifolds with their respective chart/atlas system (W_j, ζ_j) contained within that of $(M; (U, \phi))$:

$$\begin{aligned} (N_1; (W_1, \zeta_1)) &\subset (N_2; (W_2, \zeta_2)) \\ \dots &\subset (N_n; (W_n, \zeta_n)) \subseteq (M; (U, \phi)), \end{aligned} \quad (6.7)$$

also complete with inclusions through their respective chart/atlas systems

$$W_1 \xrightarrow{\rho_1} W_2 \xrightarrow{\rho_2} \dots \xrightarrow{\rho_{n-1}} W_n \xrightarrow{\rho_n} U \quad (6.8)$$

satisfying, for $1 \leq j \leq n$, the composition

$$\zeta_j = \phi \circ \rho_n \circ \rho_{n-1} \dots \circ \rho_j. \quad (6.9)$$

Likewise, we have a projection of influence as in (6.4), and each (N_j, ζ_j) can be appended with an equivalence relation “ \sim_j ”. The sequence then represents the hierarchy

Perceptual Contexts \implies Conceptual Contexts \implies Goal Contexts

The main difference between this sequence $\{N_j\}$ and the $\{M_i\}$, is that *at a sufficiently low level in the hierarchy*, the former represents a *competing sequence* to the latter within a dominating system $(M; (U, \phi))$ subject however to the projection $\Pi : X \rightarrow M$. In a Workspace setting, the sequences $\{N_j\}, \{M_i\}$ each with limited capacity, contend for recognition in a central or main processor (as represented by the larger canonical model $(M; (U, \phi))$ with its rate distortion characteristics).

Some suggested conditions are in order (but most likely not exhaustive). These include the possibilities that over some range of indices $1 \leq \ell < \min\{i, j\}$, we have:

- (i) $\dim M_\ell \geq \dim N_\ell$ [same level competition].
- (ii) $M_\ell \cap N_\ell \neq \emptyset$ [same level competition].
- (iii) For charts $\psi_\ell|V_\ell \cap W_\ell \neq \zeta_\ell|V_\ell \cap W_\ell$, when (ii) above occurs. The charts ψ_ℓ and ζ_ℓ are distinct functions on $V_\ell \cap W_\ell$ [competition].

In other words, the contexts near to, or at the bottom of, the hierarchy compete to dominate the Workspace by

means of their intrinsic cognitive mechanisms as represented by the geometric structure of the $\{M_i\}$ and $\{N_i\}$ via their respective chart/atlas system. There are other possible interpretations. For instance, as in [125], stress can be viewed as a socially devised cultural characteristic involving a schemata of languages each with its own grammar/syntax. This may be represented in mathematical terms by the high dimensional information source X . The dimensional reduction (given by projection $\Pi : X \rightarrow M$) along with the embeddings of the chain of the M_k into M , can then be viewed as the relevant interacting cognitive modules within some environment: for instance, how embedded psycho-social stress influences mind-body interactions. For ‘noise’ related purposes, we may also consider under appropriate conditions, the various embeddings $M_i \rightarrow M_j$ as linked in a semimartingale process by Stratonovich operators $\mathfrak{E}_{ij} : TM_i \times M_j \rightarrow TM_j$ as in (3.43), and likewise for the chain $N_i \rightarrow N_j$.

These cognitive modules, as reflective of the mathematical description of the model, so adhere to the asymptotic limit theorems of information theory as we have outlined them. For many persons, information overload, ‘noise’ and ‘heat engines’, are the fiendish perpetrators of such stress ailments as the modules so describe. The all-or-none competing stimuli creating a bottleneck in the central processing of neural information is considered in [38, 128] as a Workspace explanation of inattentive blindness for which the above schemata of rate distortion manifolds may serve as a blueprint.

6.4 Higher order tangency

The above setting holds further prospects for applying state-of-the-art techniques from geometry and topology. In particular, we have mentioned that higher order rate distortion manifolds are likely to produce better approximations than those of lesser order. A formal explanation of these terms is as follows. If M is a smooth manifold in the traditional sense, then a classical example is how tangent vectors coalesce with osculating curves such as the local geodesics. More generally, one may consider the *higher order tangency* of submanifolds of M where the maps in question admit *osculating spaces* to certain orders. Following [102], the idea revolves around *p -th order tangent bundles* $TM^{[p]}$ whose typical fiber consists of a p -th order *osculating vector*. The latter can be related to the classical osculating spaces of order p of a submanifold of some affine space. These higher order tangent bundles comply with the exact sequence

$$0 \rightarrow TM^{[p-1]} \rightarrow TM^{[p]} \rightarrow S^p(TM^{[1]}) \rightarrow 0, \quad (6.10)$$

where S^p denote the p -th symmetric tensor product. Such osculating spaces seem relevant to ‘higher order retinal tuning’ in the context of institutional multitasking [131], a topic for further development.

6.5 Hierarchy of organization and complexity

Complex systems (whether they are genetic, neuronal or cultural–social) can be organized into various levels of complexity each equipped with their sovereign mechanisms for managing the prevailing environment or community. Recall also how neurons communicate via synapses by means of a synchronous assembly of encoded neurons, thus realizing a mental event with some degree of plasticity. At a higher order there exists a semantic memory which influenced by referential experiences within its environment, induces the development of personal memory as characteristic of humans and other higher order species. Quite often the question is to determine how the higher levels evolve from those lower without being directly reducible to them, in a way similar to how percolation techniques of lower to higher level operations are realized in the theory of graphs and networks. A similar problem has been discussed in [46] in the context of *memory evolutive systems* (MES) which depend on types and classes of selection procedures. An internal feature is that iterated complexification can induce a hierarchy of objects of strictly increasing orders of complexity each with its own characteristics which allow a switching across the constituent organisms. This can be explained in categorical terms, in particular, in terms of ‘colimits’ as we will briefly describe.

Firstly, let us give a short but quite abstract notion of a *graph* as a family of objects $\{A_i\}$ together with a collections of arrows $f : A_1 \rightarrow A_2$ between objects. In the absence of the strict definition of a *category* (see e.g. [19, 91]), let us say for now, and informally, that a standard notion of a categorical structure can be defined on a graph in terms of objects and an internal rule of composition ‘ \circ ’ associating to pairs $f : A_1 \rightarrow A_2$, $g : A_2 \rightarrow A_3$, the composition $g \circ f : A_1 \rightarrow A_3$, satisfying the rule of associativity, together with the identity morphism $\text{id}_A : A \rightarrow A$.

If we regard objects as labeled in terms of ordered states $A < A'$, a transition functor $F(A, A') : F_A \rightarrow F_{A'}$, represents a change in states $A \rightarrow A'$, and satisfies

$$F(A, A'') = F(A, A') \circ F(A', A''). \quad (6.11)$$

Following e.g. [46], if we have a system as represented by a graph, it is said to be *hierarchical* if the objects can be divided into specified complexity levels representative of the embeddings of contexts.

Further, we can speak of a pattern of linked objects A as a family of objects A_i with specified links (edges) between them. Consider another object B to which we can associate a collective link from A to B by a family of links $f_i : A_i \rightarrow B$. We can picture then a cone with a base consisting of $A = \{A_1, A_2, \dots\}$ and with B as the vertex. The pattern is said to admit a *colimit* denoted C , if there exists a collective link $A \rightarrow C$ such that any other collective link $A \rightarrow B$ admits a unique factorization through C . If such a colimit C exists, then locally C is well–defined by the nature of the pattern to which it is attached, and globally,

C enjoys a universal property determined by the totality of the possible collective links of the pattern. In other words, C effectively binds the pattern objects while at the same time functions as the entire pattern in the sense that the collective links to B are in a one–to–one correspondence with those to C . Further, a category can be said to be *hierarchical* if its objects can be partitioned into different levels of complexity, with an object C of level $n + 1$ say, being the colimit of at least one pattern of linked objects of (strictly) lower levels $n, n - 1, \dots$.

In our situation the concept is particularly useful when the objects A_i comprise a pattern or network of rate distortion manifolds and the collective links $f_i : A_i \rightarrow B$ are morphisms to B which may, for instance, model a central processor. The colimit C then functions as the binding agent for the respective channels of information.

Remark 6.1. The concept of a colimit in a category generalizes that of forming the union $A \cup B$ of two overlapping sets, with intersection $A \cap B$. However, rather than concentrating on the actual sets A, B , we place them in context with the role of the union as permitting the construction of functions $f : A \cup B \rightarrow C$, for any C , by specifying functions $f_A : A \rightarrow C$, $f_B : B \rightarrow C$ agreeing on $A \cap B$. Thus the union $A \cup B$ is replaced by a property which describes, in terms of functions, the relationship of this construction to all other sets. In practical terms [15, 22] it is how we might compare input and output. In this respect, a colimit has ‘input data’, viz a *cocone*. For the union $A \cup B$, the cocone consists of the two functions $i_A : A \cap B \rightarrow A$ and $i_B : A \cap B \rightarrow B$.

An *evolutive system* [46] is viewed as a family of categories indexed by a suitable parameter t (usually time), together with a family of transition functors. The internal organization of a complex component C can then be modeled in relationship to a pattern of linked objects such that the actions of C on any other component are determined completely by the collective links (of the pattern), thus characterizing C as the colimit. The above model can describe an evolutionary autonomous system (or organism) with a hierarchy of components managing organized exchanges within an environment. Thus a *hierarchical evolutive system* is then an evolutive system in which the state category at each value t is hierarchical and the transition functors preserve the levels. By means of such a network of learning, the system re–adapts to changing conditions within the environment, thus leading to a MES, a characteristic which can be related to the embedding of contexts. We re–iterate that these components could be realized as specific types of rate distortion manifolds which collectively model information relay within an interactive context.

6.6 Scheduling of paths

Methods of concurrency involving directed homotopies, scheduling, n –categories/ n –complexes and related topics may well be suited to developing certain aspects of cognitive/institutional multi–tasking. The geometric perspective

is outlined in [62, 105] and an application to the Global Workspace is discussed in [132].

A particular idea starts by recalling the notion of a *partially ordered space* (a *po-space* for short) with respect to a partial ordering “ \leq ” on M . A *local po-space* is a Hausdorff space M with a covering $\mathcal{U} = \{U_\alpha, \leq_\alpha, \alpha \in J\}$ where each $U_\alpha \subseteq M$ is open and \leq_α is a partial order on U_α . We may assume here that M is a rate distortion manifold corresponding to some cognitive process. There is some scope as to how “ \leq ” may be linked to the rate distortion theorem (locally) on the U_α , in terms of channel capacity, etc. We keep in mind that our rate distortion manifolds may be subjected to ‘direction’ as would be required within a setting of channeled consciousness; this would involve some further analysis and grounds for a separate discussion.

Remark 6.2. As pointed out in [103], given a smooth manifold M , a po-space structure on M may be defined in terms of ordering of Morse functions $f : X \rightarrow \mathbb{R}$ as mentioned in §6.2. Briefly, for $x, x' \in M$, we decree an ordering by $x \leq x' \iff f(x) < f(x')$, or $x = x'$. The theory of po-spaces is one of several abstract methods employed for analyzing concept structures in theoretical computer science (others, such as *Chu spaces* [106] incorporate strict logical structures and are innately different to the ‘thermodynamic’ features of rate distortion manifolds).

Nevertheless, it is possible that a hierarchy of contexts may be executed concurrently and the ensuing transition states may be subjected to a ‘schedule’ which the (cognitive) organism may inter-impose via evolution. We give a brief mathematical description following [41, 51], but for now restricting to the non-directed case. Consider the path space $\mathcal{P}(M)$ of paths

$$\gamma : \mathbb{R}_{\geq 0} \rightarrow M, \tag{6.12}$$

of finite length. Given a covering $\mathcal{U} = \{U_a : a \in A\}$ of M by open sets indexed by a set A , a *schedule* is an element of the monoid $SA = (A \times \mathbb{R}_{\geq 0})^*$, where elements are pairs of words of the same length $(a_1 a_2 \cdots a_n, t_1 t_2 \cdots t_n)$. We say that a path γ fits the schedule $(a_1 a_2 \cdots a_n, t_1 t_2 \cdots t_n)$, if

- i) $\gamma(t) = \gamma(t_1 + \cdots + t_n)$, for $t \geq t_1 + \cdots + t_n$
- ii) $\gamma([t_1 + \cdots + t_i, t_1 + \cdots + t_{i+1}]) \subset U_{a_{i+1}}$
- iii) $\gamma([0, t_1]) \subset U_{a_1}$.

Also, there is an equivalence relation on schedules generated by

$$(a_1 a_2 \cdots a_n, t_1 t_2 \cdots t_n) \simeq (a_1 a_2 \cdots a_{i-1} a_{i+1} \cdots a_n, t_1 t_2 \cdots t_{i-1} t_{i+1} \cdots t_n), \text{ if } t_i = 0. \tag{6.13}$$

The main result of [41] states that for certain coverings, the schedules may be assigned continuously to all paths up

to the latter equivalence and this can be used towards *globalizing* locally continuous fibrations over a given space. It seems workable to apply this concept to the directed case, which we finesse for now. However, for our rate distortion manifolds the main point is that the open sets of the covering each should contain neighborhoods of points constrained by the estimate (4.4) of the Rate Distortion Theorem. A slight word of caution is necessary here: the rate distortion manifold (M, s_M) as we have described it, is already a canonical model for the ‘semantic path space’ (X, s_X) , so the above path space $\mathcal{P}(M)$ has to be considered somewhat apart from the space X . It is appealing that this notion of path scheduling may be linked to the study of universal algorithms where there is an intention to establish lower bounds on the running times of computational procedures (cf [88]). Furthermore, ‘paths in a space of paths’ is a potentially useful concept since eventually one may wish to consider an approach similar to the Jamesian ‘processes of processes’ upon which we will comment later (see §8.7).

7 Further towards groupoids

7.1 Convolved path space

It is suggested in [125]§3 that a pattern of sensory input mixed in some way with an internal ongoing activity induces a path $\gamma = (\psi_0, \psi_1, \dots, \psi_n, \dots)$, where each ψ_k may represent a composition of internal/external signals. Guiding this path into some kind of decision process, yields an output $h(\gamma)$ which belongs to one of two (disjoint) sets of system response depending on whether the pattern is recognized or not. If it is the case, the appropriate response-action may assumed as initiated. This is quite general, but for the sake of classifying cognitive modules it is suggested that commencing from the input level, one may actually classify the paths themselves in preference to specifying the output.

Typically, an input x representing an information source, is tied to an output y via some path; for example, a path representing the transition probability $p(x|y)$, or a channel of information. In another sense, one can define *equivalence classes* in a convoluted path space, such as (X, s_X) , according to which a state ψ_k is path-connected to a source state ψ_S . In this way, two states $\psi, \hat{\psi}$ are said to be *equivalent*, denoted $\psi \mathcal{R} \hat{\psi}$, if they lie on the same path γ with source ψ_S seen as varying. In this way, the path space is decomposable into (relatively) disjoint sets of equivalence classes. Such an equivalence relation defines a category known as a *groupoid*, a ‘small’ category G with all morphisms invertible, represented by

$$G \overset{r}{\underset{s}{\rightrightarrows}} M \tag{7.1}$$

where M denotes the space of *objects* and r, s denote the range (or target) and source maps, respectively (see e.g. [21, 143]). Such disjoint equivalence classes are applicable to disjoint ‘cognitive modules’ for which the equiva-

lence relation is defined by the existence of a *high probability meaningful path* connecting two points. Later we will discuss a network groupoid in which the vertices of the network (or graph) will represent different information sources dual to a cognitive process. Certainly, the study of equivalence classes for dealing with e.g. response versus sensory input, is an attractive option to analyzing an overwhelmingly complex network since the key principle would be to reduce the latter to manageable configurations involving only the (equivalence) classes.

Specifically, in our case (M, s_M) is viewed as the canonical model of (X, s_X) which, as we proposed, could be replaced by the groupoid G under path equivalence. An action on M by G , induces an equivalence relation \mathcal{R} , together with a convolution product

$$(a * b)(\gamma) = \sum_{\gamma_1 \circ \gamma_2 = \gamma} a(\gamma_1)b(\gamma_2). \tag{7.2}$$

By using general means (see e.g. [31]) we can form a corresponding *convolution algebra* $\mathcal{C}(G)$ over G of which many special cases can be realized in a systems–response mechanism. Typically, in a response to an environmental stimulus, a ‘response’ function $h : \mathcal{C}(G) \rightarrow B$, mostly nonlinear, can be defined where B is such an extensible set as before, and which could be the underlying set of some semantic/syntactical algebra, or that of an algebra of some class of operators. The simple epidemic model as a ‘generalized stochastic resonance’ mentioned in §3.11 is an example that immediately comes to mind.

7.2 Geometric phase and holonomy

The concept of holonomy in a physical sense could be loosely described by the following scenario: imagine walking along a path of some gradient flow. You may observe that neighboring flow paths tend to veer off; but as you progress steadily further, other flow paths appear to approach asymptotically. The explanation is well-known to anyone who has taken a first course in differential geometry: holonomy is essentially the parallel translation of vectors around a closed path, thus leading to a representation of the space of closed paths into a group of prevailing symmetries. The classic example involves the *Poincaré first-return map* of a dynamical system (see e.g. [1, 77, 96]). The holonomy concept embraces the sense of phase transition throughout the physical and biological sciences in whatever the context and wherever the internal states of a system are tracked in relationship to the latter’s spatiotemporal orientation. Notable physical examples include the Berry phase, whereby a slowly evolving quantum system in returning to its original state retains a memory of its motion via a geometric phase in the wave function, a phase as given by $\exp(i \int_{\gamma} A)$, where A is a suitable potential and γ is the path in question. Likewise in the Born–Oppenheimer approximation, as nuclei describe a closed path in a certain parameter space, the electronic wave function acquires such a phase. There is the more mundane example of a

cat held upside down and then released from a reasonable height. The cat usually lands safely on its feet but with its orientation reversed [97], thus realizing holonomy as incorporated within a certain innate cognitive–physiological skill.

Parallel transport induced by a ‘flat’ connection/potential having zero field strength (curvature), but nevertheless having non-trivial holonomy, causes shifting interference patterns in electron beams in the vicinity of a solenoid (cf the Aharonov–Bohm effect and Wilson loop [5, 85]). So in a similar way, the key to understanding how seemingly disjoint cognitive modules interact lies within globalizing the iterates of such local procedures to create the associated holonomy groupoid (a technique described for topological groupoids in [4]). In the skeletal framework of graphs and networks, holonomy can be described in the context of symmetry groups.

7.3 Noise flow on a rate distortion manifold

There is a convenient approach to noise flow on a rate distortion manifold in terms of *groupoid actions*. One may consider a system of noise variables $\mathcal{B} = (s_1, s_2, \dots, s_\ell)$ associated to an informational process associated modeled by a rate distortion manifold M . If the noise is network related, then it is reasonable to speak of ‘equivalence classes of noise’ and hence an associated groupoid B whose set of objects would be a network of paths in M . The essential point here is that given the groupoid B acts on M , the equivalence classes ‘foliate’ M in some way. This is simply the principle that a foliation on M corresponds to a groupoid, and conversely, the foliation is induced by the action of B :

$$\left\{ \text{Groupoid Action } B \times M \rightarrow M \right\} \implies \left\{ \text{Foliation } (M, \mathcal{F}) \right\}. \tag{7.3}$$

Generally, this foliation will be *singular* as, for instance, the leaf dimensions may jump up and down. For a study of the general theory of foliations see [26] (cf [92, 96] which deal with groupoid actions). On the other hand, the noise equivalence classes may in practice be 1-dimensional, in which case we have a *noise-induced flow* (M, \mathcal{F}) on M that is essentially stochastic in nature, and most likely singular in the general case as well.

Relative to a metric \mathcal{M} on M , we have the simplest type of Onsager relation $d\mathcal{M}/dt = LdS/d\mathcal{M}$ to which a noise term can be added:

$$\frac{d\mathcal{M}}{dt} = L \frac{dS}{d\mathcal{M}} + \sigma W(t), \tag{7.4}$$

where $W(t)$ is a function representing white noise and σ is a constant. In this way a stochastic differential equation is induced on \mathcal{M} [134]:

$$d\mathcal{M}_t = L \left(\frac{\partial}{\partial t}, \frac{\partial S}{\partial \mathcal{M}} \right) dt + \sigma \left(\frac{\partial}{\partial t}, \frac{\partial S}{\partial \mathcal{M}} \right) dB_t, \tag{7.5}$$

where L and σ are now regarded as suitable functions of t and $ds/d\mathcal{M}$, and dB_t represents the noise structure derivable from the noise variables \mathcal{B} as above. Thus the above groupoid action is manifestly the noise flow on M engendered by the 1-form component $\sigma(\frac{\partial}{\partial t}, \frac{\partial S}{\partial \mathcal{M}}) dB_t$.

7.4 Global actions and groupoid atlases

We expect that more general descriptive mechanisms of cognitive modules may eventually involve features such as atlases modeled on groupoids themselves. Once instance involves an *orbifold* for which the associated orbifold atlas corresponds to a proper (Lie) groupoid (see [96] and Appendix §11.2). Thus a ‘rate distortion orbifold’ would then be a fitting term when a rate distortion space has to admit certain singularities. However, an even broader concept is that of a *groupoid atlas*[16]. The latter, loosely speaking, entails the pasting together of local groupoid actions with the net effect of a ‘global’ groupoid, a concept which may prove to be particularly significant for logically inscribing processors or sensors (the ‘multi-agents’) within a cognitive module or a communication network.

Following [16], one commences from a family of groupoids $\{G_1, G_2, \dots\}$ where each groupoid has the same set of objects; this family is called a *single domain* or *multiple groupoid*. A *groupoid atlas* is then defined as a set with a covering by patches, each of which comprise a single domain with global action. An advantage of using this sort of atlas is that, in general, it admits a weaker structure compared with that of a conventional manifold since no condition of compatibility between arbitrary overlaps of the patches is necessary. This is an attractive option for studying cognitive modules geared to equivalence class representations. For instance, in a (single domain) global action by a group, the graph representation of intersecting orbits yields a configuration of various types of circuits (loops, etc.), and from such an action, a corresponding global action can be formulated so that group actions, in particular, can be more generally extended to groupoid actions that encode the actions of the various equivalence relations.

On taking a group G , the motivation for defining a groupoid atlas comes from considering a *global action* \mathcal{A} which consists of a set $X_{\mathcal{A}}$ together with a family

$$\{(G_{\mathcal{A}})_{\alpha} \curvearrowright (X_{\mathcal{A}})_{\alpha} : \alpha \in \Psi_{\mathcal{A}}\} = \{(G_{\mathcal{A}})_{\alpha} \times (X_{\mathcal{A}})_{\alpha} \longrightarrow (X_{\mathcal{A}})_{\alpha} : \alpha \in \Psi_{\mathcal{A}}\}, \tag{7.6}$$

of group actions ‘ \curvearrowright ’ on subsets $(X_{\mathcal{A}})_{\alpha} \subseteq X_{\mathcal{A}}$, where the local groups $(G_{\mathcal{A}})_{\alpha}$ and the corresponding subsets $(X_{\mathcal{A}})_{\alpha}$ are indexed by an indexing set $\Psi_{\mathcal{A}}$ called *the coordinate system of \mathcal{A}* , satisfying the conditions:

- (a) If $\alpha \leq \beta$ in $\Psi_{\mathcal{A}}$, then $(X_{\mathcal{A}})_{\alpha} \cap (X_{\mathcal{A}})_{\beta}$ is $(G_{\mathcal{A}})_{\alpha}$ -invariant.
- (b) For each pair $\alpha \leq \beta$, there is given a group homomorphism

$$(G_{\mathcal{A}})_{\alpha \leq \beta} : (G_{\mathcal{A}})_{\alpha} \longrightarrow (G_{\mathcal{A}})_{\beta},$$

such that given elements $\sigma \in (G_{\mathcal{A}})_{\alpha}$, and $x \in (X_{\mathcal{A}})_{\alpha} \cap (X_{\mathcal{A}})_{\beta}$, we have $\sigma x = (G_{\mathcal{A}})_{\alpha \leq \beta}(\sigma)x$.

The categorical assignment

$$G_{\mathcal{A}} : \Psi_{\mathcal{A}} \longrightarrow \text{Groups} \tag{7.7}$$

is called the *global group of \mathcal{A}* , and the set $X_{\mathcal{A}}$ is called the *enveloping set* or the *underlying set of \mathcal{A}* .

Suppose we have a group action $G \curvearrowright X$. Then we have a category $\text{Act}(G, X)$ with object set X and $G \times X$ its arrow set. It is straightforward to show that $\text{Act}(G, X)$ is actually a groupoid [16](see also Appendix I). Effectively, given an arrow (g, x) , we have source and range defined respectively by $s(g, x) = x$, and $r(g, x) = g \cdot x$, represented by

$$x \xrightarrow{(g,x)} g \cdot x. \tag{7.8}$$

The composition of (g, x) and (g', x') is defined when the range of (g, x) is the source of (g', x') such that $x' = g \cdot x$. This yields a composition $(g'g, x)$:

$$x \xrightarrow{(g,x)} g \cdot x \xrightarrow{(g',g \cdot x)} g'g \cdot x. \tag{7.9}$$

We have an identity at x given by $(1, x)$, and for any element (g, x) its inverse is $(g^{-1}, g \cdot x)$. A key point in this construction is that the orbit of a group action then becomes a connected component of a groupoid.

The above account motivates the following. A *groupoid atlas* \mathcal{A} on a set $X_{\mathcal{A}}$ consists of a family of local groupoids $(G_{\mathcal{A}})$ defined with respective object sets $(X_{\mathcal{A}})_{\alpha}$ taken to be subsets of $X_{\mathcal{A}}$. These local groupoids are indexed by a set $\Psi_{\mathcal{A}}$, again called *the coordinate system of \mathcal{A}* which is equipped with a reflexive relation denoted by \leq . Writing $(X_{\mathcal{A}})_{\alpha\beta} = (X_{\mathcal{A}})_{\alpha} \cap (X_{\mathcal{A}})_{\beta}$, this data is to satisfy the following conditions [16]:

- (1) If $\alpha \leq \beta$ in $\Psi_{\mathcal{A}}$, then $(X_{\mathcal{A}})_{\alpha\beta}$ is a union of components of $(G_{\mathcal{A}})$, that is, if $(X_{\mathcal{A}})_{\alpha\beta}$ and $g \in (G_{\mathcal{A}})_{\alpha}$ acts as $g : x \longrightarrow y$, then $y \in (X_{\mathcal{A}})_{\alpha\beta}$.
- (2) If $\alpha \leq \beta$ in $\Psi_{\mathcal{A}}$, then there is a groupoid morphism defined between the restrictions of the local groupoids to intersections

$$(G_{\mathcal{A}})_{\alpha}|_{(X_{\mathcal{A}})_{\alpha\beta}} \longrightarrow (G_{\mathcal{A}})_{\beta}|_{(X_{\mathcal{A}})_{\alpha\beta}}, \tag{7.10}$$

and which is the identity morphism on objects.

We can briefly exemplify matters as follows. Let us recall the projection of information sources $\Phi : X \longrightarrow Y$, from the higher (faster) X to the lower (slower) Y , and recall from (4.6) we defined a rate distortion manifold $\Gamma_{rd} \subseteq \Gamma_{\Phi}$ on the graph of Φ , by

$$\Gamma_{rd} := \{ (x, \Phi(x)) \in \Gamma_{\Phi} : d(x, \Phi(x)) < \epsilon \}. \tag{7.11}$$

Let G_Z be a group whose elements are, for instance, matrix components suitably representing those of a culture or environment via a slowly interacting source $Z = \{Z_k\}$ [134].

We consider an action $G_Z \curvearrowright \Gamma_{rd}$ along the previous lines, and subsequently obtain an action groupoid

$$\text{Act}(\Gamma_{rd}, G_Z) \rightrightarrows \Gamma_{rd} . \tag{7.12}$$

The significance of this construction is that the components of the above action groupoid may be related to the mutual information splitting criterion

$$I(X|Y_1, \dots, Y_m | Z_1, \dots, Z_n), \tag{7.13}$$

an essential ingredient for representing such interactions as those of a biosocio–culture [128].

8 The underlying simplex of a rate distortion manifold and network groupoids

Here we devote some attention to how the simplicial methods of graphs and networks, which are some of mainstream tools of information theory, multi–agent systems and concurrency, can be related to the often ‘continuous’ structures of rate distortion manifolds.

8.1 Simplicial methods towards networks

Given our rate distortion manifold M , we may want to recover an underlying decomposition of M into ‘networks’ so as to incorporate useful graph–theoretic techniques. A standard topological way of achieving this is by introducing simplicial methods and in particular, by taking a *triangulation* of M (see Appendix II) that serves as a conceptual mechanism towards tracing the discrete nature of the various graphs and networks functioning as specialized processors within the Workspace. Specifically, a *triangulation* (K, ϕ) of a space M means we have a simplicial complex K together with a homeomorphism

$$\phi : |K| \longrightarrow M, \tag{8.1}$$

where $|K|$ denotes the *polyhedron* or *geometric realization* of M (see e.g. [117] and Appendix II). Observe that simplicial methods often deal with choices of an open covering and for a given rate distortion manifold M , such a covering may be achieved by a collection of length spaces $\{U_\alpha, d_\alpha\}$, each of can be taken to be isometric to some simplex.

The triangulation (K, ϕ) of M that we have described above leads to identifying M with an associated polyhedral space, and so there follows a number of ‘discrete’ possibilities leading from a ‘continuous’ to a ‘discrete’ coarse–graining approach. Other possibilities may include the comparison theorems of spaces of bounded curvature with their negative curvature characteristics such as the $CAT(k)$ –spaces (*Cartan–Alexandrov–Toponogov spaces* where the (k) denotes that a value k is imposed as a curvature bound [24]). This leaves open the possibility that some class of rate distortion manifolds, realized in the category

of $CAT(k)$ spaces, might, for instance, admit a ‘hyperbolic structure’ in a suitable sense.

A topological graph can be converted to a metric graph by an assignment of ‘lengths’ to edges, although for infinite graphs this may result in a topology change [24]. More specifically, a simplicial decomposition of M permits an internal, skeletal–like representation of M in terms of a graph or network Γ , and then subsequently to an associated groupoid model. We can start with a categorical representation $\text{FreeCat}(\Gamma)$ of the graph Γ : regarding the vertices of Γ as objects, then between two vertices v, w , we take $\text{FreeCat}(\Gamma)(v, w)$ to denote the set of paths or edges in Γ commencing at v and ending at w . To a path $v \mapsto w$, we can assign a sequence of edge–labels (a_1, a_2, \dots, a_n) . The composition in $\text{FreeCat}(\Gamma)$ is by the usual concatenation of paths where for each edge a between v and w , a reciprocal (or reverse) edge a^{-1} between w and v exists.

In forming path sequences, the latter can be reduced by removal from the sequence of any adjacent edges of the form (a, a^{-1}) , or (a^{-1}, a) . In this way the graph Γ leads to a groupoid structure, namely the free groupoid $\text{FreeGpd}(\Gamma)$ over Γ (see e.g. [21, 146]). Thus many of the graph–theoretic and network analysis models relating to phase transition, percolation and epidemic processes, etc., can be reduced to a combinatorial groupoid setting for which there is already available a broad range of algebraic concepts that can be applied.

8.2 Modular networks and the giant component

Between disjoint cognitive modules one assumes that linkages occur randomly and the latter represent ‘cross–talk’ as a (non–zero) measure of mutual information. A descriptive method for studying this influence of cross–talk uses random graph theory (mainly following Erdős–Rényi [50]; we also refer to the exposition in [2]). One of the key concepts is that of a *giant component*, that is, a subnetwork that dominates the entire network (of cognitive modules) and which can capture up most of the smaller subnetworks.

More specifically, suppose we consider c elements of the equivalence class algebra of languages (that is, c disjoint cognitive modules) dual to some cognitive process as represented by the vertices of a graph. If a graph with c vertices has $\ell = \frac{1}{2}ac$ edges chosen at random, for $a > 1$, then it will have a giant connected component with approximately $g(a)c$ vertices with

$$g(a) = 1 + \frac{1}{a}W(-a \exp(-a)), \tag{8.2}$$

in which W denotes the Lambert W –function defined implicitly by $W(x) \exp(W(x)) = x$. An example depicted in [126], reveals a sharp phase transition occurring at $a = 1$ that initiates the Global Workspace as a ‘tunable blackboard’ defined by a set of cross–talk mutual information measures between interacting modules. The cross–talk connection corresponds to random linkages in the case of

[50] and the entropy H of the language dual to the cognitive process, will grow as some monotonic function of the giant component. Such a phase transition in a network affords some correlation between the size of the component and the richness of the language to which it is associated. Thus tuning the giant component by altering the network topology leads to further insight through a geometric representation.

Standard analysis of critical clustering reveals that uplifting the clustering coefficient increases the average number of edges necessary for the formation of a giant component. For instance, [126] on applying [124], shows that for a random network with parameter a , at cluster value C , there is a critical value given by

$$a_C = \frac{1}{1 - C - C^2}. \tag{8.3}$$

One easily sees that for $C = 0$, the giant component forms when $a = 1$. The case $C \geq \frac{1}{2}(\sqrt{5}-1)$, which is the Golden Section, shows that no giant component is definable, for any a . Thus not every network topology can support a giant component. As pointed out in [126], some cognitive network models cannot then represent consciousness and this poses some big questions ranging from the evolution of the latter to the actual nature of the sleeping state (cf [127]). Institutional or machine-based cognition no less necessitates the synchronization of information relay between giant components [134, 135].

A governing principle is to define an interaction parameter $\omega_C > 0$ that will define a regime of giant components of network elements linked via mutual information $\geq \omega_C$. Then following [134, 135], the idea is to *invert the argument*: namely, a given topology of the giant component will in turn define some value of ω_C , so that network elements interacting by mutual information less than ω_C will be blocked from conscious perception (see the example of inattentive blindness below). Thus ω_C is a syntactically dependent detection limit which depends on the giant component topology for an individual cognitive framework. Thus the variation of ω_C is one example of a topological shift. This opens up the possibility that the level sets of ω_C may be defined in terms of Morse theory. Accordingly, a parameter space may be characterized by the critical points of ω_C to ensure a fundamental shift in the high level cognitive topology.

8.3 Inattentive blindness

We briefly describe a situation from [128, 135] that has already been mentioned. An intensive focus on a task involving interactive cognitive modules may necessitate the giant component to be sustained at an optimum level within the topology of the network in question. In this way, a high limit may be placed on the magnitude of mutual information signals which can intrude into the Workspace. When the focus of attention on a single aspect of a complicated perceptual field or programmed environment precludes the

detection of intervening events which may or may not be essential to the original task, a condition known as *inattentive blindness* occurs [115]. An example of this condition might be that of a person conducting an on-line business transaction while oblivious of occasional ‘pop-ups’. In this scenario it may be that intervening signals fall below a threshold in syntax in order to intrude markedly on consciousness; alternatively, it fails to be an enduring competitor in the Workspace (cf §6.1). Further, it is expected that slower acting information sources represent the embedding sociocultural factors across the environment. In the context of institutional/directed cognition, the intense focus on economic and data-driven programs may often result in a blind-side to other essential factors (such as the sociological consequences of planned shrinkage, industrial expansion, etc.). Given a fixed topology of the Workspace, the condition of inattentive blindness thus emerges as a thermodynamic limit on the overall processing capacity [128, 135].

8.4 Network phase transitions via connections on graphs and groupoids

The finer study involves the nature of phase transitions within the simplicial network/graph structure underlying the geometry of a rate distortion manifold which we will proceed to describe. Typically, various types of percolation processes exhibit phase transitions. For instance, in [84] network percolation techniques are used to analyze phase transitions of dynamic neural systems such as those embedded within segments of cortical neuropil. But for a large class of networks there are available means for measuring phase transitions and differences in terms of parallel transport and holonomy which are analogous to the standard differential-geometric means employed on a differentiable manifold. We shall briefly discuss some of these.

Firstly, for graph-theoretic models there are certain combinatorial notions which can be used to replicate a ‘differential’ structure as realized on a standard differentiable manifold. Let $\Gamma = (V, E)$ be a graph with V denoting a finite vertex set, E an edge set with an oriented edge $e = (u, v)$ (accordingly, $e^{-1} = (v, u)$) such that $u = i(e)$ is the initial vertex and $v = t(e)$ is the terminal vertex. The *star of a vertex* $st(v)$ is the set of edges emanating from v , that is

$$st(v) = \{e : i(e) = v\}, \tag{8.4}$$

(see also Appendix II). Given that the star of a vertex is sometimes viewed as the combinatorial version of the tangent space to a manifold at a point, in [18] is defined a *connection* ∇ on a graph Γ as defined via a set of one-to-one functions $\nabla(u, v)$, one for each oriented edge $e = (u, v)$ of Γ satisfying:

- (1) $\nabla(u, v) : st(u) \rightarrow st(v)$
- (2) $\nabla(u, v)(u, v) = (v, u)$
- (3) $\nabla(v, u) = (\nabla(u, v))^{-1}$.

Assuming a graph Γ admits such a connection ∇ , in [18] is defined the notion of a 3-geodesic as a sequence of four vertices (u, v, w, z) with edges $\{u, v\}$, $\{v, w\}$ and $\{w, z\}$ for which $\nabla(v, w)(v, u) = (w, z)$. Subsequently, a k -geodesic is defined inductively as a sequence of $(k+1)$ vertices. The three consecutive edges $\{d, e, f\}$ of a 3-geodesic is referred to as an edge chain. A closed geodesic can then be specified as a sequence of edges e_1, \dots, e_n such that each consecutive triple $(e_\alpha, e_{\alpha+1}, e_{\alpha+2})$ is an edge chain for each $1 \leq \alpha \leq n$ (modulo n). This leads to a convenient way of defining a ‘totally geodesic subgraph’ [18], that is, given (Γ, ∇) , a subgraph $\Gamma_0 = (V_0, E_0) \subset \Gamma$, is said to be totally geodesic if all geodesics commencing at E_0 remain within E_0 . In other words, for every two adjacent vertices u, v in Γ_0 , we have

$$\nabla(u, v)(\text{st}(u) \cap E_0) \subseteq E_0. \tag{8.5}$$

8.5 The covariant derivative of entropy along a network path

Suppose now the above vertices $(e_1, e_2, \dots, e_{k+1})$ are interpreted and renamed as $k + 1$ information sources $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{k+1})$ in accordance with the APSE condition (see Remark 3.2), where the \mathbf{X}_i act with the set of tuning parameters associated to a set of giant components. We consider a connection ∇ acting

$$\nabla(\mathbf{X}_i, \mathbf{X}_j) : \text{st}(\mathbf{X}_i) \longrightarrow \text{st}(\mathbf{X}_j), \tag{8.6}$$

with the indicated properties (for $1 \leq i, j \leq k + 1$) as before. With respect to the metric $\mathcal{M} = \mathcal{M}(\mathbf{X}_i, \mathbf{X}_j)$ applied to these information sources, the above ‘connection’ map in (8.6) implements on the underlying network, the covariant differentiation along the path $\mathbf{X}_i \longrightarrow \mathbf{X}_j$, just as in (3.30):

$$dH/d\mathcal{M} = \lim_{\mathbf{X}_j \rightarrow \mathbf{X}_i} \frac{H(\mathbf{X}_j) - H(\mathbf{X}_i)}{\mathcal{M}(\mathbf{X}_i, \mathbf{X}_j)}. \tag{8.7}$$

Now relative to each \mathbf{X}_i , a maximized channel capacity C_i is assigned, in accordance with the estimate of (3.15), that is, $H(\mathbf{X}_i) \leq C_i$, holds for $1 \leq i \leq k + 1$, and in respect of the Rate Distortion Theorem along paths $\mathbf{X}_j \longrightarrow \mathbf{X}_i$. This apparent optimality in terms of the estimate (3.15) motivates decreeing the information sources $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{k+1})$ to be a k -geodesic (there is no loss in generality by supposing that these actually form a closed geodesic). In fact, [18] shows that the set of all such geodesics in a given network determines the connection ∇ which accordingly can be implemented as a form of covariant derivative along the remaining paths.

Remark 8.1. In order to realize how geodesics may arise in applied graphs and networks, there is the example of [73] where in the ever competing US telecommunications industry, extensive and rapid switching of networks between nodes (vertices), the large-scale use of fiber optics (reducing transmission costs) and network nodes ramifying to interconnecting subnetworks, has lead to the reduction of

a vast pyramidal-like network to a structure with many geodesic subgraphs. Thus the pyramid transforms to a structure somewhat akin to Buckminster Fuller’s ‘geodesic dome’. The principles are analogous to those expected in huge networks of parallel computation which in turn contribute efficiency to the skeleton of some institutional cognitive mechanism.

8.6 The holonomy groupoid and path connections

Given $C = \{e_1, \dots, e_n\}$ is any cycle in Γ , for which $t(e_\alpha) = i(e_{\alpha+1})$ modulo n , then the connection around C leads to a permutation

$$\nabla_C = \nabla_{e_n} \circ \dots \circ \nabla_{e_1} \circ \nabla_{e_0}, \tag{8.8}$$

of $\text{st}(u)$. The upshot is that the notion of a holonomy group at a vertex can be defined [18]: the holonomy group $\text{Hol}(\Gamma_u)$ at a vertex u of Γ , is the subgroup of the permutation group of $\text{st}(u)$ generated by the permutations ∇_C over all cycles C that pass through the vertex u . In this way, holonomy contributes to the geometry of the graph.

On the other hand, as noted in [82] every groupoid G leads to a reflexive symmetric graph (RSG). Loosely speaking, the objects form the vertices, arrows form the edges, inversion in the groupoid leads to symmetry, and the identity leads to the pointwise identity arrow. For instance, on a manifold M the set $\mathcal{P}(M)$ of (smooth) Moore paths $\gamma : [a, b] \longrightarrow M$, has the structure of a RSG with M the vertex set and $\gamma(a), \gamma(b)$ defined to be the domain and codomain of γ respectively. Taking $M_{(1)}$ to denote the first neighborhood of the diagonal of M [82], a connection ∇ on $G \rightrightarrows M$ is a morphism of a RSG from $M_{(1)}$ to the underlying graph of G . Let $(x, y) \in M_{(1)}$, then similar to above, $\nabla(x, y)$ is an arrow $x \mapsto y$ in G , such that

- i) $\nabla(x, x) = \text{id}_x$
- ii) $x \sim y \implies \nabla(y, x) = (\nabla(x, y))^{-1}$.

A path connection on G is a morphism of RSGs in the case of $\mathcal{P}(M) \longrightarrow G$ that satisfies certain rules of reparametrization, representation and subdivision (we refer to [82] §6 which follows in part [123]).

Many groupoids with connection

$$(G \rightrightarrows M, \nabla) \tag{8.9}$$

have the property that unique partial integrals exist along any map $[a, b] \longrightarrow M$. Thus we may say that (8.9) admits path integration. Consequently, a connection ∇ is then ‘flat’ along any path, that is, $\nabla(x, y) \circ \nabla(y, z) = \nabla(x, z)$.

For $u \sim v$ in $[a, b]$, let us set

$$\left(\int_\gamma \nabla\right)(u, v) = \nabla(\gamma(u), \gamma(v)). \tag{8.10}$$

Following [82] we define the holonomy $\text{hol}_\nabla(\gamma)$ along a path $\gamma : [a, b] \longrightarrow M$, as the arrow $(\int_\gamma \nabla)(a, b)$ with domain

$\gamma(a)$ and codomain $\gamma(b)$. Thus a map

$$\text{hol}_{\nabla} : \mathcal{P}(M) \longrightarrow G, \tag{8.11}$$

is obtained. If (8.9) admits path integration, then the above map in (8.11) is considered to be a path connection.

There appears to be a close relation of these ideas to [61, 118] wherein are considered n -cell systems based on systems of n ordinary differential equations describing the dynamics of some (possibly) synchronized network. This is important because the synchronous coupling of a cell system to its close environment affects a change in the latter as well as in the collective organism whose task it is to square-up to those of the higher, multi-parallel, institutional types. The coupling and equivalence of cells leads to a natural groupoid structure of a resulting coupled cell network $\Gamma = (V, E, \sim_v, \sim_e)$ with its intrinsic equivalence classes $[v]_V$ and $[e]_E$. Here, the vertices or nodes of the network are taken to be representative of such cells. Synchrony of such cell systems may be dependent on groupoid symmetries which, as pointed out in [131] in the context of institutional cognition, can be broken by an impinging rapid crosstalk internal to the system while the latter attempts to manage a slower external crosstalk.

For each $v \in V$, a *cell phase space* P_v is defined. Usually P_v is a finite dimensional vector space and a *total phase space* is defined as $P = \prod_{v \in V} P_v$. A vector field may then be characterized in terms of a map $f : P \rightarrow P$, that in principle should be related to the above permutation subgroups of $\text{st}(u)$ thus leading to a suitable notion of *parallel transport* within the system.

We have at this stage arrived at a formalism for obtaining a network/graph theory underlying a typical rate distortion manifold, similar to taking an X-ray picture of an essential organism. Beyond the example of Remark 8.1 there are many possible applications such as in areas where one considers the passage from an iteration of local processes towards global structures. For instance, the situation described in Remark 8.1 is likely to have analogs in the study of social networks. These may be the ‘small world’ graphs having low density and which are highly clustered thus giving rise to the likelihood of networks of geodesic subgraphs [142]. ‘Small world’ relationships are studied in [109] in a similar way to how strong ties (families, cliques, etc.) with large clustering are bridged by weak ties [63]. The corresponding social networks are likely to involve more intricate topologies and statistical fluctuations, and where ‘simplicial’ Nash equilibria may provide optimal predictions within the resulting framework of games [120]. Additionally, the graph holonomy concept and the ‘giant component’ thus provide formal criteria in which to specify the essential phase transitions leading to higher orders of complexity. For small world networks in the context of Global Workspace Theory, steps in this direction have been taken in [60].

8.7 2-Groupoids and Stacks

To some degree the cognitive modules we have considered should afford a Jamesian characteristic of “processes of processes”. The key is to take a step up in ‘categorical dimension’. Loosely speaking, a 2-category \mathcal{C}_2 can be described in a ‘cellular’ sense: \mathcal{C}_2 consists of a class of objects \mathbb{O} (0-cells), a class of 1-morphisms \mathbb{A}_1 (1-cells), a class of 2-morphisms (2-cells) with ‘horizontal’ composition defined between 1- and 2-cells, along with a separate ‘vertical’ composition between 2-cells. In other words \mathcal{C}_2 affords the extra mechanism of *morphisms between morphisms*. When the 2-morphisms of \mathcal{C}_2 are invertible and the 1-morphisms invertible (up to homotopy), then \mathcal{C}_2 shapes up as a 2-groupoid. Suitable reference to this subject are e.g. [19, 76, 78, 91].

The Cartesian closed category Cat of small categories is a 2-category in which the 2-morphisms are the natural transformations for which the vertical composition is given via composition in the codomain category. Also, the category of groupoids Gpd is a (full) 2-subcategory of Cat . A 2-functor $F : \mathcal{C}_2 \rightarrow \mathcal{D}_2$ is an enriched functor in Cat that preserves the 2-category structure of \mathcal{C}_2 on taking objects, 1- and 2-morphisms of \mathcal{C}_2 to those in \mathcal{D}_2 .

In relationship to manifold structures, the Yoneda lemma says that any space or manifold M is uniquely determined by the categorical functor

$$\text{Map}(\cdot, M) : \text{Mnf} \longrightarrow \text{Sets} \tag{8.12}$$

A *stack* S is a (2)-functor between categories of manifolds and groupoids (with categories)

$$S : \text{Mnf} \longrightarrow \text{Gpd} \subset \text{Cat} \tag{8.13}$$

where for any manifold M , we obtain a corresponding category $S(M)$ in which all morphisms are isomorphisms, for any morphism $f : N \rightarrow M$, we have a functor $f^* : S(M) \rightarrow S(N)$, and for any $Z \xrightarrow{g} N \xrightarrow{f} M$, there is a natural transformation $T_{f,g} : g^* f^* \cong (g \circ f)^*$ which is associative on a trio of compatible morphisms.

As shown in e.g. [68, 86, 91], such a functor S also enjoys the properties of glueing together all of the objects and morphisms. Furthermore, S can itself admit an chart/atlas description generalizing that for a manifold, thus leading to a potentially useful concept for a further large-scale study of interactive cognitive modules in the same way as the groupoid atlas has been proposed. The above account is one categorically formal means of representing a rate distortion manifold as groupoid (or, to consider a stack of groupoids upon the former). In the 2-categorical sense, one then contemplates a next step up from ‘meaningful paths’ to ‘meaningful membranes’ towards realizing a higher order Global Workspace continuum (see Remark 8.3 below). In another, but related context, the idea of ‘morphisms between morphisms’ may be relevant to the passage from 1st order to 2nd order complexity of information in terms of ‘referents’, in so far that the 2nd order houses the sense of ‘meaning’ [10].

Remark 8.2. In view of earlier remarks concerning parallel processing, it would be reasonable to append to the model a chain of rate distortion submanifolds $M_1 \subseteq M_2 \subseteq \dots \subseteq M_n = M$, on each of which there is a groupoid structure corresponding to an equivalence relation “ \sim_i ” (cf *path equivalence*), for $1 \leq i \leq n$:

$$\begin{aligned}
 (G_1, \sim_1) &\rightrightarrows M_1 \\
 (G_2, \sim_2) &\rightrightarrows M_2 \\
 &\dots \dots \dots \\
 (G_n, \sim_n) &\rightrightarrows M_n
 \end{aligned}
 \tag{8.14}$$

A broader framework could be related to the ‘stack’ functor S as previously, where any such nested sequence (of information) would yield a corresponding nested groupoid sequence as stacks over the manifolds M_k .

Remark 8.3. We have mentioned that geometric concepts such as parallel transport and holonomy may be realized within graphs and networks. One instance of a 2–categorical approach to 2–parallel transport using simplicial methods is described in [11] by means of a ‘sweeping functor’. There is also the related work of [23] which is relevant to surface holonomy.

9 Some applications towards cognition at–large

In previous sections we described the mathematical architecture of the possible rate distortion manifolds and network related ideas. Next we discuss the motivating informational background from the point of view of immunology–language and several classes of cognition with possible ramifications.

9.1 The Atlan–Cohen perspective

The immunology–information principle as outlined in [8] starts with sets of strings of amino acids in an antibody molecule poised to influence the quantity of information in the corresponding protein. Recall that protein synthesis as a channel of information is transcribed into the protein amino acid sequence which acknowledging the genetic code whereby DNA stores information in the nucleotide bases A(Adenine), C(Cytosin), G(Guanine), T(Thymine).

Biological interactive networks as a class of complex networks consist of local cellular communities organized and managed by their characteristic selection procedures. In such a partitioning of the structure, it is necessary to regulate the local properties of the organism rather than the global mechanism while genetic switches operate as transcription factors encoding and switching on other genes within this hierarchy. Moreover, one can include systems which by their intrinsic structure interact via noncommutative relationships. More specifically, inter–regulatory systems of genetic networks via activation or inhibition of

DNA transcription can be modeled at several differing levels where various factors influence distinct states usually by some embryonic process or by the actual network structure itself. For each gene it is important to understand the dynamics of inter–regulatory genetic groups which of themselves create hierarchial systems with their own characteristics. A gene positively (or negatively) regulates another when the protein coding of the former activates (respectively, inhibits) the properties of the latter. In this way, genetic networks are comprised of interconnecting positive and negative feedback loops. The DNA binding protein is encoded by a gene at a vertex i say, activating a target gene j where the transcription rate of i is realized in terms of a function of the concentration x_j of the regulatory protein. Acting towards a given gene, the regulating genes are protein coded and induce a transcription factor. Subsequent modeling techniques can be drawn from a variety of mathematical sources : graph theory, stochastic differential equations, and Boolean networks are examples (specific approaches are realized by de Jong et al.[35]). An overall exposition of these ideas from the categorical viewpoint and that of higher dimensional algebra is presented in [15].

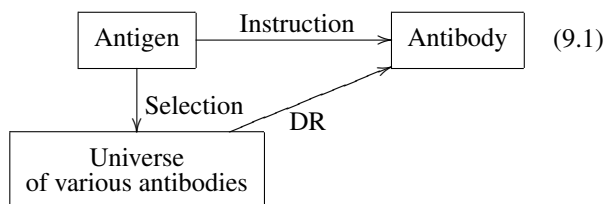
Immune networks had been proposed by Jerne [75] as networks of mutually interacting and cooperative ‘idiotypes’ and ‘anti–idiotypes’ as regulators of immune response towards antigenic approaches through which the antigen itself reveals a ‘meaning’. In relationship to this hypothesis, a main premise of [8] is that an antigen should be viewed as a fundamental unit of information. However, it is postulated that noise prevails in the system, thus interfering with and faulting the transmission of information. Any ‘meaning’ then attributed to an antigen is dependent on the kind of immune response it generates and one which, as proposed in [8], operates via the molecular structure in some accordance with the Shannon–theoretic principles of information. Consequently, the system has several options in responding to an antigen: a finely tuned cognitive system organizes the information as it is induced by the antigen and devises the ‘format’ for internal processing and release into the biochemical language of the immune system.

One instance is where the system engages different response cytotoxic T–cells, where ‘helper’ T–cells secrete mixtures of cytokines while lymphocytes navigate several cell types. Subject–predicate type of communication occurs when an antigen cell communicates an immune sentence to a T–cell which is unable to recognize the antigen totally. Thus the T–cell antigen receptor (TCR) requires the antigen to register with a superficial major histocompatibility complex (MHC) whereby a peptide functions as the ‘subject’ (of the immune sentence), and the way in which the T–cell responds to the peptide in the MHC is said to define the *meaning* of the antigen. The various response/non–response options are germline predicate signals comprising of cell–interaction/adhesion molecules. The predicate signals assess the antigen–presenting cells (APC) and tissues, thus registering the potential threat posed by the anti-

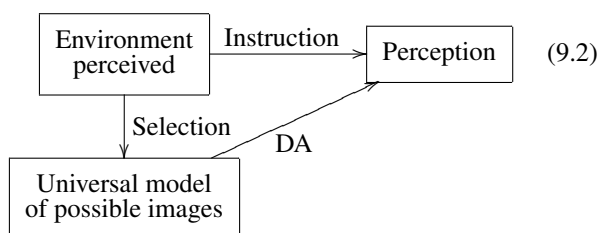
gen. The T-cells read off the context of the antigen subject through integration of the germline signals with antigen recognition and react accordingly by dispatching a team of cytokines and other molecules. In their overall function, they resemble typical cognitive processes – the response of the immune system to an antigen so reflects a function of the entire community (network). This is the essence of the immuno-cognitive function: assimilating a perceived signal with respect to a learned association with the environment, and then upon comparison, initiating a select response mechanism from a large repertoire of possibilities. As incorporated into higher animals, the immune system becomes patently a deeper cognitive organism due to the increased complexity of factors of social mechanisms and environmental management each exerting their characteristic cross-talk and tendency to noise. In certain respects, the breakdown of a given immune system so results from a disorder in transmission of information.

A possible scenario tied to the paradigm of [8] describes the activation of T-cell development and the immunological synapse via adhesion of the T-cell and APC which in theory could be initiated by the peptide MHC [40]. The viewpoint of [40] is that, in practice, the process of signaling response is influenced by certain classes of integrins in which the actin cytoskeleton provides a suitable structural mechanism for assimilating the signaling input. Since the immunological synapse is sensitive to the overall quality of the MHC peptide, the formation of this synapse depends upon the T-cell surface and the actin/myosin cytoskeletal systems in composing a cellular structure out of the transient interaction of the TCR and peptide MHC.

The ‘Collective Efficacy’ of [110] is one source of analogies between immune cognition and socio-environmental neighborhood cognition. Similarly, the diagram below based on [48](§3.3 Fig. 8) shows a mapping from immune recognition schemes in terms of a *determination through universals*



where ‘DR’ is short for ‘differential reproduction’. Associated to this interpretation via immunology, the corresponding ‘cognitive’ interpretation may be represented by [48](§3.4 Fig. 11):



where ‘DA’ is short for ‘differential amplification’.

9.2 Comparisons with neural networks

The underlying processes of institutional/directed cognition and intelligent machine operation can in part be compared with the functioning of neural networks, thalamocortical and olfactory systems, as examples. Recall that the acclaimed Hodgkin–Huxley model, together with several allied models provide a descriptive base for studying a variety of neuronal cell complexes in which informational patterns can be analyzed on codes based on the temporal properties of impulses: statistical intervals, frequencies, amplitude and phase variation. Accordingly, operative functions that will determine the number of possibilities, depend mainly on the statistical structure of the information sources and the specific nature of the codes in question. In a related setting, the theory of differentiable dynamics is applicable for modeling the effects of neuronal activity (such as spiking and bursting) in terms of homoclinic/periodic orbits in relationship to stable (or unstable) manifolds of critical elements, saddle node-bifurcations, hyperbolic sets, and the applications of the major theorems of Smale and others (see e.g. [52, 72]). In this respect, rate distortion manifolds are suitable models for analysis of such concepts while at the same time they afford the special features of adhering to the Shannon–McMillan theorems. We recall the Poincaré first return map relative to the phase portraits (see e.g. [72]) that originally led to the holonomy concept, an essential descriptive mechanism of neural and cognitive transition states as we have pointed out in the context of groupoids.

Rather than by individual cells, quanta of information can be considered as encoded by communities of the former. Typically, place cells are representative of encoding information within an environmental frame of reference whereupon a quorum of cells responds to the demands within a given location. Each constituent putatively breaks down its response in terms of an average, plus a variation in noise (neurons can be typically noisy and in turn can cause noisy synaptic inputs, oftentimes impeding transmission relay) thus contributing to sequences of spiking, in turn encoding information within the period of stimulus. Eventually, there results an overall cumulative response to the environment in relationship to the direction of motion, color, shape, form etc. as they are encoded into the appropriate regions of the visual cortex.

Recall that in the pioneering work on holography (and later wavelets) Gabor [56] postulated an ‘uncertainty’ – a quantum of information corresponding to a limit to which both frequency modulations and spatial information can be simultaneously measured. Pribram [107] in the context of neural networks and brain transition states, developed analogous ideas of holography/uncertainty, to some extent based upon the Gabor theory. Within neuronal systems, dendritic-processing employs analogous uncertainty principles in order to optimize the relay of information by

micro-processing. Both time and spectral information (frequencies) are considered as stored in the brain which supposedly maintains a process of self-organization in order to minimize the uncertainty through a wide-scale regulatory system of phase transitions, the origin of which involves the various computational neuroscientific mechanisms of (hyper) polarizing action potentials, spiking, bursting and phase-locking, etc. These contribute to a multitude of network cells that register and react to an incoming perceptive signal. Thus it is claimed that cognitive processes up to consciousness may emerge from the neural level, but this emergence necessitates the integration of lower levels evolving from the successive cultural complexifications through phase transitions within a hierarchy of which the model of a colimits structured MES is one such example.

It is understood that the maintenance of a cell membrane potential depends not only upon inter-cellular communication, but also upon spiking and bursting rates: usually fast K^+ for transmission between nerve and muscle cells, and slow Ca^{2+} for contraction of muscle fibers [58]. Periodic inputs give rise to spike trains, but stochastic resonance through noise is needed in order to surpass a threshold for an action potential [57]. Granted a noisy environment, one expects a suitable noise level for the maximum signal transmission in correlation to the rate distortion theorem. It is pertinent to the question of neuronal computation by population coding, gating and phase-locking in the presence of stochastic resonance; altogether a different informational scenario to the language/immunology of cellular systems where maximum likelihood methods can involve substantial data accumulation leading to implement an ‘electoral system’ for predicting vectors by regarding the activity of a given cell as a vote for taking a preferred direction [104] and thus initiating its cognitive response, quite in tune with the Atlan-Cohen model.

9.3 The thalamocortical model

In an analogous way, the viewpoint of [12], as we have mentioned, is to regard the nervous system as a collection of specialized unconscious processors complete with its own squad of perceptual analyzers, output systems, etc. These are considered as performing cooperatively and efficiently within their locale, but since the system is characteristically decentralized, such qualities may not automatically function at a ‘global’ level. Thus within the system (or community) the interaction, control and coordination of squads of unconscious specialists depend on a central information exchange somewhat like a typical broadcasting system (such as the Global Workspace). Take for instance a cognitive basis for emotion, complete with its own language/grammar/syntax as a framework for individual adjustment to a challenging psycho-social environment and a mechanism for implementing various response categories towards the latter [125] – a further slant on the Atlan-Cohen perspective. Whereas some functions can be performed habitually, special operations require a combined

team effort, the strategy and implementation of which is somehow relayed throughout the environment/community (cf §9.1). It appears to be a characteristic ubiquitous to a number of commonly studied neuro-cognitive and immuno-biological models. For instance, the neurobiological hypothesis of [13] is that intralaminar nuclei as a subset of the thalamus comprises the broadcasting network for the Global (Neuronal) Workspace. A main premise is that the reticular nucleus of the thalamus is instrumental for gating attention in an information-theoretic capacity and thus constitutes an agent towards consciousness.

In many regions within the various cortical zones, neuronal groups from one zone can arouse those in another so to produce a relatively organized re-projection of signals back to the former, thus creating a network of reverberating loops as are realized in the hippocampus, the olfactory system and cortical-thalamus. A riding assumption is that there is a certain synchronization of neurons through resonance and periodic oscillations of the neighboring population activity. Let us dwell on a particular scenario. Suppose \mathfrak{X} and \mathfrak{Y} denote surfaces consisting of neurons and receptor cells respectively, and let $f : \mathfrak{X} \rightarrow \mathfrak{Y}$ be a mapping of points of \mathfrak{X} to assigned points of \mathfrak{Y} under f .

In the *maps/re-entry* model [42, 43, 44], such a mapping should be considered as a component of the cerebral anatomy which is equipped and genetically coded with such mapping networks, as for instance, the operational part of the visual cortex. Re-entry is a selective process whereby a multitude of neuronal groups interact rapidly by two-way signaling (reciprocity) where parallel signals are inter-relayed between maps; take for instance the field of reverberating/signaling cycles active within the thalamocortical meshwork. *A priori*, such a process is not a feedback system since there are many parallel streams operating simultaneously and re-entry channels serve to link, in a sense, the compositions $f_1 \circ f_2 \circ f_3 \dots$ of distinct maps. In general these mappings are defined locally throughout, where a global mapping can be considered as defining a perceptual category. The maps/re-entry processes comprise a representational schemata for external stimuli on the nervous system, ensuring the context dependence of local synaptic dynamics at the same time mediating conflicting signals. Thus re-entrant channels between hierarchial levels of cortical regions assist the synchronous orchestration of neural processes. Impediments and general malfunctioning of information in the re-entry processes (possibly due to some biochemical imbalance) may then be part explanation for various mental disorders such as depression and schizophrenia. The association of short-term memory tied to consciousness within an architecture of thalamocortical reverberatory loops is proposed in [33]. Further support for the thalamocortical model as an essential component of the Workspace is provided in [38] in the context of a neuronal basis for inattentional blindness, the cognitive malfunction we had described earlier. From our perspective, the nested sequences of rate distortion manifolds considered in §6 and the processing via groupoids in (8.14) as descriptive mech-

anisms for such interactions, hence appear strikingly relevant.

The efficiency of re-entry is dependent on widespread variation in strength of connection, orientation and the potential convergence/divergence of paths in the rate distortion sense. Suggestive of the 2-categories interpretation of the Jamesian sense of consciousness through processes of processes, *the dynamic core hypothesis* [44] concerns the strength–framework of neural interactions within a functional cluster, mainly prevalent in the thalamocortical meshwork. A point here is that the dynamic core defines a neuronal state space (space of objects) and paths connecting points in this space represent a sequence of conscious states over time. We suggest that morphisms between the paths themselves should be admissible. Information relayed to the Workspace is proposed in [13] from the intralaminar nuclei comprising certain collections of thalamic regions. The reticular nucleus of the thalamus is considered in [13] as instrumental in gating attention. Under the premise that an orchestrated thalamus is a key component towards consciousness, the reticular nucleus is one leading factor to which a network–theoretic analysis/ rate distortion theory seems applicable.

Our discussion of groupoids concerning the reciprocity in relay of signaling (invertibility) in such networks, is a motivation for representing neuronal (groups) clusters by an appropriate categorical–algebraic structure (much weaker than the conventional notion of a ‘group’ in a mathematical sense). Such categorical representations in the terminology of [46] are called ‘categorical neurons’ (or *cat-neurons* for short). Consciousness loops [43], the Global Neuronal Workspace of [12] are among an assortment of models that have such a categorical representation. Among other things, there is proposed several criteria for studying the binding problem via the overall integration of neuronal assemblies and concepts such as *the archetypal core*: the cat–neuron resonates as an echo propagated to target concepts through series of thalamocortical loops. Analogous to how neurons communicate mainly through synaptic networks, cat–neurons interact in accordance with certain linking procedures and can be studied in the context of categorical logic which in turn may be applied to semantic modeling for neural networks [66, 67]. In this respect (neuro)groupoids with their invertibility property for all morphisms, provide the descriptive sub–mechanisms for reciprocity within the constituent assemblies.

9.4 Autopoietic systems and Distributed Cognition

The viewpoint of Maturana and Varela [93], as supported by several accounts in this paper, is that cognition is fundamentally a biological process and that living systems inhabit a cognitive domain through the autopoiesis of structurally coupled unitary (self-reproducing) systems that influence the organization and maintenance between both themselves and their environment over time. Many types of

systems, be they biological or social, are realized through the autopoiesis of their various components and the totality of their interactive relations forming a medium in which these components realize their ontogeny. If anything, this may simply be for the sake of getting their survival mechanisms straight. It is through participation alone that an autopoietic system determines a social system by realizing the relations that are characteristic of that system, and it is reasonable to view their ‘cellular’ models as described in terms of the information spaces we have considered. The ‘cellular’ organization of cognition adjusts and adapts to the ever-changing thermodynamic phase transitions of the environment and subsequent levels of complexity; accordingly, the latter induces by reciprocation a re-adjustment within the former. Davia [34] suggests defining the range of thermodynamic conditions in which an organism can mediate transitions as a catalyst to be its “environmental survival space”.

The descriptive and causal notions which can be described in terms of our groupoid (and other categorical) structures may be guided by the following principles [93] (Chapter III):

- (1) Relations of constitution that determine the components produced constitute the topology in which the autopoiesis is realized.
- (2) Relations of specificity that determine that the components produced be the specific ones defined by their participation in the autopoiesis.
- (3) Relations of order that determine that the concatenation of the components in the relations of specification, constitution and order be the ones specified by the autopoiesis.

In this respect, concepts such as the Atlan–Cohen model and Institutional Cognition would appear to have partial overlap with autopoietic systems whereby the dynamics of their constituent cognitive cells can be modeled in terms of rate distortion manifolds as component representations of the Global Workspace architecture. Davia [34] argues that the concept of a soliton wave is ubiquitous to representing an autopoietic self–sustaining dynamic process. It is interesting to hypothesize that such wavefronts permeating through a given cognitive cell may actually be represented by a ‘foliation’ on a corresponding rate distortion manifold (such as a ‘noise flow’ as was discussed earlier).

9.5 Distributed and Institutional cognition

Closely related are systems of *Distributed Cognition* [70] (as discussed in [134] which dynamically inter-arrange and marshal their subsystems for task–implementation within the broader context of cognitive ethnography. In a similar way to autopoietic transformations, distributed cognition applies not only in relationship to individual human cognition, but extends to the broader institutional/machine–based cognition where humans undertake the task of con-

trolling and navigating through multi-tasking machine worlds, implementing policies, etc. while embedded in the ambient ‘memetic’ environment of that culture. Hollan et al. [70] exemplify task-oriented activities of tightly-knitted groups in relationship to their working environment, and address the social organization and structure of activity that induces an information flow as part of the cognitive process necessary for the completion of a given operation. This may entail certain perceptual inferences within an evolved ‘conceptual space’ of the tasking environment (such as realized in the handling of digitally regulated flight instrumentation [70]). Concerning the interactions between distributed cognition, ethnography, experiment, work place and work materials, Hollan et al. [70] identify several widely applicable core principles such as:

- people establish and coordinate different types of structure in their environment
- it takes effort to maintain coordination
- people off-load cognitive effort to the environment whenever practical
- there are improved dynamics of cognitive load-balancing available in social organization.

The ‘culture’ of oceanic navigation, such as described in [74], with its exclusive range of techniques of measurement, skills, etc., itself becomes a cognitive process. This way of thinking about how such computational mechanisms are essentially cognitive, is discussed at philosophical/complexity levels in [119].

Somewhat related to the apparent corporate teamwork of distributed cognition are other information oriented systems of cognitive interaction. These may be viewed in a dynamical systems context which incorporates ‘embodiment’ within the context of cultural, linguistic factors, physical motion, and so on [30]. One such example is that of *social prosthetic systems* [83] which describe how deficiencies in individual (cognitive) capacity can be compensated via participation with the brain–fusion of socio-environmental networks. The argument is based on how supposed “selfish” genetic programming, aware of limitations on information handling, motivates reaching into the environment to attain to conceptual management within the latter. Loosely speaking, the brain uses the world and “enduring relationships” as extensions of itself [83].

Once such systems can be represented by their corresponding equivalence classes, configurations of interacting groupoids can be realized for which the discussion of §7.4 has relevant applicability. These can be compared with, and applied to, the network analysis and geodesic subgraph evolution via ‘small world’ partnerships as discussed in e.g. [2, 109, 142] and where ‘weak ties’ permit the formation of Global Workspaces and inter-communication between them [63, 135]. In this respect, the underlying graph of a groupoid and the concept of a groupoid atlas may well become essential techniques for delving further into the descriptive mechanisms of such systems (see [60]).

Whereas the disciplines of neurophysiology/biology provide some explanation to the underlying mechanisms of human consciousness (but often curbed by the strictures of the ‘mereological fallacy’ [17]), it is of growing importance to further study the interactive–reciprocity of the individual body/brain with the environment, as in the way autopoietic and social prosthetic systems profess to do. Likewise, some brave new world of consciousness machines will interact with their embedding systems thus creating new strains of epidemics and cognitive failures [136].

One may also consider how related social factors on a more global scale can physically determine and shape the environment created through the cognitive mechanisms of its inhabitants. Within a framework of spatial syntax and information, this has been addressed in [69]. Such factors lead to multifarious forms of development (and those quite clearly tied to the influence of institutional cognition) and are manifest at many levels. For instance, we have the concepts of ‘street’, ‘terrace’, ‘lane’ and how these civil structures eventually do shape the physical appearance, the cultural character and ethos of a city while reflecting its order of wealth, industry, affluence, ethnic divisions, and so on. As much as this development might once have been viewed as positive over decades, centuries even, inevitably several ‘institutional cognitive modes’ that assisted the creation of the city in the first place, often are destined to go into reverse gear. Consequently, the features of urban atrophy begin to set in: derelict housing, the demise of public services (health, transport, education, etc.), planned shrinkage and an upsurge in societal epidemics (HIV, AIDS, obesity, depression, tuberculosis, etc.). A Markov game thus unfolds between city and suburbs [135, 137].

The ‘wrench in the works’ of social networking as foreseen by [63] is often the cause of certain epidemics as a recent report [29] on obesity suggests: from the embedding in a network, ‘social distance’, friendship (perhaps more so than within a family) and the network tolerance towards obesity appear at least as influential as heredity factors (such as an under-active thyroid gland). One might also argue that obesity is one of several epidemics realized at the negative end of social prosthetic systems, as much as toxic waste is to some ‘thriving’ chemical industry somewhere on the planet.

9.6 Red Queen versus the Pentagon Ratchet

Lewis Carroll’s ‘Red Queen’ has been taken metaphorically to describe an evolutionary system which “keeps running” in order to co-evolve with ambient competing systems. The analogy seems to be ubiquitous to modes of institutional cognition, economic game theory, arms races and predator–prey type models where advanced capabilities in one system are aimed to decrease those in the other. For instance, how a slowly evolving cognitive system has to gear itself to the constant threat of infectious epidemics. In a similar way it can be viewed as a contest between the internal cultures of a system (corporation, whatever) on one

hand, and associated external technologies, policies, legislation, outsourcing etc., on the other. As discussed in [135, 136] network (giant component) analysis reveals sudden and sharp phase transitions on passing critical points thus forecasting the most efficient co-evolutionary structure as gaining competitive strength. The Red Queen influences a multiple Workspace environment (e.g. one that is socio-economic, institutional or directed cognitive) and simultaneously interacts with a powerful mutual crosstalk creating a ‘ratchet-down’ effect. The latter has been coined the ‘Pentagon Ratchet’ [135, 136] a term suggestive of the legislative (re)allocation of major resources from the civilian into the military sectors. Likewise, the language of large-scale cognitive systems within an interactive environment may undergo a phase transition induced by intense crosstalk in reversal of their evolution. Thus the Ratchet gradually breaks down the competitive function of the Red Queen hence causing its sectors to become fragmented or to disintegrate altogether. Techniques involve the critical manifolds of differential game theory and explicit examples of renormalization modes leading to embeddings into state spaces, are exhibited in [135](§4.3) and [136](Chapter 4). Let us remark that the corresponding social network of the game may develop towards a ‘small worlds’ situation, and as previously mentioned, affording an enrichment of topological and statistical properties within a graph theoretic interpretation (cf [60, 120]).

The Red Queen (RQ) and Pentagon Ratchet (PR) are deemed to be interacting ‘principal environments’ for each other. In more general mathematical terms, let G_{RQ}, G_{PR} denote the corresponding groupoids of path equivalences and M_{RQ}, M_{PR} denote their respective set of objects (‘acquired characteristics’). Then in the competition we may regard the groupoids as acting on each other’s set of objects via crosstalk

$$\begin{aligned} G_{PR} &\curvearrowright M_{RQ} \longrightarrow M_{RQ} \\ G_{RQ} &\curvearrowright M_{PR} \longrightarrow M_{PR} \end{aligned} \tag{9.3}$$

thus yielding orbit spaces of generally lower dimension which, for instance, symbolize the curtailment/policy effect of one upon the other (the notion of a *groupoid action* is made specific in Appendix I §11.1). In view of the fundamental homology with ‘thermodynamic’ processes, the RQ along with small world networks are exemplified by certain distributed and institutional cognitive systems in [60].

There are several evolutionary scenarios tied to RQs and rate distortion theory that deserve mention. One such concerns a proposal by Eigen [47] of an evolutionary model which involves selections as a condensation in an information space. Some complications arise regarding the matter of genetic complexity since information has to be encoded in longer gene sequences by using replication with optimal fidelity. However, in order to do this, it is necessary to have a complex replication enzyme which just happens to be elusive, since such an enzyme will itself require a longer gene and the latter would violate an error threshold [71]. With the aim of resolving this paradox [141] employ a rate dis-

tortion argument coupled with a RQ coevolutionary ratchet toward establishing an evolutionary condensation that results in an effective error-correction mechanism. We refer to [141] for complete details.

9.7 Optimal coding and physiology: examples

For most species, and whether for predator or prey in particular, interaural time difference (ITD) is a characteristic property geared to localizing sound sources as crucial to the survival mechanism. Case studies have revealed optimal coding strategies depend not only on sound frequency ranges, but also on evolutionary driven physiological factors such as cranial size and form. Within groups of coincidence-detector neurons encoding ITD, each constituent member may be tuned for ITD in relationship to the ambient physiological range whereby exact tuning is determined by a time interval of axonal conduction in the auditory system. For pure tones, there is for each ITD an interaural phase difference (IPD) whereby an optimal coding strategy is seen to depend significantly on the relative width of the physiological range of individual IPDs in comparison to their corresponding tuning curves [65]. At the same time, we expect such strategies are significantly influenced by the behavioral patterns of the environmental stimuli; in this respect (auditory) receptors attain to optimal rather than average performance for most survival purposes [90].

Such systems are expected not to be free of corruption by noise. On the other hand, we have noted that noise, particularly in the case of ‘population’ based phenomena’, can engender a stochastic resonance which may favorably enhance and/or optimize the transmission of a weak signal via sensorimotor integration as shown, for instance, in certain cognitive studies of controlled visual stimulation [80] or ‘randomly enhanced’ human gaming strategies [144]. In [114] is considered the response of a neuron (in relationship to the cat primary visual cortex) on the linear filtering of the stimulus (luminance) values S by a linear receptive field L over space-time. In the usual network setting, a groupoid structure G can be revealed and the convolution $L \star S$ defined accordingly thus leading to a convolution algebra $\mathcal{C}(G)$ over a suitable class (of continuous) function on G . The output of the filter is then passed through a nonlinear function $h : \mathcal{C}(G) \rightarrow B$ such that the neurons response $R(t)$ is specified by $h(L \star S)$. Similar principles may be applied for explaining how activation receptors on registering a certain stimulus, transmit pulsations to the sensory cortex and assimilate the resulting meshwork of convoluted signals [54].

10 Conclusion

In this paper we have described a structural framework upon which rate distortion manifolds as representing cer-

tain cognitive modules, can be constructed by a variety of state-of-the-art mathematical concepts. We expect that implementing these concepts will lead to more exact, conceptually-centered, information-based models of cognition-at-large. The associated techniques as we have presented them, provide a method for the construction of a variety of information spaces structured by the Shannon coding and rate distortion theorems besides the means of describing globalization through local procedures. We have shown that the flexible, less rigid structures afforded to us by the notion of an atlas-manifold topology (or more generally, a groupoid atlas) along with simplicial/graph theoretic methods, can be adapted to a wide range of cognitive situations operative within the Global Workspace. This affords greater elegance and meaning to how these processes can be modeled without recourse to the traditional rigid, data-driven techniques which quite often can obscure some deeper underlying meaning.

In several instances we have employed the groupoid method as a category theory technique that allows one to reduce a vast labyrinthine configuration of networks to their corresponding sets of equivalence classes. The latter are computationally more user-friendly and create their own kind of dynamical systems via groupoid actions, (path) holonomy, etc. In particular, we should observe that the techniques we have outlined in the manifold/groupoid setting, are those suited to the description of ‘local-to-global processing’, seen for instance in the case of scheduling of paths and in the construction of the holonomy groupoid. In this way, the dynamics of cognitive processes (particularly those of the distributed and institutional type) can be aptly encoded in terms of groupoid actions as revealed, for instance, in the coevolutionary contest between the Red Queen and the Pentagon Ratchet. Likewise, we have seen how symmetry breaking of the network groupoid of linked cognitive modules cultivates a giant component which eventually emerges as a phase transition. In this respect, the fundamental homology describes close analogies between evolutionary modes (e.g. punctuated equilibria) influencing most cognitive processes, and the underlying dynamics of certain statistical-physical systems; more specifically, how alterations in the information network topology can induce phase-transitional states.

The geometry/topology of a rate distortion manifold thus represents the shape and form of information flow with respect to its syntactic-semantic content within the cultural environment of the particular Workspace through which it passes. In so far that the message transmits the channel, the former may be susceptible to cultural and evolutionary impingement. In a related way, a computational scheme of a cognitive process, may itself be deemed as a form of cognition. This leads us to questions of ‘higher categorical’ cognition rather befitting the ‘processes of processes’ as was alluded to earlier—clearly a matter that warrants further investigation.

Acknowledgement

We wish to thank the Overview Editor, Dr. Maria Ganzha, for her consideration of this paper.

11 Appendix I: Groupoids and their actions

Since groupoids and related actions have been pinpointed in the text, we provide the basic definitions and refer to [21, 31, 92, 143] for further details. Recall that a *groupoid* G is, loosely speaking, a small category with inverses over its set of objects $\text{Ob}(G)$. More specifically :

A groupoid consists of a set G with a distinguished subset denoted $G^{(0)} = \text{Ob}(G) \subset G$, called *the set of objects* of G , together with maps

$$r, s : G \overset{r}{\underset{s}{\rightrightarrows}} G^{(0)} \tag{11.1}$$

called the *range* and *source maps* respectively, together with a law of composition

$$\circ : G^{(2)} = \{(\gamma_1, \gamma_2) \in G \times G : s(\gamma_1) = r(\gamma_2)\} \longrightarrow G \tag{11.2}$$

on the set of ‘arrows’ $G^{(2)}$, such that the following hold:

- (1) $s(\gamma_1 \circ \gamma_2) = s(\gamma_2)$, $r(\gamma_1 \circ \gamma_2) = r(\gamma_1)$, for all $(\gamma_1, \gamma_2) \in G^{(2)}$.
- (2) $s(x) = r(x) = x$, for all $x \in G^{(0)}$.
- (3) $\gamma \circ s(\gamma) = \gamma$, $r(\gamma) \circ \gamma = \gamma$, for all $\gamma \in G$.
- (4) $(\gamma_1 \circ \gamma_2) \circ \gamma_3 = \gamma_1 \circ (\gamma_2 \circ \gamma_3)$.
- (5) Each γ has a two-sided inverse γ^{-1} with $\gamma\gamma^{-1} = r(\gamma)$, $\gamma^{-1}\gamma = s(\gamma)$.

Often one denotes by $G_x^y = s^{-1}(x) \cap r^{-1}(y)$ the set of morphisms in G from x to y , and G_x^x denotes the isotropy group at $x \in G^{(0)}$.

Example 11.1. An equivalence relation \mathcal{R} on a set X can constitute a groupoid in the following way. Specifically, $\mathcal{R} \subset X \times X$ is identifiable with the set of ordered pairs (x, y) satisfying $x\mathcal{R}y$, whereby the morphisms are

$$\mathcal{R}_y^x = \begin{cases} \{(x, y)\} , & \text{if } x\mathcal{R}y, \\ 0, & \text{otherwise.} \end{cases} \tag{11.3}$$

The composition is given by

$$\begin{aligned} \circ : \mathcal{R}_y^x \times \mathcal{R}_z^y &\longrightarrow \mathcal{R}_z^x, \\ (x, y) \circ (y, z) &= (x, z), \end{aligned} \tag{11.4}$$

where (x, x) is the identity and $(x, y)^{-1} = (y, x)$. Accordingly, the orbit $\mathcal{R}(x)$ is the equivalence class of $x \in X$.

Conversely, a groupoid G may induce an equivalence relation \mathcal{R} on the set X , for which the equivalence classes $\mathcal{R}(x)$ are the orbits $G(x)$, for all x in X . This is subject to a forgetful functor $F : G \rightarrow \mathcal{R}$, such that $F(g) = (y, x)$ if and only if $g \in G_x^y$.

Example 11.2. Clearly, any group is a groupoid whereby the object set consists of the single element $\{e\}$ the identity (e.g the fundamental group $\pi_1(M)$ of a manifold M , is a groupoid). Thus groupoids may be seen as consisting of ‘multiple identities’. Indeed, any manifold M can be viewed as a groupoid over itself where all morphisms are units (that is, the arrow set of M is M itself). We also have the pair groupoid $M \times M \rightrightarrows M$ where the natural projections from each factor comprise the range and source maps.

11.1 Groupoid actions

Let X be a topological space admitting an action ‘ \curvearrowright ’ of a group G . Specifically $\curvearrowright : X \times G \rightarrow X$, with $\curvearrowright(x, g) = xg$ and $x(g_1g_2) = (xg_1)g_2$, for all $x \in X$ and $g \in G$. Here we have a natural groupoid $G = X \times G$ with $G^{(0)} = X \times \{1\}$, and for which the following conditions hold:

- (1) $r(x, g) = x, s(x, g) = xg$, for all $(x, g) \in X \times G$.
- (2) $(x, g_1)(y, g_2) = (x, g_1g_2)$ if $xg_1 = y$.
- (3) $(x, g)^{-1} = (xg, g^{-1})$, for all $(x, g) \in X \times G$.

Consider a groupoid $G \rightrightarrows B$ over its set of objects $B = G^{(0)}$. Let M be a topological space and $f : M \rightarrow B$ a continuous map. Consider the set

$$G \curvearrowright M = \{ (g, u) \in G \times M : sg = f(u) \} \subset G \times M. \tag{11.5}$$

An action of G on (M, f, B) is a continuous map $G \times M \rightarrow M$, given by $(g, u) \mapsto gu$ satisfying:

- (1) $f(gu) = rg$, for all $(g, u) \in G \curvearrowright M$.
- (2) $h(gu) = (hg)u$, for all $(h, g) \in G \times G, (g, u) \in G \curvearrowright M$.
- (3) $f(\widetilde{u})u = u$, for all $u \in M$, where \sim denotes the corresponding groupoid isomorphism.

We call $G \curvearrowright M \rightarrow M$ the action groupoid. For $u \in M$, the subset $G[u] = \{gu : g \in G\}$, is the orbit of u under G . These concepts generalize the notion of a group action on a topological space.

11.2 Proper groupoids and orbifolds

Firstly, G is said to be a Lie groupoid when $G^{(0)}$ and $G^{(2)}$ have the structures of differentiable (Hausdorff) manifolds, the map s is a differentiable submersion (with Hausdorff fibers), and all other structure maps are differentiable. A

Lie groupoid is said to be proper if it is Hausdorff and the map $(s, r) : G^{(2)} \rightarrow G^{(0)} \times G^{(0)}$ is proper (that is, each inverse image of a compact subset is compact).

An orbifold atlas of dimension n of a topological space Q is a collection of pairwise compatible orbifold charts

$$\mathcal{U} = \{(U_i, G_i, \phi_i)\}_{i \in \mathcal{I}}, \tag{11.6}$$

of dimension n on Q , where the $G_i \subset \text{Diff}(U_i)$ are finite subgroups, such that $\bigcup_{i \in \mathcal{I}} \phi_i(U_i) = Q$. Two orbifold atlases of Q are equivalent if their union is an orbifold atlas. Then an orbifold of dimension n is a pair (Q, \mathcal{U}) when Q is a (second countable) Hausdorff topological space and \mathcal{U} is a maximal orbifold atlas of dimension n of Q . For further details see [96]. In particular, there is an associated pseudogroup of transitions $\Psi(\mathcal{U})$ and an effective proper groupoid $\Gamma(\mathcal{U}) = \Gamma(\Psi(\mathcal{U}))$ associated to $\Psi(\mathcal{U})$ (see [96] §5.6).

12 Appendix II: Briefly simplicial complexes and triangulations

Let K be a simplicial complex, that is, K contains a set of objects $V(K)$ called vertices and a set of non-empty subsets of $V(K)$ called simplices. If $\sigma \subset V(K)$ is a given simplex and $\kappa \subset \sigma, \kappa \neq \emptyset$, then κ is also a simplex. The geometric realization (or polyhedron) of K , denoted $|K|$, comprises the set of all functions $V(K) \rightarrow [0, 1]$, such that:

- 1. If $\alpha \in |K|$, the set $\{v \in V(K) : \alpha(v) \neq 0\}$ is a simplex of K .
- 2. $\sum_{v \in V(K)} \alpha(v) = 1$.

If $s \in K$, we let $|s|$ denote the set

$$|s| = \{\alpha \in |K| : \alpha(v) \neq 0 \Rightarrow v \in s\}, \tag{12.1}$$

and

$$\langle s \rangle = \{\alpha \in |K| : \alpha(v) \neq 0 \Leftrightarrow v \in s\}. \tag{12.2}$$

We call $\alpha(v)$ the v -th barycentric coordinate of α and $p_V(\alpha) = \alpha(v) : |K| \rightarrow [0, 1]$ is the corresponding v -th barycentric projection. A metric $d(\cdot, \cdot)$ can be defined on $|K|$ by setting

$$d(\alpha, \beta) = \left[\sum_{v \in V(K)} (p_v(\alpha) - p_v(\beta))^2 \right]^{\frac{1}{2}}, \tag{12.3}$$

with the resulting initial topology for barycentric projections.

Often it is useful to pass to a subdivision of a given simplicial complex K . A subdivision of K is a simplicial complex K' satisfying:

- a) the vertices of K' are identified as points of $|K|$;
- b) if s' is a simplex of K' , there exists a simplex s in K such that $s' \subset |s|$; and

c) the map $|K'| \rightarrow |K|$ extending the map of vertices of K' to their corresponding points of $|K|$, is a homeomorphism.

Let K_p denote the set of p -simplices of K . If $\sigma = \{v_0, \dots, v_p\} \in K_p$, then its *barycenter* $b(\sigma)$ is the point

$$b(\sigma) = \sum_{0 \leq i \leq p} \frac{1}{p+1} v_i \in |K|. \tag{12.4}$$

Accordingly, we may speak of a *barycentric subdivision* \widehat{K} of K as the simplicial complex with vertices the barycenters of the simplices of K , and whose simplices are finite non-empty collections of barycentres of simplices totally ordered by the face relations of K .

Suppose now $\mathcal{U} = \{U_\lambda : \lambda \in \Lambda\}$ is an open covering of M . We assign an ordering to the indexing set Λ of the cover and then let

$$\Lambda^{(n)} = \{(\lambda_0, \lambda_1, \dots, \lambda_{n-1}) \in \Lambda : \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}\}. \tag{12.5}$$

The *nerve* $N\mathcal{U}$ of the cover \mathcal{U} is defined as follows. Firstly, let

$$(N\mathcal{U})_n := \bigsqcup_{\nu \in \Lambda^{(n+1)}} U_\nu, \tag{12.6}$$

where $U_\nu = U_{\lambda_0} \cap \dots \cap U_{\lambda_n}$. In this way an element of $(N\mathcal{U})_n$ consists of an $(n+2)$ -tuple $(x, \lambda_0, \dots, \lambda_n)$, where $x \in U_\nu$ and $\nu = (\lambda_0, \dots, \lambda_n) \in \Lambda^{(n+1)}$. Then the nerve of \mathcal{U} is given by $N\mathcal{U} := \lim_n (N\mathcal{U})_n$.

12.1 Triangulations

A *triangulation* (K, ϕ) of a space M means we have a simplicial complex K together with a homeomorphism $\phi : |K| \rightarrow M$. For any vertex v in K , we define its (*open*) *star* by

$$\text{st}(v) = \{\alpha \in |K| : \alpha(v) \neq \emptyset\}. \tag{12.7}$$

Alternatively,

$$\text{st}(v) = \bigcup \{s : v \text{ is a vertex of } s\}, \tag{12.8}$$

that is, the union of interiors of all simplices having s as a vertex. Note that $\mathcal{U} = \{\text{st}(v) : v \in K\}$ provides an open covering of $|K|$. References to these topics are [20, 76, 117].

13 Appendix III: Differentiable structures on path space

13.1 Plots and iterated integrals

Let us recall the state space E^Γ over the alphabet Γ . In general, we do not expect E^Γ to have a differentiable structure in the conventional sense of classical calculus, but one of several concepts of abstract ‘differentiable spaces’, might

be applicable. One such structure uses an abstract notion of ‘plots’ [27], permitting a ‘differentiable space’ structure on E^Γ in terms of the following conditions. We consider a collection of maps $f : \mathbb{R}^n \rightarrow E^\Gamma$ (where n can be arbitrarily large), called *plots*, such that:

1. If $f : \mathbb{R}^n \rightarrow E^\Gamma$ is a plot, and if $g \in C^\infty(\mathbb{R}^m, \mathbb{R}^n)$ is a smooth map in the usual sense, then $f \circ g$ is a plot.
2. If $g_\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a collection of embeddings whose images cover \mathbb{R}^n , and $f : \mathbb{R}^n \rightarrow E^\Gamma$ is a map such that $f \circ g_\alpha$ is a plot, then f is also a plot.
3. Every map $f : \mathbb{R}^0 \rightarrow E^\Gamma$ is a plot.

Given another such differentiable space Y , a map $\psi : E^\Gamma \rightarrow Y$ is said to be *differentiable* if for every plot f in E^Γ , $f \circ \psi$ is a plot in Y .

Relevant here is that the information (path) space $X = \mathcal{P}(E^\Gamma)$, supposedly as a length space, with its rate distortion measure s_X , could be assumed as endowed with a differentiable space structure as well. Effectively, we can view a path in E^Γ as a plot of the type $\gamma : I \rightarrow E^\Gamma$. For every set map $\alpha : U \rightarrow X = \mathcal{P}(E^\Gamma)$, there is a corresponding suspension map

$$\lambda_\alpha : I \times U \rightarrow E^\Gamma \tag{13.1}$$

$$(t, \xi) \mapsto \alpha(\xi)(t).$$

Then $X = \mathcal{P}(E^\Gamma)$ can be viewed as a differentiable space when assigned plots of the type $\alpha : U \rightarrow X$, such that λ_α is a plot of E^Γ .

Suppose that \mathcal{A} denotes some suitable (alphabetical or events) algebra and that E^Γ admits some choice of algebra \mathcal{A} -valued 1-forms $w_1, \dots, w_r \in \Omega^1(E^\Gamma, \mathcal{A})$, then once given a path $\gamma : I \rightarrow E^\Gamma$ of sufficient differentiability, [27] introduces the notion of *iterated integrals*

$$\int_\gamma w_1, \dots, w_r := \int_0^1 f_1 dt_1, \dots, f_r dt_r, \tag{13.2}$$

where $f_i(t) = w_i(\gamma(t), \gamma'(t))$, or in terms of the pull-back, $\gamma^* w_i = f_i(t) dt$. Subsequently, this defines a map

$$\int_\gamma : (X, s_X) \rightarrow \mathcal{A}. \tag{13.3}$$

This makes the same sense if we replace (X, s_X) by its canonical model (M, s_M) :

$$\int_\gamma : (M, s_M) \rightarrow \mathcal{A}. \tag{13.4}$$

Higher degree (differential) forms can be treated accordingly.

13.2 Fröhlicher spaces

The above notion of ‘differentiability’ via plots and iterated integrals is a relatively weak one that may be suited to rate

distortion theory. There are other possibilities that provide an approach to calculus on spaces more ‘pathological’ than standard differentiable manifolds (such as to be found in e.g. [55, 81]). For instance in [55], a Fröhlicher space X consists of a triple $(X, \mathcal{C}_X, \mathcal{F}_X)$ where X is a set, \mathcal{C}_X is a subset of all mappings $\mathbb{R} \rightarrow X$, and \mathcal{F}_X is the set of all functions $X \rightarrow \mathbb{R}$, satisfying the properties:

- (1) A function $f : X \rightarrow \mathbb{R}$ belongs to \mathcal{F}_X if and only if $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$, for all $c \in \mathcal{C}_X$.
- (2) A map $c : \mathbb{R} \rightarrow X$ belongs to \mathcal{C}_X if and only if $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$, for all $f \in \mathcal{F}_X$.

A morphism of $\varphi : X \rightarrow Y$ of Fröhlicher spaces X, Y is said to be *smooth* when the following three equivalent conditions hold:

- (1) For each $c \in \mathcal{C}_X$, the composition $\varphi \circ c \in \mathcal{C}_Y$.
- (2) For each $f \in \mathcal{F}_X$, the composition $f \circ \varphi \in \mathcal{F}_Y$.
- (3) For each $c \in \mathcal{C}_X$, and for each $f \in \mathcal{F}_X$, the composition $f \circ \varphi \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$.

In short, the space X admits an admissible family of curves \mathcal{C}_X together with an admissible family of functions \mathcal{F}_X whose respective elements satisfy the above conditions.

References

- [1] R. Abraham and J. E. Marsden (1978), *Foundations of Mechanics* (2nd Ed.) Benjamin/Cummings Publ., Reading MA.
- [2] R. Albert and A.–L. Barabási (2002), Statistical mechanics of complex networks, *Rev. Mod. Phys.* 74, 47–97.
- [3] D. Alonso, A. J. McKane and M. Pascual (2007), Stochastic amplification in epidemics, *J. R. Soc. Interface* 4, 575–582.
- [4] M. E. – S. A. Aof and R. Brown (1992), The holonomy groupoid of a locally topological groupoid, *Top. Appl.* 47, 97–113.
- [5] J. Anandan (1992), The geometric phase, *Nature* 360, 307–313.
- [6] R. Ash (1990), *Information Theory*, Dover Publications, New York.
- [7] H. Atlan (1972), *L’Organisation biologique et la théorie de l’information*, Hermann, Paris.
- [8] H. Atlan and I. R. Cohen (1998), Immune information, self–organization and meaning, *International Immunology* 10, 711–717.
- [9] H. Atlan (1979), *Entre le cristal et la fumée*, Seuil, Paris.
- [10] H. Atmanspacher and R. G. Jahn (2003), Problems of reproducibility in complex mind matter systems, *J. of Scientific Exploration* 17(2), 243–270.
- [11] R. Attal (2004), Combinatorics of non–abelian gerbes with connection and curvature, *Ann. Fond. Louis de Broglie* 29 (4), 609–633.
- [12] B. J. Baars (1988), *A Cognitive Theory of Consciousness*, Cambridge University Press.
- [13] B. J. Baars and J. Newman (1994), A neurobiological interpretation of the Global Workspace theory of consciousness, in *Consciousness in Philosophy and Cognitive Neuroscience* (A. Revonsuo and M. Kumpinen, eds.) Erlbaum, Hillsdale NJ.
- [14] B. J. Baars and S. Franklin (2003), How conscious experience and working memory interact, *Trends in Cognitive Science* 7, 166–172.
- [15] I. C. Baianu, R. Brown, G. Georgescu, and J. F. Glazebrook (2006), Complex nonlinear biodynamics in categories, higher dimensional algebra and Lukaszewicz–Moisil topos: transformations of neuronal, genetic and neoplastic networks, *Axiomathes* 16 Nos. 1–2, 65–122.
- [16] A. Bak, R. Brown, G. Minian and T. Porter (2006), Global actions, groupoid atlases and related topics, *J. of Homotopy and Related Structures* 1, 1–54.
- [17] M. Bennett and P. Hacker (2003), *Philosophical Foundations of Neuroscience*, Blackwell Publishing, London.
- [18] E. D. Bolker, V. W. Guillemin and T. S. Holm, How is a graph like a manifold?, (to appear) <http://arxiv:math.CO/0206103>
- [19] F. Borceux (1994), *Handbook of Categorical Algebra (1 and 2)– Encyclopedia of Mathematics and its Applications* 50 & 51, Cambridge University Press.
- [20] R. Bott and L. W. Tu (1982), *Differential Forms in Algebraic Topology*, Grad. Texts in Math. 82, Springer Verlag.
- [21] R. Brown (2006) *Topology and Groupoids*, Book-Surge LLC.
- [22] R. Brown, R. Paton and T. Porter (2004), Categorical language and hierarchical models for cell systems, in *Computation in Cells and Tissues - Perspectives and Tools of Thought*, Paton, R.; Bolouri, H.; Holcombe, M.; Parish, J.H.; Tateson, R. (Eds.) Natural Computing Series, Springer Verlag, 289–303.
- [23] R. Brown and İ. İcen (2003), Towards a 2–dimensional notion of holonomy, *Advances in Math.* 178, 141–175.

- [24] D. Burago, Y. Burago and S. Ivanov (2001), *A Course in Metric Geometry*, Grad. Studies in Math. 33, Amer. Math. Soc., Providence RI.
- [25] D. J. Chalmers (1996), *The Conscious Mind—In search of a fundamental theory*, Oxford University Press.
- [26] A. Candel and L. Conlon (2000), *Foliations I*, Graduate Studies in Mathematics 23, Amer. Math. Soc., Providence RI.
- [27] K. T. Chen (1977), Iterated path integrals, *Bull. Amer. Math. Soc.* 83 No. 5, 831–879.
- [28] D. Chigirev and W. Bialek (2004), Optimal manifold representation of data : an information theoretic approach, in *Advances in Neural Information Processing Systems* 16 161–168, MIT Press, Cambridge MA.
- [29] N. A. Christakis and J. H. Fowler (2007), The spread of obesity in a large social network over 32 years, *New England J. Medicine* 357 (4), 370–379.
- [30] A. Clark (1997), *Being there: Putting brain, body and world together again*, MIT Press, Cambridge MA.
- [31] A. Connes (1994), *Noncommutative Geometry*, Academic Press.
- [32] T. Cover and J. Thomas (1991), *Elements of Information Theory*, Wiley, New York.
- [33] F. Crick and C. Koch (1990), Toward a neurobiological theory of consciousness, *The Neurosciences*, 263–275.
- [34] C. J. Davia (2006), Life, catalysis and excitable media: A dynamic systems approach to metabolism and cognition, in (J. A. Tuszynski, ed.) *The Emerging Physics of Consciousness*, Springer Berlin–Heidelberg, 255–292.
- [35] H. de Jong and M. Page (2000), Qualitative simulation of large and complex genetic regulatory systems, in *Proc. of Fourteenth European Conference on Artificial Intelligence, ECAI 2000*, W. Horn (ed.), IOS Press, Amsterdam, 141–145.
- [36] A. Dembo and O. Zeitouni (1993), *Large Deviation Techniques and Applications*, Jones and Bartlett Publ. Boston–London.
- [37] S. Dehaene and L. Naccache (2001), Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework, *Cognition* 79, 1–37.
- [38] S. Dehaene and J.-P. Changeux (2005), Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentive blindness, *PLoS Biology* 3 (5), 1–18.
- [39] F. Dretske (1981), *Knowledge and the Flow of Information*, MIT Press, Cambridge MA.
- [40] M. L. Dustin and J. A. Cooper (2000), The immunological synapse and the actin cytoskeleton: molecular hardware for T-cell signaling, *Nature Immunology* 1 No. 1, 23–29.
- [41] E. Dyer and S. Eilenberg (1988), Globalizing fibrations by schedules, *Fund. Math.* 130, 125–136.
- [42] G. Edelman (1992), *Brilliant Air, Brilliant Fire – On the Matter of the Mind*, Basic Books, New York.
- [43] G. Edelman (1989), *The Remembered Present*, Basic Books, New York.
- [44] G. Edelman and G. Tononi (2000), *A Universe of Consciousness*, Basic Books, New York.
- [45] J. Eells and L. Lemaire (1983), *Selected topics in harmonic maps*, CBMS Regional Conf. Series, Amer. Math. Soc., Providence RI.
- [46] A. C. Ehresmann and J.-P. Vanbremeersch (2007), *Memory Evolutionary Systems: Hierarchy, Emergence, Cognition*, Elsevier.
- [47] M. Eigen (1996), *Steps towards life: A perspective on evolution*, Oxford University Press, New York.
- [48] D. Ellerman (2007), Adjoints and emergence: applications of a new theory of adjoint functors, *Axiomathes* 17, 19–39.
- [49] M. Emery (1989), *Stochastic Calculus in Manifolds*, Springer Verlag, Berlin Heidelberg.
- [50] P. Erdős and R. Rényi, A. (1960), On the evolution of random graphs. Reprinted in *The Art of Counting* 1973 (pp. 574–618). Cambridge, MA: MIT Press. Also in *Selected Papers of Alfred Rényi* (pp. 482–525) (1976). Budapest: Akadémiai Kiado.
- [51] L. Fajstrup (2005), Dipaths and dihomotopies in a cubical complex, *Adv. in Appl. Math.* 35, 188–206.
- [52] U. Feudel et al. (2000), Homoclinic bifurcation in a Hodgkin–Huxley model of thermally sensitive neurons, *Chaos* 10 No. 1, 231–239.
- [53] R. Feynman (1996), *Feynman Lectures on Computation*, Addison–Wesley, Reading MA.
- [54] W. J. Freeman (1999), Consciousness, intentionality and causality, *J. Consciousness Studies* 11 (1999), 143–172.
- [55] A. Frölicher and A. Kriegl (1988), *Linear Spaces and Differentiation Theory*, J. Wiley, Chichester.
- [56] D. Gabor (1946), Theory of communication *J. IEE (London)* 93 (III), 429–457.

- [57] L. Gammaitoni, P. Hänggi, P. Jung and F. Marchesoni (1998), Stochastic resonance, *Rev. Mod. Phys.* 70 No. 1, 223–287.
- [58] W. Gerstner and W. Kistler (2002), *Spiking Neuron Models*, Cambridge University Press.
- [59] T. Giordano, I. Putnam and C. Skau (2004), Affable equivalence relations and orbit structure of Cantor dynamical systems, *Ergodic Theory Dynam. Systems* 24 No. 2, 441–475.
- [60] J. F. Glazebrook and R. Wallace (2009), Small Worlds and Red Queens in the Global Workspace: an information–theoretic approach, *Cognitive Systems Research*, to appear.
- [61] M. Golubitsky and I. Stewart (2006), Nonlinear dynamics of networks: the groupoid formalism, *Bull. Amer. Math. Soc.* 43, No. 3, 305–364.
- [62] E. Goubault (2003), Some geometric perspectives in concurrency theory, *Homology, Homotopy and Applications* 5 (2), 95–136.
- [63] M. S. Granovetter (1973), The strength of weak ties, *Amer. J. Sociol.* 78, 1360–1380.
- [64] S. Gudder (2004), Noncommutative probability and applications, *Real and Stochastic Analysis—New Perspectives* (M. Rao, ed.), pp 199–238, Birkhäuser, Boston–Basel–Berlin.
- [65] N. S. Harper and D. McAlpine (2004), Optimal neural population coding of an auditory spatial cue, *Nature* 430, 682–686.
- [66] M. J. Healy and T. P. Caudell (2006), Ontologies and worlds in category theory: Implications for neural systems, *Axiomathes* 16 Nos. 1–2, 165–214.
- [67] M. J. Healy and T. P. Caudell (2004), Neural networks, knowledge and cognition: A mathematical semantic model based upon category theory. Technical report EECE-TR-04-20, Dept. of Electrical and Computer Engineering, University of New Mexico.
- [68] J. Heinloth (2004), Some notes on differentiable stacks, *Math. Inst. Sem.* (Y. Tschinkel, ed.), 1–32, Univ. Göttingen.
- [69] B. Hillier (2003), The knowledge that shapes the city, in *Proc. 4th International Space Syntax Symposium*, London.
- [70] J. Hollan, E. Hutchins and D. Kirsh (2000), Distributed cognition: Toward a new foundation for human–computer interaction research, *ACM Transactions on Computer Interaction* 7 (2), 174–196.
- [71] E. Holmes (2005), On being the right size, *Nature Genetics* 37, 543–546.
- [72] F. C. Hoppensteadt and E. M. Izhikevich (1997), *Weakly Connected Neural Networks*, Appl. Math Sciences 126, Springer Verlag.
- [73] P. W. Huber (1987), The Geodesic Network – 1987 Report on competition in the telephone industry, *U.S. Dept. of Justice*, U. S. Gov. Printing Office, Washington DC.
- [74] E. Hutchins (1994), *Cognition in the Wild*, MIT Press, Cambridge MA.
- [75] N. K. Jerne (1974), Towards a network theory of the immune system, *Ann. Immunology (Inst. Pasteur)*, 125C, 373–389.
- [76] H. Kamps and T. Porter (1996), *Abstract Homotopy and Simple Homotopy Theory*, World Scientific, Singapore.
- [77] A. Katok and B. Hasselblatt (1995), *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press.
- [78] G. M. Kelly and R. Street (1974), Review of the elements of 2–categories, in *Category Seminar (Proc. Sem. Sydney, 1972/1973)*, *Lect. Notes in Math.* 420, Springer, Berlin, 75–103.
- [79] A. I. Khinchin (1957), *Mathematical Foundations of Information Theory*, Dover Publ.
- [80] K. Kitajo, D. Nozaki, L. M. Ward and Y. Yamamoto (2003), Behavioral stochastic resonance within the human brain, *Phys. Rev. Lett.* 90 No. 21, 218103, 1–4.
- [81] A. Kock (2006), *Synthetic Differential Geometry*, Cambridge Univ. Press.
- [82] A. Kock (2006), Connections and path connections in groupoids, Aarhus Univ. Math. Preprint No. 10.
- [83] S. M. Kosslyn (2007), On the evolution of human motivation: The role of social prosthetic systems, in (Platek et al. eds) *Evolutionary Cognitive Neuroscience*, MIT Press.
- [84] R. Kozma, M. Puljic, P. Balister, B. Bollobas and W. J. Freeman (2004), Neuropercolation: a random cellular automata approach to spatio–temporal neurodynamics, *Lecture notes in Computer Science* 3305, 435–443.
- [85] N. P. Landsman (1998), *Mathematical Topics between Classical and Quantum Mechanics*, Springer Verlag, New York.
- [86] G. Laumon and L. Moret-Bailly (2000), Champs algébriques, *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge.*, vol. 39, Springer-Verlag, Berlin.

- [87] J. Lee (2000), *Introduction to topological manifolds*, Graduate Texts in Math. 202, Springer-Verlag, New York.
- [88] M. Li and P. Vitányi (1993), *An Introduction to Kolmogorov complexity and its applications*, Texts and Monographs in Computer Science, Springer-Verlag, New York.
- [89] D. Lind and B. Marcus (1995), *An Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press.
- [90] C. K. Machens, T. Gollisch, O. Kolesnikova and A. V. Herz (2005), Testing the efficiency of sensory coding with optimal stimulus ensembles, *Neuron* 47 (3), 447–456.
- [91] S. Mac Lane and I. Moerdijk (1992), *Sheaves in Geometry and Logic: A first Introduction to Topos Theory*, Springer Verlag, New York.
- [92] K. C. H. Mackenzie (1987), *Lie Groupoids and Lie Algebroids in Differential Geometry*, London Math. Soc. Lect. Notes 124, Cambridge University Press.
- [93] H. R. Maturana and F. J. Varela (1980), *Autopoiesis and Cognition—The Realization of the Living*, Boston Studies in the Philosophy of Science Vol. 42, Reidel Pub. Co., Dordrecht.
- [94] J. Mayor and W. Gerstner (2005), Noise-enhanced computation in a model of cortical column, *NeuroReport* 16 (11), 1237–1240.
- [95] J. Milnor (1963), *Morse Theory*, Princeton University Press, Princeton NJ.
- [96] I. Moerdijk and J. Mrčun (2003), *Introduction to Foliations and Lie Groupoids*, Cambridge Studies in Adv. Math. 91, Cambridge Univ. Press.
- [97] R. Montgomery (1990), Isoholonomic problems and applications, *Commun. Math. Phys.* 128 (3), 565–592.
- [98] T. Murayama and M. Okada (2003), One step RSB scheme for the rate distortion function, *J. Phys. A: Math. Gen.* 36, 11123–11130.
- [99] R. Nisbett (2003), *The Geography of Thought: How Asians and Westerners think differently ... and why*, Free Press, New York.
- [100] J. Peebles, I. F. Putnam and I. F. Zwiars, A survey of orbit equivalence for Cantor minimal dynamics, to appear.
- [101] G. Pistone and C. Sempì (1995), An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one, *Annals of Statistics* 25 (5), 1543–1561.
- [102] W. F. Pohl (1962), Differential geometry of higher order, *Topology* 1, 169–211.
- [103] T. Porter (2008), Enriched categories and models for spaces of evolving states, *Theoretical Computer Science* 405, No. 1-2, 88–100.
- [104] A. Pouget, P. Dayan and R. Zemel (2000), Information processing within population codes, *Nature Rev. Vol 1*, 125–132.
- [105] V. Pratt (1991), Modeling concurrency with geometry, in *Proceedings of the 18th ACM Symposium on Principles of Programming Languages*.
- [106] V. Pratt (2003), Chu spaces as a semantic bridge between linear logic and mathematics, *Theoretical Comp. Sci.* 294, 439–471.
- [107] K. H. Pribram (1991), *Brain and Perception: Holonomy and Structure in Figural Processing*, Lawrence Erlbaum Assoc., Hillsdale NJ.
- [108] I. Prigogine (1980), *From Being to Becoming: Time and Complexity in the Physical Sciences*, W. H. Freeman and Co., San Francisco.
- [109] G. Robins, P. Pattison and J. Woolcock (2005), Small and other worlds: Global structures from local processes, *Amer. J. Sociology* 110 (4), 894–936.
- [110] R. Sampson (2004), Networks and neighborhoods, in *Network Logic: Who governs in an Interconnected World?* (H. McCarthy, P. Miller and R. Skidmore, eds.), Demos, London.
- [111] E. Schneidman, W. Bialek and M. J. Berry II (2002), An information theoretic approach to the functional classification of neurons, *Advances in Neural Information Processing Systems* 15 197–204, MIT Press, Cambridge MA.
- [112] C. E. Shannon (1948), A mathematical theory of communication, *Bell Syst. Tech. J.* 30, 379–423, 623–656.
- [113] V. V. Sharko (1993), Functions on manifolds, algebraic and topological aspects, *Amer. Math. Soc. Translations* 131, Amer. Math. Soc., Providence RI.
- [114] T. Sharpee et al. (2006), Adaptive filtering enhances information transmission in visual cortex, *Nature* 439, 936–942.
- [115] D. J. Simons and C. F. Chabris (1999), Gorillas in our midst: Sustained inattentive blindness for dynamic events, *Perception* 28, 1059–1074.
- [116] N. Surlas (1989), Spin-glass models as error-correcting codes, *Nature* 339, 693–695.
- [117] E. H. Spanier (1966), *Algebraic Topology*, McGraw Hill.

- [118] I. Stewart, M. Golubitsky and M. Pivato (2003), Symmetry groupoids and patterns of synchrony in coupled cell networks, *Siam J. Appl. Dynamics* 2 No. 4, 609–646.
- [119] R. Sun (2009), Theoretical status of computational cognitive modeling, *Cognitive Systems Research* 10 No. 2, 124–140.
- [120] G. Szabó and G. Fáth (2007), Evolutionary games on graphs, *Phys. Rep.* 446, 97–216.
- [121] J. B. Tenenbaum, V. de Silva and J. C. Langford (2000), A global geometric framework for nonlinear dimensionality reduction, *Science* 290, 2319–2323.
- [122] J. T. Townsend, B. Solomon and J. S. Smith (2001), The perfect Gestalt: Infinite dimensional Riemannian face spaces and other aspects of face perception. In M. J. Wenger and J. T. Townsend (Eds.) *Computational geometric and process perspectives of facial cognition: Contexts and challenges*, 39–82, Mahwah, NJ: Lawrence Erlbaum Associates.
- [123] J. Virsik (1971), On the holonomy of higher order connections, *Cah. Top. Géom. Diff. Cat.* 12, 197–212.
- [124] E. Volz (2004), Random networks with tunable degree distribution and clustering, *Phys. Rev. E* 70, 056115.
- [125] R. Wallace (2005), *Consciousness: A Mathematical Treatment of the Global Neuronal Workspace*, Springer, New York.
- [126] R. Wallace (2005), A global workspace perspective on mental disorders, *Theoretical Biology and Medical Modelling* 2, 49.
- [127] R. Wallace (2005), The sleep cycle: a mathematical analysis from a global workspace perspective. <http://cogprints.org/4517/>
- [128] R. Wallace (2007), Culture and inattention blindness: a global workspace perspective, *Journal of Theoretical Biology*, 245, 378–390.
- [129] R. Wallace (2006), Public policy, institutional cognition and the geographic diffusion of multiple-drug resistant HIV in the United States, preprint.
- [130] R. Wallace (2006) New mathematical foundations for AI and Alife: Are the necessary conditions for animal consciousness sufficient for the design of intelligent machines?, preprint.
- [131] R. Wallace (2006), Institutional cognition, preprint.
- [132] R. Wallace (2006), Varieties of animal consciousness: the spectrum of mathematical Global Workspace models and its implications, preprint.
- [133] R. Wallace (2002), Adaptation, punctuation, and information: a rate distortion approach to non-cognitive ‘learning plateaus’ in evolutionary process, *Acta Biotheoretica* 50, 101–116.
- [134] R. Wallace (2008), Toward formal models of biologically inspired, highly parallel machine cognition, *International Journal of Emergent, Parallel and Distributed Systems* 23, 367–408.
- [135] R. Wallace and M. Fullilove (2008), *Collective Consciousness and Its Discontents: Institutional distributed cognition, racial policy, and public health in the United States*, Springer, New York.
- [136] R. Wallace and R. G. Wallace, *Psychopathica Automatorum: A cognitive neuroscience perspective on highly parallel computation and its dysfunctions*, to appear.
- [137] D. Wallace and R. Wallace (1998), *A Plague on Your Houses*, Verso Press, New York.
- [138] R. Wallace and R. G. Wallace (1998), Information theory, scaling laws and the thermodynamics of evolution, *J. Theor. Biology* 192, 545–559.
- [139] R. Wallace and D. Wallace (2004), Structured psychosocial stress and therapeutic failure, *Journal of Biological Systems* 12, 335–369.
- [140] R. Wallace, D. Wallace and R. G. Wallace (2003), Towards cultural oncology: the evolutionary information dynamics of cancer, *Open systems and Information Dynamics* 10, 159–181.
- [141] R. Wallace and R. G. Wallace (2008), On the spectrum of prebiotic chemical systems: An information theory treatment of Eigen’s paradox, *Origins of Life and Evolution of Biospheres* 38 No. 5, 419–4551.
- [142] D. J. Watts (1999), *Small Worlds: The Dynamics of networks between Order and Randomness*, Princeton Univ. Press, Princeton, NJ.
- [143] A. Weinstein (1996), Groupoids: unifying internal and external symmetry, *Notices of the Amer. Math. Soc.* 43, 744–752.
- [144] R. L. West, T. C. Stewart, C. Lebiere and S. Chandrasekharan (2005), Stochastic resonance in human cognition: ACT-R versus game theory, associative neural networks, recursive neural networks, Q-learning and humans, in *Proc. 27th Annual Meeting of The Cognitive Science Soc.*, 2353–2358.
- [145] J. Zhang and P. Hästö (2006), Statistical manifold as an affine space: A functional equation approach, *J. Math. Psychology* 50(1), 60–65.
- [146] R. Živaljević (2008), Combinatorial groupoids, cubical complexes, and the Lovász conjecture, *Discrete and Computational Geometry* 41(1), 135–161.

Comparative Study of Tripartite Identity-Based Authenticated Key Agreement Protocols

Marko Hölbl, Tatjana Welzer and Boštjan Brumen

Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ulica 17, 2000 Maribor, Slovenia

marko.holbl@uni-mb.si

Keywords: authentication, identity-based, key agreement, pairing, security

Received: April 8, 2008

Key agreement protocols are used to exchange keys between two or multiple entities. The exchanged key can be later used to assure confidentiality through encryption. Additionally authenticated key agreement protocols offer implicit authentication. In this paper we conduct a security and efficiency comparison of tripartite authenticated identity-based key agreement protocols and review all of the protocols from the group. From the security perspective the protocols are compared with respect to the level to which they comply with defined security properties for authenticated key agreement protocols and the number of known attacks, whereas from the efficiency perspective the protocols are compared regarding computational effort. The comparative study enables in-depth analysis of existing protocols and the development of new ones.

Povzetek: Podana je primerjava protokolov za izmenjavo ključev.

1 Introduction

In key agreement protocols two or more entities agree upon a session key to be used for assuring a confidentiality or similar cryptographic goals. In 1976, Diffie and Hellman proposed the first key agreement protocol [13]. However, the basic Diffie-Hellman protocol does not authenticate the two communication entities, thus is susceptible to the man-in-the-middle attack. Later, different approaches and protocols have been developed to solve the problem [20, 12].

A research direction in the field of key agreement protocols are key agreement protocols for multi-party settings. A special case of multi-party key agreement protocols are tripartite (or three-party) protocols, which are of special interest as they are applicable to many practical scenarios such as e-commerce (two users and a merchant). Moreover, their implementation is easier and often more efficient than in case of multi-party protocols which are often very complex. The pioneer work by Joux [17] has shown how to implement a tripartite key agreement protocol employing pairings. In the protocol only one broadcast is required for each entity. However, just like the basic Diffie-Hellman protocol, Joux's protocol does not provide authentication of the three communicating entities and thus is vulnerable to the man-in-the-middle attack. To solve the problem with Joux's protocol, Al-Riyami et al. presented several protocols [1] which assure authenticity through use of certificates issued by a Certificate Authority (CA). The session keys are generated by both ephemeral (short-term) keys and static (long-term) keys. The signature of the CA assures that only the entities who possess the static keys are able to compute the session keys. However, in a certificate system, before using the public key of a user, the participants must

first verify the certificates which requires a large amount of computing time and storage. The set of key pairs, certificates and certification authorities is referred to as public key infrastructure (PKI).

As an alternative to certificate-based PKIs, Shamir introduced the concept of an identity-based cryptosystem [24] in which the user's public key is an easily calculated function of her identity (e.g. social security number), while the user's private key is calculated by a trusted authority referred to as Key Generation Center (KGC). Shamir provided the first identity-based key construction based on the RSA problem, and presented an identity-based signature scheme [24]. The identity-based public key cryptosystem simplifies the process of key management, therefore can be an alternative for certificate-based public key infrastructure (PKI). In such cryptosystems, entity A can send encrypted messages to entity B by using her identity information even before B obtains her private key from the KGC. Hence, the idea also provides a way to construct authenticated key agreement protocols.

Recently, bilinear pairings have found positive application in cryptography [3, 6, 17, 29]. They can also be applied for constructing identity-based cryptographic protocols. Many identity-based cryptographic protocols for two and three-party setting have been proposed using the bilinear pairings. Some examples are Boneh-Franklin's identity-based encryption scheme [3], identity-based authentication key agreement protocol by Smart [28], McCullagh-Barreto [18] and several identity-based signatures schemes [29, 23, 9].

In this paper we will conduct a comparative study of identity-based authenticated key agreement protocols using

pairing operations for three-party settings. As far as we are aware, no tripartite identity-based authenticated key agreement protocol without pairings were proposed and that is why the comparative study includes only protocols employing pairing operations. In addition, we review all the protocols. The comparative study of the protocols is conducted as to security and efficiency. Both comparisons will be conducted using defined criteria. Thus the security criteria is defined by the fulfillment of security properties as described in [4, 7] and existence of attacks on the protocols. The efficiency comparison is realized using efficiency criteria; i.e., the number of computational operations required by a protocol. Even if a protocol fulfills all the security properties, its usage is questionable in case attacks for the protocol were published.

The rest of the paper is organized as follows: the next section briefly explains the identity-based public key infrastructure and the corresponding concepts (bilinear maps, the Weil pairing and the associated computational problems). Section 3 gives details on the security properties desired for a sound authenticated key agreement protocol. In Section 4, tripartite identity-based authenticated key agreement protocols using pairing operations are reviewed. For every protocol a description of the phases, security and efficiency discussion are given. A comparative study of the reviewed protocols regarding security and efficiency is conducted in section 5. Finally, a conclusion is made in section 6.

2 Identity-based Public Key Infrastructure employing pairing

In this section, we briefly describe the basic definition and properties of the bilinear pairing, the Weil pairing and the computational problems which form the basis for identity-based public key infrastructure employing pairings.

Traditional PKI (public key infrastructure) is expensive mainly because of the infrastructure needed to manage and authenticate public keys, and the difficulty in managing multiple communities. It is not believed that identity-based public key cryptography would replace the conventional PKIs, but can be rather seen as an alternative solution. In identity-based public key cryptography, one's public key is predetermined by information that uniquely identifies them. The idea of this concept, that was first proposed by Shamir [24], was to simplify certificate management in e-mail systems. When A sends e-mail to B , she encrypts the message using the public key string of B 's e-mail (e.g. Bob@email.com). No public key certificate for B has to be obtained by A . When B receives the encrypted mail she contacts the key generation center (KGC), authenticates herself and thus can obtain the private key from the KGC, which enables her to decrypt the e-mail. In contrast to existing PKI, A is able to send encrypted mail to B even if B has not setup her public key certificate yet. A special case of identity-based public key cryptography (PKC)

can be implemented using bilinear pairings, which will be described next.

2.1 Bilinear Maps

In this section we describe bilinear maps, pairings and their properties. More details can be found in Joux [17] and Boneh-Franklin [3].

Let \mathbb{G}_1 and \mathbb{G}_2 denote two groups of prime order q . \mathbb{G}_1 is an additive group and \mathbb{G}_2 a multiplicative group. Let P be a generator of \mathbb{G}_1 . A pairing is a computable bilinear map between these two groups. Two pairings have been studied for cryptographic use, namely the Weil pairing [19, 27, 31, 3] and the Tate pairing [14, 15, 16].

For our purpose, let \hat{e} denote a general bilinear map $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$, which satisfies the following three properties:

1. *Bilinear*: If $P, Q \in \mathbb{G}_1$ and $a, b \in \mathbb{Z}_q^*$, then $\hat{e}(aP, bQ) = e(P, Q)^{ab}$.
2. *Non-degenerative*: There exist non-trivial points $P, Q \in \mathbb{G}_1$ both of order q such that $\hat{e}(P, Q) \neq 1$.
3. *Computable*: If $P, Q \in \mathbb{G}_1$, $\hat{e}(P, Q) \in \mathbb{G}_2$ is efficiently computable (in polynomial time).

We say that \mathbb{G}_1 is a bilinear group if the group action in \mathbb{G}_1 can be computed efficiently and there exists a group \mathbb{G}_2 and an efficiently computable bilinear map $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ as above. Weil and Tate pairings associated with super singular elliptic curves or Abelian varieties can be modified in order to create such bilinear maps. Concrete examples and details are given in [3], [17], [5].

2.2 The Weil Pairing

Let \mathbb{G}_1 be a subgroup of the group of points on the Elliptic curve E over the finite field \mathbb{F}_q . Let the order of \mathbb{G}_1 be denoted by l , and define k to be the smallest integer such that $l/q^k - 1$. In practical implementations we will require k to be small and so we will usually take E to be a super singular curve over \mathbb{F}_q . The Weil pairing [31, 3] is a map $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ which satisfies the properties given in section 2.1 (bilinearity, non-degeneration and computability).

2.3 Computational Problems

Many pairing-based cryptographic protocols are based on the hardness of the BDHP (Bilinear Diffie-Hellman Problem) for their security [3, 10]. Some computational problems related to the elliptic curve cryptography:

– Bilinear Diffie-Hellman Problem (BDHP)

Let \mathbb{G}_1 and \mathbb{G}_2 be two groups of prime order q . Let $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$, be a bilinear map and let P be a generator of \mathbb{G}_1 . The BDH problem in $\langle \mathbb{G}_1, \mathbb{G}_2, \hat{e} \rangle$ is defined as: Given $(P, xP, yP, zP) \in$

\mathbb{G}_1 for some x, y, z chosen at random from \mathbb{Z}_q^* , compute $\hat{e}(P, P)^{xyz} \in \mathbb{G}_2$.

– **Discrete Logarithm Problem (DLP)**

Given $P, Q \in \mathbb{G}_1$, find an integer n such that $P = nQ$.

– **Computational Diffie-Hellman Problem (CDHP)**

Given a tuple $(P, aP, bP) \in \mathbb{G}_1$ for $a, b \in \mathbb{Z}_q^*$, find the element abP .

3 Security properties

In order to get a sound key agreement protocol, we need to define properties, which are described in detail in [4]. Here we assume A, B and C are three honest entities. It is desired for authenticated key agreement protocols to possess the following security attributes [4, 7, 10, 22]:

- **Implicit Key Authentication.** A key agreement protocol provides *implicit key authentication* if each entity A is assured that no other entity besides entities B and C can determine the value of a particular secret key. A protocol which provides implicit key authentication for entities A, B , and C is called an authenticated key agreement protocol (AK).
- **Known-Key Security.** In each round of a key agreement protocol, A, B and C should generate a unique secret key. Each key generated in one protocol round is independent and should not be exposed if other secret keys are compromised, i.e. the compromise of one session key should not compromise other session keys.
- **Forward Secrecy.** If the long-term private keys of one or more of the entities are compromised, the secrecy of previously established session keys should not be affected. We say that a system has *partial forward secrecy* if some but not all of the entities' long-term keys can be corrupted without compromising previously established session keys, and we say that a system has *perfect forward secrecy* if the long-term keys of all the entities involved may be corrupted without compromising any session key previously established by these entities.
- **Unknown Key-Share resilience.** After the protocol run, entity A believes she shares a key with B and C , whereas B and C mistakenly believe that the key is instead shared with an adversary. Therefore, a sound authenticated key agreement protocol should prevent the unknown key-share situation.
- **Key-Compromise Impersonation.** Assume that A, B and C are three principals. Suppose A 's secret key is disclosed. Obviously, an adversary who knows this secret key can impersonate A to B and C . However,

it is desired that this disclosure does not allow the adversary to impersonate other entities (e.g. B and C) to the real A .

- **Key Control.** The key should be determined jointly by all A, B and C . Neither A, B nor C can control the key alone.

4 Review of tripartite identity-based authenticated key agreement protocols employing pairings

In this section we will review tripartite identity-based authenticated key agreement protocol employing pairings. Some protocols derive multiple keys for later encryption, like [32], [30], whereas others compute just one key [21], [21], [26] and simplified variant of [32]. All protocols consist of three phases, namely the *system setup*, *private key extraction* and *key agreement* phase. Furthermore, each protocol requires three entities (e.g. A, B and C) and a key generation center (KGC) that is relied upon to create and deliver private keys to entities and to not abuse its knowledge of those keys.

When describing the *key agreement* phase we will give only examples of computations performed by entity A . Observe that entities B and C perform almost identical computational operations in the particular key agreement phases of the reviewed protocols, except for the change of indexes in the equations.

All key agreement schemes feature a key derivation function kdf defined as $kdf = \mathbb{F}_q^* \rightarrow \{0, 1\}^*$. The key derivation function is needed in every scheme because the session keys are subsequently used for encrypting data with it usually realized using block ciphers. These require bit strings as keys.

All of the reviewed protocols features the same *system setup* and *private key extraction* phases. Therefore the first two phases will be reviewed here, whereas the *key agreement* phase will be described for every protocol separately.

System Setup. The Key Generation Center (KGC) constructs two groups \mathbb{G}_1 and \mathbb{G}_2 and a map $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$. Next it computes a cryptographic hash function $H : \mathbb{Z}_q^* \rightarrow \mathbb{G}_1$, a generator (primitive root) $P \in \mathbb{G}_1$, a random integer $s \in \mathbb{Z}_q^*$ as KGC's private key and KGC's public key as $P_{KGC} = sP$. All elements are of order q . Finally, the following parameters are published: $\langle \mathbb{G}_1, \mathbb{G}_2, \hat{e}, P, P_{KGC}, H \rangle$ and the master key is s .

Private Key Extraction. For user with identity ID_i the public key is derived as $Q_i = H(ID_i)$ and the private key as $S_i = sQ_i$. Both parameters are computed by the KGC and afterwards S_i is issued to the entity via a secure channel.

4.1 Zhang-Liu-Kim’s Protocol

In 2002, Zhang, Liu and Kim proposed the first tripartite identity-based authenticated key agreement protocol [32]. Each instance of the protocol results in multiple session keys. The way the session key is produced makes use of the Weil pairing and the identity-based static public keys.

Key Agreement. To establish a session key, the three communication entities, A , B and C must proceed as follows.

1. $A \rightarrow B, C: P_A = aP, P'_A = a'P, T_A = H(P_A, P'_A)S_A + aP'_A.$
2. $B \rightarrow A, C: P_B = bP, P'_B = b'P, T_B = H(P_B, P'_B)S_B + bP'_B.$
3. $C \rightarrow A, B: P_C = cP, P'_C = c'P, T_C = H(P_C, P'_C)S_C + cP'_C.$

A verifies: $\hat{e}(T_B + T_C, P) = \hat{e}(H(P_B, P'_B)Q_B + H(P_C, P'_C)Q_C, P_{KGC}) \cdot \hat{e}(P_B, P'_B) \cdot \hat{e}(P_C, P'_C).$

If the above equation holds, then A computes the 8 session keys:

$$K_A^{(1)} = \hat{e}(P_B, P_C)^a, K_A^{(2)} = \hat{e}(P_B, P'_C)^a, K_A^{(3)} = \hat{e}(P'_B, P_C)^a, K_A^{(4)} = \hat{e}(P'_B, P'_C)^a, K_A^{(5)} = \hat{e}(P_B, P_C)^{a'}, K_A^{(6)} = \hat{e}(P_B, P'_C)^{a'}, K_A^{(7)} = \hat{e}(P'_B, P_C)^{a'}, K_A^{(8)} = \hat{e}(P'_B, P'_C)^{a'}.$$

Each entity takes the eight values $K_{ID}^{(i)}, i = 1, 2, \dots, 8$, as the final session keys. The correctness of the protocol can be easily checked by the bilinear property of the pairing:

$$K_A^{(1)} = \hat{e}(P_B, P_C)^a = \hat{e}(abP, cP) = \hat{e}(aP, cP)^b = \hat{e}(P_A, P_C)^b = K_B^{(1)} = \hat{e}(bP, aP)^c = \hat{e}(P_B, P_A)^c = K_C^{(1)}. \text{ Similarly, we get } K^{(i)} = K_A^{(i)} = K_B^{(i)} = K_C^{(i)}, i = 2, 3, \dots, 8.$$

Security and attacks. From the security point of view the protocol has the following security properties: known key security, perfect forward secrecy, key control, key-compromise impersonation and unknown key-share.

Shim and Woo developed an attack on Zhang-Liu-Kim’s protocol [30]. They showed that the protocol is insecure against an unknown key-share (UK-S) which enables the adversary to make entity a believe she shares a key with B and C , whereas B and C mistakenly believe that the key is instead shared with an adversary.

In the UK-S attack scenario B and C compute the same 8 session keys, while A computes his session keys, from which 4 of 8 keys are equal. Thus A, B and C share the first four session keys and A thinks that the session keys are shared with B and C , while B (resp. C) mistakenly believes that she shares the keys with E and C (resp. E and B). Moreover, both A and B come to share the same eight session keys. The weakness of the protocol against the unknown-key share attack is due to the fact that anyone who does not know an ephemeral private key a corresponding to $P_A = aP$ can generate

her own signature on P_A and the lack of explicitness in cryptographic messages, i.e., the signed messages of the protocol do not include some information to confirm that the sender is identical to a genuine communicating entity.

Efficiency. In the protocol each entity uses 4 pairings for verification of the broadcast messages from the other two entities, and 4 pairings to compute the 8 session keys. Additionally, each entity has to compute 6 scalar multiplication and 8 exponentiations. Because there are 8 keys derived, the computational overhead per derived key for each entity is 1 pairing operation, 0,75 scalar multiplication and 1 exponentiations.

4.2 Simplified Zhang-Liu-Kim’s Protocol

In the same paper [32], Zhang-Liu-Kim also published a simplified version of identity-based tripartite authenticated key agreement, i.e., the 3 entities agree to 1 session key instead of 8 keys.

Key Agreement. A, B and C compute and broadcast the following:

1. $A \rightarrow B, C: P_A = aP, T_A = H(P_A)S_A + aP_A.$
2. $B \rightarrow A, C: P_B = bP, T_B = H(P_B)S_B + bP_C.$
3. $B \rightarrow A, B: P_C = cP, T_C = H(P_C)S_C + cP_C.$

A verifies: $\hat{e}(T_B + T_C, P) = \hat{e}(H(P_B)Q_B + H(P_C)Q_C, P_{KGC}) \hat{e}(P_B, P_B) \hat{e}(P_C, P_C).$

If the above equation holds, then A computes: $K_A = \hat{e}(P_B, P_C)^a.$

Then the session key is $K_A = K_B = K_C = \hat{e}(P, P)^{abc}.$

Security and Attacks. The authors claim that their protocol has the following security properties: known key security, perfect forward secrecy, key control, key-compromise impersonation and unknown key-share. No attacks on the protocols are known so far.

Efficiency. With the simplified version of the protocol, an entity needs to compute 5 pairings, 4 for verification and 1 for the generation of the session key.

4.3 Nalla-Reddy’s Protocol

Nalla and Reddy proposed their identity-based tripartite authenticated key agreement protocol employing pairings in 2003 [22]. They employ ideas by Shim two-party identity-based authenticated key agreement protocol employing pairings [25] and Joux’s tripartite identity-based authenticated key agreement protocol [17].

The authors present 3 protocols: ID-AK-1 (Identity-based Authenticated Key Agreement Protocol 1), ID-AK-2 and ID-AK-3, which will be reviewed separately.

Key Agreement

ID-AK-1. Each user generates a random number a, b and

c . The ephemeral (or short term) public keys would be aP , bP and cP , and the ephemeral or short term private keys would be a , b and c .

1. $A \rightarrow B, C: aP$.
2. $B \rightarrow A, C: bP$.
3. $B \rightarrow A, B: cP$.

User A computes $K_A = \hat{e}(bP, cP)^a \cdot \hat{e}(Q_B, P_{KGC}) \cdot \hat{e}(Q_C, P_{KGC}) \cdot \hat{e}(S_A, P) = \hat{e}(P, P)^{abc} \cdot \hat{e}(Q_A, P)^s \cdot \hat{e}(Q_B, P)^s \cdot \hat{e}(Q_C, P)^s$.

The session key is computed as $K_{ABC} = \hat{e}(P, P)^{abc} \cdot \hat{e}(Q_A, P)^s \cdot \hat{e}(Q_B, P)^s \cdot \hat{e}(Q_C, P)^s = \hat{e}(P, P)^{abc} \cdot \hat{e}((Q_A + Q_B + Q_C), P_{KGC})$ and hence depends on the identities of the three entities Q_A , Q_B , Q_C , and the three ephemeral private keys a , b and c .

ID-AK-2. Similarly as in ID-AK-1, each user generates a random number a , b and c . The ephemeral (or short term) public keys would be aP_{KGC} , bP_{KGC} and cP_{KGC} , and the ephemeral or short term private keys would be a , b and c .

1. $A \rightarrow B, C: aP_{KGC}$.
2. $B \rightarrow A, C: bP_{KGC}$.
3. $B \rightarrow A, B: cP_{KGC}$.

User A computes $K_A = \hat{e}(aS_A, P) \cdot \hat{e}(Q_B, bP_{KGC}) \cdot \hat{e}(Q_C, cP_{KGC}) = \hat{e}(aQ_A + bQ_B + cQ_C, sP)$ and user B computes $K_B = \hat{e}(Q_A, aP_{KGC}) \cdot \hat{e}(bQ_B, P) \cdot \hat{e}(Q_C, cP_{KGC}) = \hat{e}(aQ_A + bQ_B + cQ_C, sP)$ Similarly C computes $K_C = \hat{e}(Q_A, aP_{KGC}) \cdot \hat{e}(Q_B, bP) \cdot \hat{e}(cS_C, P_{KGC}) = \hat{e}(aQ_A + bQ_B + cQ_C, sP)$.

Hence the session key is computed as $K_{ABC} = K_A = K_B = K_C = \hat{e}(aQ_A + bQ_B + cQ_C, sP)$.

ID-AK-3. Each user generates random $a, b, c \in \mathbb{Z}_q^*$, which are the ephemeral private keys of A , B and C . The data flows of the protocol are as follows.

1. $A \rightarrow B: aP, aQ_C; A \rightarrow C: aP, aQ_B;$
2. $B \rightarrow A: bP, bQ_C; B \rightarrow C: bP, bQ_A;$
3. $C \rightarrow A: cP, cQ_B; C \rightarrow B: cP, cQ_A;$

A computes the key $K_A = \hat{e}(a(Q_B + Q_C), P_{KGC}) \cdot \hat{e}(S_A, (bP + cP)) \cdot \hat{e}(bQ_C, P_{KGC}) \cdot \hat{e}(cQ_B, P_{KGC})$.

Hence the session key is derived as $K_{ABC} = K_A = K_B = K_C = \hat{e}(a(Q_B + Q_C) + b(Q_A + Q_C) + c(Q_A + Q_B), sP)$.

Security and Attacks. The authors claim different security properties fulfillment for each of the three protocols. *ID-AK-1* complies to forward secrecy, key control and unknown key-share. The *ID-AK-2* protocol conforms to the properties of forward secrecy, key control, key-compromise impersonation and unknown key-share. The third protocol, *ID-AK-3*, fulfills the following security properties: known key security, forward secrecy, key control, key-compromise impersonation and unknown key-share.

However, passive attacks on ID-AK-2 and ID-AK-3 protocols were published by Chen [8] in 2003. In a passive attack, the adversary is able to derive the session keys just eavesdropping on the communication line and use the intercepted data to compute the key.

The passive attack on ID-AK-2 is carried out as follows: since in the ID-AK-2 protocol the key is computed as $K_{ABC} = \hat{e}(aQ_A + bQ_B + cQ_C, sP) = \hat{e}(Q_A, aP_{pub}) \cdot \hat{e}(Q_B, bP_{pub}) \cdot \hat{e}(Q_C, cP_{pub})$ and Q_A, Q_B, Q_C and P_{KGC} are publicly known, a passive attacker can eavesdrop aP_{pub}, bP_{pub} and cP_{pub} , and is able to compute K_{ABC} .

Additionally to the presented attack on ID-AK-2, Chen also demonstrated attack on ID-AK-3 [8]. In the protocol the key is computed as $K_{ABC} = \hat{e}(a(Q_B + Q_C) + b(Q_A + Q_C) + c(Q_A + Q_B), sP)$. Q_A, Q_B, Q_C and P_{KGC} are publicly known, and a passive attacker can know them. In a protocol run, the passive attacker can eavesdrop $aQ_b, aQ_C, bQ_C, bQ_A, cQ_A$, and cQ_B and is able to compute K_{ABC} .

Additionally, Shim published a man-in-the-middle attack on Nalla-Reddy's ID-AK-1 protocol [26]. In the attack the adversary is able to compute and share session keys with all three entities by intercepting the original messages aP, bP, cP and inserting her own messages $a'P, b'P, c'P$. At the end E is can compute K_A, K_B and K_C and therefore shares a key with A, B and C .

Efficiency. Because of the different computational task performed by each of the protocol (ID-AK-1, ID-AK-2 and ID-AK-3), we will discuss efficiency of each of them separately.

In the *ID-AK-1* protocol each user needs to compute 4 Weil pairings and 1 scalar multiplication. However, 3 of the Weil pairings can be precomputed and only 1 pairing needs to be computed for each session. To sum up, each entity has to perform: 5 pairing operation and 5 scalar multiplication.

In the second protocol, *ID-AK-2*, each user is required to compute 2 scalar multiplications and 3 Weil pairings.

The last of the three presented protocols (*ID-AK-3*) is role symmetric since each participant executes the same number of operations. It requires each participant to compute 2 additions, 4 scalar multiplications, and 4 Weil pairings.

4.4 Nalla's Protocol with Signatures

Nalla, proposed another tripartite key agreement protocol for identity-based systems employing identity-based signatures in 2003 [21]. Because some identity-based tripartite key agreement protocols proposed in Nalla-Reddy's previous work [22] suffered passive attacks, and Joux's protocol [17] suffered man-in-the-middle attack, Nalla proposed a new protocol including signature in Joux's protocol. It resulted in much simpler identity-based key agreement protocols.

Key Agreement. Let A, B and C be the three parties wishing to compute a session key. First, A, B and C select

random number a, b and $c \in \mathbb{Z}_q^*$ and perform the following actions:

1. $A \rightarrow B, C: P_A = aP, T_A = a^{-1}(H(P_A)S_A)$
2. $B \rightarrow A, C: P_B = bP, T_B = b^{-1}(H(P_B)S_B)$
3. $C \rightarrow A, B: P_C = cP, T_C = c^{-1}(H(P_C)S_C)$

A verifies: $\hat{e}(P_B, T_B) \cdot \hat{e}(P_C, T_C) = \hat{e}(P_{KGC}, H(P_B)Q_B + H(P_C)Q_C)$ and computes $K_A = \hat{e}(P_B, P_C)^a = \hat{e}(P, P)^{abc}$.

This verification ensures the authenticity of the senders. The session key is the value $K_{ABC} = K_A = K_B = K_C = \hat{e}(P, P)^{abc}$.

Security and Attacks. The author claims that his protocol has the following security properties: known key security, perfect forward secrecy, key control, key-compromise impersonation and unknown key-share.

In 2003, Shim published an impersonation attack on the Nalla’s protocol with signatures [26]. According to [26], the adversary is able to broadcast such messages with help of which she can impersonate an entity (in the paper an example for entity A is given). The messages sent by the adversary E impersonating A are successfully verified by B and C . Additionally, E can compute the session key K_A and finally succeed to impersonate A to B and C and compute the session key.

Shim claims [26] that Nalla’s protocol is insecure against the man-in-the-middle attack because of the impersonation attack. She further claims that the weakness of the protocol against the attack is due to the fact that anyone who does not know each other’s private key (S_{ID}) can generate a valid pair (P_{ID}, T_{ID}) .

Efficiency. Regarding efficiency, in each protocol run the following operations have to be computed: 4 pairing operation, 5 scalar multiplication and 1 exponentiations.

4.5 Shim’s Protocol with Signatures

Due to the flaws in Nalla-Reddy’s and Nalla’s protocols, Shim proposed a modified identity-based tripartite key agreement protocol with signatures [26].

Key Agreement. Let A, B and C be the three parties wishing to compute a session key. A, B and C select random number a, b and $c \in \mathbb{Z}_q^*$ and exchange the following messages:

1. $A \rightarrow B, C: P_A = aP, T_A = H(P_A)S_A + aP_{KGC}$
2. $B \rightarrow A, C: P_B = bP, T_B = H(P_B)S_B + bP_{KGC}$
3. $C \rightarrow A, B: P_C = cP, T_C = H(P_C)S_C + cP_{KGC}$

A verifies: $\hat{e}(T_B + T_C, P) = \hat{e}(P_{KGC}, H(P_B)Q_B + H(P_C)Q_C + P_B + P_C)$.

If the equation holds, then A computes $K_A = \hat{e}(P_B, P_C)^a = \hat{e}(P, P)^{abc}$.

This verification ensures the authenticity of the senders. The session key is the value $K_{ABC} = K_A = K_B =$

$$K_C = \hat{e}(P, P)^{abc}.$$

Security and Attacks. From the security perspective the protocol features known key security, perfect forward secrecy, key control, key-compromise impersonation and unknown key-share. No attacks on the protocols are known so far.

Efficiency. In the reviewed protocol the computation effort includes 3 pairing operations, 5 scalar multiplications and 1 exponentiation.

4.6 Shim-Woo’s Protocol

Recently, Shim and Woo proposed a more efficient identity-based tripartite multiple-key agreement protocol which satisfies all the required security attributes and does not use any one-way hash functions.

Key Agreement. Suppose three communication entities, A, B and C want to establish a secret session key. To achieve this, they perform:

1. $A \rightarrow B, C: P_A = aP, P'_A = a'P, T_A = S_A + a^2P + a'P_{KGC}$.
2. $B \rightarrow A, C: P_B = bP, P'_B = b'P, T_B = S_B + b^2P + b'P_{KGC}$.
3. $B \rightarrow A, B: P_C = cP, P'_C = c'P, T_C = S_C + c^2P + c'P_{KGC}$.

A verifies $\hat{e}(T_B + T_C, P) = \hat{e}(Q_B + Q_C + P'_B + P'_C, P_{KGC}) \cdot \hat{e}(P_B, P_B) \cdot \hat{e}(P_C, P_C)$.

If the above equation holds, then A computes the 8 session keys:

$$K_A^{(1)} = \hat{e}(P_B, P_C)^a, K_A^{(2)} = \hat{e}(P_B, P'_C)^a, K_A^{(3)} = \hat{e}(P'_B, P_C)^a, K_A^{(4)} = \hat{e}(P'_B, P'_C)^a, K_A^{(5)} = \hat{e}(P_B, P_C)^{a'}, K_A^{(6)} = \hat{e}(P_B, P'_C)^{a'}, K_A^{(7)} = \hat{e}(P'_B, P_C)^{a'}, K_A^{(8)} = \hat{e}(P'_B, P'_C)^{a'}.$$

Each entity takes the eight values $K_{ID}^{(i)}, i = 1, 2, \dots, 8$, as the final session keys. The correctness of the protocol can be easily checked by the bilinear property of the pairing:

$$K_A^{(1)} = \hat{e}(P_B, P_C)^a = \hat{e}(abP, cP) = \hat{e}(aP, cP)^b = \hat{e}(P_A, P_C)^b = K_B^{(1)} = \hat{e}(bP, aP)^c = \hat{e}(P_B, P_A)^c = K_C^{(1)}. \text{ Similarly, we get } K^{(i)} = K_A^{(i)} = K_B^{(i)} = K_C^{(i)}, i = 2, 3, \dots, 8.$$

Security and Attacks. From the security point of view, the protocol features known key security, perfect forward secrecy, key control, key-compromise impersonation and unknown key-share.

However, Chou-Lin-Shiu published an impersonation attack on Shim-Woo’s protocol [11] in 2005. As a result, the adversary can share the 4 keys $K^{(1)}, K^{(2)}, K^{(5)}, K^{(6)}$ of the 8 session keys. Under this situation, two of the three entities (e.g. A and C) involved in the protocol, think

these 4 session keys are shared the third entity (e.g. *B*), but indeed, they are shared with the adversary. Besides, both *A* and *C* come to share the same 8 session keys. Thus, the impersonation attack on 4 of the 8 session keys can be successfully mounted. More precisely, the adversary can use the 4 session keys to communicate with *A* and *C*, and he can have one half of the probability to realize what the communication contents are between *A* and *C*.

Efficiency. According to the authors, their protocol requires the following computational operations: 8 pairing operation, 4 scalar multiplication, 8 exponentiations. Additionally, since 8 key are derived, the effort per key for each entity is 1 pairing operation, 0,5 scalar multiplication, 1 exponentiations.

5 Comparative study

In the following section we compare the reviewed protocols with respect to security and performance. From the security point of view the criteria to compare security of the protocols is given by the extent to which a specific protocol fulfills the security properties as discusses in section 3. Additionally, attacks for each protocol are analyzed and included in the criteria for comparing security. From the performance point of view the criteria for comparing efficiency is defined as the number of computational operations required per protocol run.

In general authenticated key agreement protocols have to be secure and at the same time as efficient as possible. Therefore the security factor is more important when assessing and comparing the reviewed protocols.

5.1 Security Comparison

The security comparison of the reviewed protocols is conducted as to two criteria: the fulfillment of security properties as defined in section 3 and the existence of attacks due to errors in the design of the protocols. Often, an attack on a protocol results in the not-fulfillment of specific security properties, but please observe that this is not always the case, since ID-AK-1 and ID-AK-2 are susceptible to passive attack and yet this does not violate any security property.

Further please notice that Nalla-Reddy’s ID-AK-1, ID-AK-2 and ID-AK-3 were broken due to various attacks (please refer to section 4), but are included in the comparison for completeness.

5.1.1 Security Properties

Table 1 summarizes the fulfillment of security properties for each of the reviewed protocols. For definition and details regarding the security properties the reader is referred to section 3.

The majority of protocols do not fulfill the security properties, with exception simplified ZLK and Shim’s proto-

Table 1: Security properties

Protocol	KKS	FS	KCI	UKS	KC
ZLK	+	+*	+	− ^a	+
simplified ZLK	+	+*	+	+	+
ID-AK-1	-	+	-	+	+
ID-AK-2	-	+	+	+	+
ID-AK-3	+	+	+	+	+
Nalla	+	+	− ^b	+	+
Shim	+	+	+	+	+
Shim-Woo	+	+*	− ^c	+	+

KKS - Known-Key Secrecy

FS - Forward Secrecy

* - perfect forward secrecy

KCI - Key-Compromise Impersonation

UKS - Unknown Key-Share

KC - Key Control

a - as to Shim-Woo’s unknown key-share attack [30]

b - Shim’s Impersonation attack [26]

c - Chou-Lin-Chiu’s Impersonation attack [11]

col. Additionally, ZLK’s, the simplified ZLK and Shim-Woo’s protocols offer perfect forward secrecy, whereas the rest has the property of partial forward secrecy. Nalla’s and Shim-Woo’s protocols do not fulfill particular security properties because of attacks (please refer to section 4 for further details).

5.1.2 Known Attacks

Some of the reviewed protocols have been shown to have weaknesses, which were exploited for attacks. Table 2 sums up attack for the reviewed protocols. For two of the reviewed protocols (i.e., ZLK’s and Shim’s protocol) there are no known attacks. When a protocol suffers from attacks

Table 2: Known attacks

Protocol	Attacks
ZLK	Shim-Woo’s unknown key-share attack [30]
simplified ZLK	\
ID-AK-1	Shim’s Man-in-the-middle attack [26]
ID-AK-2	Chen’s Passive attack [8]
ID-AK-3	Chen’s Passive attack [8]
Nalla	Shim’s Impersonation attack [26]
Shim	\
Shim-Woo	Chou-Lin-Chiu’s Impersonation attack [11]

it lacks security and as a consequence sometimes does not

fulfill a defined security property (see section 3).

5.2 Efficiency Comparison

The computations effort per user (number of computations performed) of the reviewed protocols is given in table 3. We compare operations which are expensive from the computational point of view - pairing operations, scalar scalar multiplications and exponentiation. Additions and hash operations are ignored since they are much less computationally expensive. As precomputing pairing operations increases the performance and lowers the computational effort, we also give data regarding precomputation. A pre-computed pairing operations must only be carried out when the three entities conduct a key agreement for the first time and can be later omitted.

Table 3: Computation effort per user.

Protocol	PairOp	ScMul	Exp	PP
ZLK	8	6	8	0
simplified ZLK	5	5	1	0
ID-AK-1	4	1	0	3
ID-AK-2	3	2	0	0
ID-AK-3	4	4	0	0
Nalla	4	5	1	0
Shim	3	5	1	0
Shim-Woo	8	4	8	0

PairOp - pairing operations

ScMul - scalar multiplications in \mathbb{G}_1

Exp - exponentiation in \mathbb{G}_2

PP - pairings that can be pre-computed

Before discussing the efficiency of the reviewed protocols, it should be noted that according to [2], the effort to evaluate one pairing operation is approximately equal to the effort of computing three scalar multiplications. As can be seen from table 3 the most efficient protocol is Nalla-Reddy's ID-AK-2, followed by Shim's protocol. In contrary, the least efficient protocol is the ZLK protocol resp. the simplified ZLK protocol.

The most robust and most efficient protocol from both the security and efficiency point of view is not straightforward. Due to the discussed attacks, we have to rule out all the protocols susceptible to attacks and not fulfilling all security requirements. This leaves us with only two protocols, namely the simplified ZLK protocol and Shim's protocol. When taking the results of the efficiency analysis into account Shim's protocols prevails as it offers best performance while fulfilling the desired security properties and immune to attacks.

6 Conclusion

Identity-based authenticated key agreement protocols can be an alternative for certificate-based protocols. This is true, especially when efficient key management and moderate security are required. In the paper we have made a review and comparative study of tripartite authenticated identity-based key agreement protocols using pairings. We have presented the state of the art in attacks on the reviewed protocols and conducted a comparative study regarding the fulfillment of security properties, attacks published and the computational effort required by each protocol. The prevailing protocol considering security and efficiency is Shim's protocol as it is efficient and at the same time offers all security properties. Future development of protocols must take the analysis results and attacks into account when developing new protocols.

References

- [1] S. Al-Riyami, and K. Paterson (2002) Authenticated three party key agreement protocols from pairings, *Cryptology ePrint Archive, Report 2002/035*.
- [2] P.S.L.M. Barreto, H.Y. Kim, B. Lynn, M. Scott (2002) Efficient algorithms for pairing-based cryptosystems, *Advances in Cryptology - Crypto'02, LNCS Vol. 2139*, Springer, UK, pp. 213 - 229 .
- [3] D. Boneh, and M. Franklin (2003) Identity-based encryption from the Weil pairing, *Advances in Cryptology - Crypto'01, LNCS Vol. 2442*, Springer, UK, pp. 354-368.
- [4] S. Blake-Wilson, D. Johnson, and A. Menezes (1992) Key agreement protocols and their security analysis (Extended abstract), *6th IMA International Conference on Cryptography and Coding, LNCS Vol. 1355*, Springer, UK, pp. 30-45.
- [5] D. Boneh, B. Lynn, and H. Shacham (2002) Short signatures from the Weil pairing, *Advances in Cryptology - Asiacrypt'01, LNCS 2248*, Springer, UK, pp. 514-532.
- [6] D. Boneh, B. Lynn, and H. Shacham (2001) Short signatures from the Weil pairing, *Advances in Cryptology-Asiacrypt 2001, LNCS 2248*, Springer, UK, pp. 514-532.
- [7] Bellare, M., Rogaway, P. (1993) Entity Authentication and Key Distribution, *Advances in Cryptology - CRYPTO '93*, Springer, UK, pp. 232-249.
- [8] Chen, Z. (2003) Security analysis on Nalla-Reddy's ID-based tripartite authenticated key agreement protocols, *Cryptology ePrint Archive, Report 2003/103*.

- [9] J.C. Cha, and J.H. Cheon (2002) An identity-based signature from gap Diffie-Hellman groups, *Cryptology ePrint Archive, Report 2002/018*.
- [10] L. Chen, C. Kudla (2003) Identity Based Authenticated Key Agreement Protocols from Pairings, *16th IEEE Computer Security Foundations Workshop*, IEEE Press, USA, pp. 219-233.
- [11] J.S. Chou, C.H. Lin, C.H. Chiu (2005) Weakness of Shim's New ID-based tripartite multiple-key agreement protocol, *Cryptology ePrint Archive, Report 2005/457*.
- [12] R. Dutta, R. Barua (2005) Overview of Key Agreement Protocols, *Cryptology ePrint Archive, Report 2005/289*.
- [13] W. Diffie, M. Hellman (1976) New directions in cryptography, *IEEE Transactions on Information Theory*, vol.22, no.6, IEEE Press, USA, pp. 644-654.
- [14] G. Frey, M. Mller and H. Rock (1999) The Tate pairing and the discrete logarithm applied to elliptic curve cryptosystems, *IEEE Transactions on Information Theory*, vol. 45, no.5, IEEE Press, USA, pp. 1717-1719.
- [15] S. Galbraith (2001) Supersingular curves in cryptography, *Advances in Cryptology - Asiacrypt' 01, LNCS Vol. 2248*, Springer, UK, pp. 495-513.
- [16] S.D. Galbraith, K. Harrison, and D. Soldera (2002) Implementing the Tate Pairing, *5th International Symposium on Algorithmic Number Theory, LNCS Vol. 2369*, Springer, UK, pp. 324-337.
- [17] A. Joux (2000) A one round protocol for tripartite Diffie-Hellman, *4th International Symposium on Algorithmic Number Theory, LNCS Vol. 1838*, Springer, UK, pp. 385-393.
- [18] N. McCullagh, and P.S.L.M. Barreto (2005) A new two-party identity-based authenticated key agreement, *Topics in Cryptology - CT-RSA 2005, LNCS Vol. 3376*, Springer, UK, pp. 262-274.
- [19] A. Menezes, T. Okamoto, and S. Vanstone (1993) Reducing elliptic curve logarithms to logarithms in a finite field, *IEEE Transactions on Information Theory* Vol. 39, IEEE Press, USA, pp. 1639-1646.
- [20] A. Menezes, P.C. Van Oorschot, and S. Vanstone (1997) *Handbook of Applied Cryptography*, CRC Press, USA.
- [21] Nalla, D. (2003) ID-based tripartite key agreement with signatures, *Cryptology ePrint Archive, Report 2003/144*.
- [22] D. Nalla, K.C. Reddy (2003) ID-based tripartite key agreement with signatures, *Cryptology ePrint Archive, Report 2003/004*.
- [23] K.G. Paterson (2002) ID-based signatures from pairings on elliptic curves, *Cryptology ePrint Archive, Report 2002/004*.
- [24] A. Shamir (1985) Identity-Based Cryptosystems and Signature Schemes, *Advances in Cryptology - CRYPTO 84*, pp. 47-53 .
- [25] K. Shim (2003) Efficient ID-based authenticated key agreement protocol based on Weil pairing, *Electronics Letters*, Vol. 39, No. 8, IEEE Press, USA, pp. 653-654.
- [26] K. Shim (2003) Cryptanalysis of ID-based Tripartite Authenticated Key Agreement Protocols, *Cryptology ePrint Archive, Report 2003/115*.
- [27] J.H. Silverman (1994) Advanced topics in the arithmetic of elliptic curves, *Graduate Texts in Mathematics Vol. 151*, Springer, UK.
- [28] N.P. Smart (2002) Identity-based authenticated key agreement protocol based on Weil pairing, *Electronics Letters*, Vol. 38, No. 13, IEEE Press, USA, pp. 630-632.
- [29] R. Sakai, K. Ohgishi, M. Kasahara (2000) Cryptosystems based on pairing, *Symposium on Cryptography and Information Security (SCIS2000)*, Japan.
- [30] K. Shim, S. Woo (2005) Weakness in ID-based one round authenticated tripartite multiple-key agreement protocol with pairings, *Applied Mathematics and Computation*, Vol. 166, No. 3, Elsevier, USA, pp. 523-530.
- [31] E. Verheul (2001) Evidence that XTR is more secure than supersingular elliptic curve systems, *Advances in Cryptology - EUROCRYPT 2001, LNCS Vol. 2045*, Springer, UK, pp. 195-210.
- [32] F. Zhang, S. Liu, K. Kim (2002) ID-based one-round authenticated tripartite key agreement protocol with pairings, *Cryptology ePrint Archive, Report 2002/122*.

Semantic Graphs Derived from Triplets with Application in Document Summarization

Delia Rusu

Technical University of Cluj-Napoca, Faculty of Automation and Computer Science,
G. Barişiu 26-28, 400027 Cluj-Napoca, Romania
E-mail: delia.rusu@gmail.com

Blaž Fortuna, Marko Grobelnik and Dunja Mladenić

Jozef Stefan Institute,
Jamova 39, 1000 Ljubljana, Slovenia
E-mail: {blaz.fortuna,marko.grobelnik, dunja.mladenic}@ijs.si

Keywords: natural language processing, text mining, semantic graph, document summarization

Received: November 12, 2008

Information nowadays has become more and more accessible, so much as to give birth to an information overload issue. Yet important decisions have to be made, depending on the available information. As it is impossible to read all the relevant content that helps one stay informed, a possible solution would be condensing data and obtaining the kernel of a text by automatically summarizing it. We present an approach to analyzing text and retrieving valuable information in the form of a semantic graph based on subject-verb-object triplets extracted from sentences. Once triplets have been generated, we apply several techniques in order to obtain the semantic graph of the document: co-reference and anaphora resolution of named entities and semantic normalization of triplets. Finally, we describe the automatic document summarization process starting from the semantic representation of the text. The experimental evaluation carried out step by step on several Reuters newswire articles shows a comparable performance of the proposed approach with other existing methodologies. For the assessment of the document summaries we utilize an automatic summarization evaluation package, so as to show a ranking of various summarizers.

Povzetek: V članku predlagamo pristop k analizi besedila in zajemanju pomembnih informacij v obliki semantičnega grafa, ki je zasnovan na predstavitvi stavka s trojkami (osebek-povedek-predmet).

1 Introduction

The accessibility of information arises mostly from the rapid development of the World Wide Web and online information services. One has to read a considerable amount of relevant content in order to stay updated, but it is impossible to read everything related to a certain topic. A feasible solution to this admitted problem is condensing this vast amount of data and extracting only the essence of the message, in the form of an *automatically generated summary*.

In this paper we describe a method of text analysis with the stated purpose of extracting valuable information from documents. We shall attach a graphical representation, called semantic graph, to the initial document. The graph is based on triplets retrieved from the document sentences. Moreover, we are going to describe an application of semantic graphs generation–text summarization – as a method for reducing the quantity of information but preserving one important characteristic – its quality.

The paper is organized as follows. Firstly, the triplet based semantic graphs generation algorithm is presented. Two steps are detailed in this phase: triplet extraction

from sentences, followed by the procedure of yielding the semantic graph of the document. In order to obtain the graph, named entity co-reference and anaphora resolution as well the semantic normalization of triplets are employed. Secondly, the summarization process is explained, followed by an evaluation of the system components. The paper concludes with several remarks.

2 Triplet based semantic graphs

In English, the declarative sentence has the basic form *subject – verb – object*. Starting from this observation, one can think of the “core” of a sentence as a *triplet* (consisting of the aforementioned three elements). We assume that it contains enough information to describe the *message of a sentence*. The usefulness of triplets resides in the fact that it is much easier to process them instead of dealing with very complex sentences as a whole.

For triplet extraction, we apply the algorithm for obtaining triplets from a treebank parser output described in [1], and employ the *Stanford Parser* [2].

The extraction is performed based on pure syntactic analysis of sentences. For obtaining semantic information, we first annotate the document with *named entities*. Throughout this paper, the term “*named entities*” refers to names of people, locations and organizations. For named entity extraction we consider *GATE (General Architecture for Text Engineering)* [3], which was used as a toolkit for natural language processing.

The *semantic graph* corresponds to a visual representation of a document’s semantic structure. The starting point for deriving semantic graphs was [4].

The procedure of semantic graph generation consists of a series of sequential operations composing a pipeline:

- *Co-reference* resolution by employing text analysis and matching methods, thus consolidating named entities.
- *Pronominal anaphora* resolution based on named entities.
- *Semantic normalization* using WordNet synsets.
- *Semantic graph generation* by merging triplet elements with respect to the synset they belong to.

The following sub-sections will further detail these pipeline components.

2.1 Co-reference Resolution

Co-reference is defined as the identification of surface terms (words within the document) that refer to the same entity [4]. For simplification, we are going to consider co-reference resolution for the named entities only. The set of operations we have to perform is threefold. Firstly we have to determine the named entity *gender*, so as to reduce the search space for candidates. Secondly, in the case of named entities composed of more than one word, we eliminate the set of English stop words (for example Ms., Inc., and so on). Thirdly, we apply the heuristics proposed in [4]: two different surface forms represent the same named entity if one surface form is completely included in the other. For example, “*Clarence*”, “*Clarence Thomas*” and “*Mr. Thomas*” refer to the same named entity, that is, “*Clarence Thomas*”. Moreover, abbreviations are also co-referenced, for example “*U.S.*”, “*U.S.A.*”, “*United States*” and “*United States of America*” all refer to the same named entity – “*United States America*” (“*of*” will be eliminated, as it is a stop word).

2.2 Anaphora Resolution

In linguistics, *anaphora* defines an instance of an expression that refers to another expression; pronouns are often regarded as anaphors. The pronoun subset we considered for anaphora resolution is formed of: {*I, he, she, it, they*}, and their objective, reflexive and possessive forms, as well as the relative pronoun *who*.

We perform a sequential search, first backward and then forward, with the purpose of finding good replacement candidates for a given pronoun, among the named entities. Firstly, we search backwards inside the

sentence where we found the pronoun. We select candidates that agree in gender with the pronominal anaphor, as suggested in [5, 6]. Next, we look for possible candidates in the sentences preceding the one where the pronoun is located. If we have found no candidates so far, we search forward within the pronoun sentence, and then forward in the next sentences, as in [4]. Once the candidates have been selected, we apply antecedent indicators to each of them, and assign scores (0, 1, and 2). The antecedent indicators we have taken into account are a subset of the ones mentioned in [5]: *givenness, lexical reiteration, referential distance, indicating verbs* and *collocation pattern preference*. After assigning scores to the candidates found, we select the candidate with the highest overall score as the best replacement for the pronoun. If two candidates have the same overall score, we prefer the one with a higher collocation pattern score. If we cannot make a decision based on this score, we choose the candidate with a greater indicating verbs score. In case of a tie, we select the most recent candidate (the one closest to the pronoun).

We summarize the anaphora resolution procedure in the algorithm in Figure 2.1.

```

function ANAPHORA-RESOLUTION (pronoun,
number_of_sentences) returns a solution, or failure
  candidates ←
  BACKWARD-SEARCH-INSIDE-SENTENCE
(pronoun) ∪ BACKWARD-SEARCH (pronoun,
number_of_sentences)
  if candidates ≠ ∅ then
  APPLY-ANTECEDENT-INDICATORS (candidates)
  else
  candidates ← FORWARD-SEARCH-INSIDE-
SENTENCE (pronoun) ∪ FORWARD-SEARCH
(pronoun, number_of_sentences)
  if candidates ≠ ∅ then
  APPLY-ANTECEDENT-INDICATORS (candidates)
  result ← MAX-SCORE-CANDIDATE (candidates)
  if result ≠ failure then return result
  else return failure

function APPLY-ANTECEDENT-INDICATORS (candidates)
returns a solution, or failure
  result ← APPLY-GIVENNESS (candidates) ∪
APPLY-LEXICAL-REITERATION (candidates) ∪
APPLY-REFERENTIAL-DISTANCE (candidates) ∪
APPLY-INDICATING-VERBS (candidates) ∪
APPLY-COLLOCATION-PATTERN-PREFERENCE
(candidates)
  if result ≠ failure then return result
  else return failure

```

Figure 2.1: The anaphora resolution algorithm.

2.3 Semantic Normalization

Once co-reference and anaphora resolution have been performed, the next step is semantic normalization. We compact the triplets obtained so far, in order to generate a more coherent semantic graphical representation. For this task, we rely on the synonymy relationships between words. More precisely, we attach to each triplet element the synsets found with WordNet. If the triplet element is composed of two or more words, then for each of these words we determine the corresponding synsets. This

feature): object, subject, verb (all of these are words), location of the sentence in the document, similarity with the centroid, number of locations in the sentence, number of named entities in the sentence, authority weight for the object, hub weight for the subject, size of the weakly connected component for the object.

The summarization process, described in Figure 3.1, starts with the original document and its semantic graph. The three types of features abovementioned are then retrieved. Further, the sentences are classified with the linear SVM and the document summary is obtained. Its sentences are labelled with SVM scores and ordered based on these scores in a decreasing manner. The motivation for doing this is presented in the next section of the paper.

4 System evaluation

The experiments that were carried out involve gender information retrieval, co-reference and anaphora resolution and finally summarization. In the following, each of these experiments are presented, highlighting the data set used, the systems selected for result comparison and the outcome.

4.1 Gender Information Retrieval

Gender related information was extracted from two GATE resource files: *person_male* and *person_female* gazetteers. For evaluation we manually annotated 15 random documents taken from the Reuters RCV1 [9] data set. The two systems that were compared with the manually obtained results are:

- Our system, henceforward referred to as *System*
- A *Baseline* system, which assigns the *masculine* gender to all named entities labelled as persons.

The results are presented in Table 4.1.

	Masculine	Feminine	Total
System	170/206 (83%)	7/14 (50%)	177/220 (80%)
Baseline	206/206 (100%)	0/14 (0%)	206/220 (94%)

Table 4.1: Gender evaluation results.

The fact that *System* correctly labelled a significant percent of masculine as well as feminine persons shows it will carry out gender retrieval better than the baseline system when the number of persons belonging to either genders will be more balanced.

4.2 Co-reference Resolution

For the evaluation of co-reference resolution the same set of 15 articles mentioned in section 4.1 was used. Named entities were extracted based on GATE, and the co-reference resolution performed by *System* was compared with the one of GATE. The results are shown in Table 4.2. There are 783 named entities extracted using GATE. The *System* performance is better than that of GATE, 750 entities compared to GATE's 646 entities co-referenced.

	Co-References
System	750/783 (96%)
GATE	646/783 (83%)

Table 4.2: Co-reference evaluation results.

4.3 Anaphora Resolution

In the case of anaphora resolution, the *System* was compared with two baseline systems. Both of them consider the closest named entity as a pronoun replacement, but one takes gender information into account, whereas the other does not. In this case, we randomly chose a subset of the Reuters data set formed of 77 articles.

Pronouns	System	Baseline-gender	Baseline-no gender
He	35/42 (83%)	18/42 (43%)	18/42 (43%)
They	7/20 (35%)	8/20 (40%)	2/20 (10%)
I	4/15 (27%)	0/15 (0%)	2/15 (13%)
She	0/0	0/0	0/0
Who	0/0	0/0	0/0
It	11/35 (31%)	11/35 (31%)	11/35 (31%)
Other	2/4 (50%)	2/6 (33%)	3/6 (50%)
Total	59/116 (51%)	39/118 (33%)	36/118 (31%)

Table 4.3: Anaphora evaluation results.

The results are listed in Table 4.3, pointing out the *System* strength where the “*he*” pronoun is concerned.

4.4 Summary Generation

For summarization evaluation, two tests were carried out. The first one involved the usage of the DUC (*Document Understanding Conferences*) [10] 2002 data set, for which the results obtained were similar with the ones listed in [4]. For the second one the DUC 2007 update task data set was used for testing purposes. The data consisted of 10 topics (A-J), each divided in 3 clusters (A-C), each cluster with 7-10 articles. For this assessment, we focused on the first part of the task – *producing a summary of documents in cluster A* – 100-words in length, without taking into consideration the topic information. In order to obtain the 100-word summary, we first retrieved all sentences having triplets belonging to instances with the class attribute value equal to +1, and ordered them in an increasing manner, based on the value returned by the SVM classifier. Out of these sentences, we considered the top 15%, and used them to generate a summary. That is because most sentences that were manually labelled as belonging to the summary were among the first 15% top sentences.

In order to compare the performance of various systems, we employed ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [11], an automatic summarization evaluation package. Our system was ranked 17 out of 25, based on the ROUGE-2 evaluation method, and 18 out of 25 based on the ROUGE-SU4 evaluation method (Figure 4.1 and Figure 4.2).

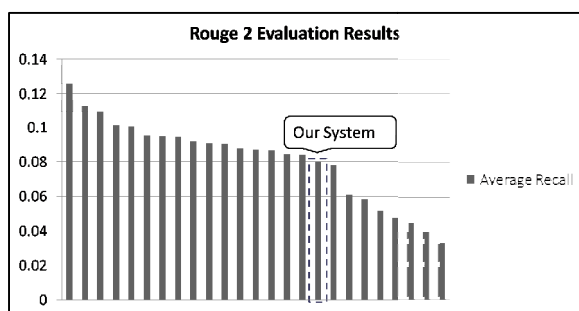


Figure 4.1: ROUGE-2 average recall results for 25 systems.

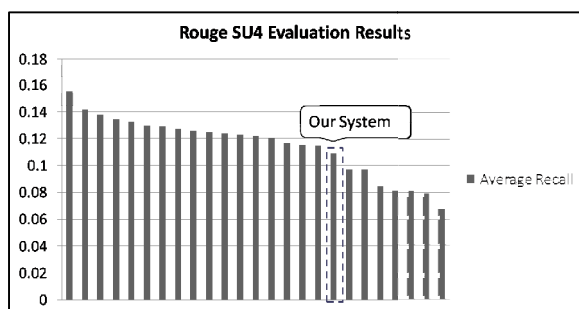


Figure 4.2: ROUGE-SU4 average recall results for 25 systems.

5 Conclusion

The stated purpose of the paper was to present a methodology for generating semantic graphs derived from logical form triplets and, furthermore, to use these semantic graphs to construct document summaries. The evaluation that was carried out showed the system in comparison to other similar applications, demonstrating its feasibility as a semantic graph generator and document summarizer.

As far as future improvements are concerned, one possibility would be to combine the document summarizer with an online newswire crawling system that processes news on the fly, as they are posted, and then uses the summarizer to obtain a compressed version of the initial story.

6 References

- [1] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, D. Mladenić. Triplet Extraction from Sentences. Ljubljana: 2007. Proceedings of the 10th International Multiconference "Information Society - IS 2007". Vol. A, pp. 218 - 222.
- [2] Stanford Parser: <http://nlp.stanford.edu/software/lex-parser.shtml>
- [3] GATE (General Architecture for Text Engineering): <http://gate.ac.uk/>
- [4] J. Lescovec, M. Grobelnik, N. Milic-Frayling. Learning Sub-structures of Document Semantic Graphs for Document Summarization. Seattle: 2004. KDD 2004 Workshop on Link Analysis and Group Detection (LinkKDD).
- [5] R. Mitkov. Robust pronoun resolution with limited knowledge. Montreal: 1998. Proceedings of the 18th International Conference on Computational Linguistics COLING'98/ACL'98. pp. 869-875.
- [6] R. Mitkov. Anaphora Resolution: The State of the Art. Wolverhampton: 1999. Working Paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution).
- [7] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. 1998. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms.
- [8] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. 1998.
- [9] D. D. Lewis, Y. Yang, T. G. Rose, F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. 2004, Journal of Machine Learning Research, Vol. 5.
- [10] DUC (Document Understanding Conferences): <http://duc.nist.gov/>
- [11] ROUGE (Recall-Oriented Understudy for Gisting Evaluation): <http://haydn.isi.edu/ROUGE/>

Improving HTML Compression

Przemysław Skibiński
 University of Wrocław, Institute of Computer Science,
 Joliot-Curie 15, 50-383 Wrocław, Poland
 E-mail: inikep@ii.uni.wroc.pl

Keywords: hypertext markup language, HTML compression, HTML transform

Received: March 6, 2008

The verbosity of the Hypertext Markup Language (HTML) remains one of its main weaknesses. This problem can be solved with the aid of HTML specialized compression algorithms. In this work, we describe a lossless HTML transform which, combined with generally used compression algorithms, allows to attain high compression ratios. Its core is a fully reversible transform featuring substitution of words in an HTML document using a static English dictionary or a semi-static dictionary of the most frequent words in the document, effective encoding of dictionary indexes, numbers, and specific patterns.

The experimental results show that the proposed transform improves the HTML compression efficiency of general purpose compressors on average by 15% in case of gzip achieving comparable processing speed. Moreover, we show that the compression ratio of gzip can be improved by up to 28% for the price of higher memory requirements and much slower processing.

Povzetek: Opisan je izvirni algoritem za izboljšavo zgoščevanja HTML.

1 Introduction¹

Since the origin of World Wide Web, the Hypertext Markup Language (HTML) is a standard for Internet web pages. HTML has many advantages. One of its main advantages is that it is a textual format, what means that HTML is human-readable and can be edited by any text editor. The textual format of HTML is also one of its main disadvantages as it introduces verbosity. Nevertheless verbosity can be coped with by applying data compression.

Currently HTML files are usually compressed with common LZ77-based compression algorithms [23] like gzip or deflate [7]. LZ77-based algorithms can be substituted with more powerful, but slower and much more memory-demanding BWT-based [3] and PPM-based [5] algorithms. All these algorithms, however, are general-purpose and much better results can be achieved with a compression algorithm specialized for dealing with HTML documents.

In recent years there is slow progress in general-purpose compression thus many researchers and practitioners have directed towards specialized compression. Specialized compression algorithms can be divided into two groups: completely novel algorithms and preprocessors coupled with general-purpose algorithms. A good example for the first group is DNA compression, where many specialized algorithms exist. The second group of algorithms transforms the input data

and then passes the transform's output to a general-purpose compressor. The transform removes redundancies and correlations not exploited by a general-purpose algorithm, what makes output data more compressible.

In our previous work we have presented algorithms specialized for text [18] and XML compression [19]. The biggest gain from our algorithms was achieved by creating dictionary of frequent words and replacing the words with shorter codewords. We have observed over 27% improvement with an LZ77-based algorithm on English texts using a fixed English dictionary. Using a fixed dictionary for XML documents was problematic because of the hardness to select a proper set of words, relevant across a wide range of real-world XML documents. Therefore we achieved the best results (over 35% improvement with an LZ77-based algorithm) with a semi-dynamic dictionary obtained in a preliminary pass over the input data. About 20% of this improvement was achieved by effective encoding of numbers, dates and times found in XML documents.

HTML format is similar to XML as both are SGML-based. On the other side, XML is mainly used as a standard for storage and interchange of structured information and HTML is mainly used to publish text-based information. In this usage HTML is more similar to texts.

The primary objective of our research was to design an efficient way of compressing HTML documents, which will reduce Internet traffic or will reduce storage requirements of HTML data. In this work we will present our two specialized algorithms for HTML compression. One uses a fixed English dictionary and the second one uses a semi-dynamic dictionary obtained in a preliminary

¹ This is an extended version of the poster: Skibiński, P.: *Improving HTML Compression*. Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA, (2008), pp. 545.

pass over the input data. Both algorithms are designed in four variants for proper general-purpose algorithms.

The map of the paper is as follows. Section 2 contains a short review of HTML format and existing HTML compression methods thus putting our work in a proper context. We also describe related word-based and XML compressors. In Section 3 we describe step-by-step our HTML transform, its main ideas, and its most significant details. The next section presents back-end compression algorithms used with our transform and details about optimizations for these algorithms. Section 5 contains implementation details, description of files used for experiments, and experimental results. Section 6 gives our conclusions and points out several issues for further study.

2 Related work

2.1 HTML description

HTML is a language that describes a structure of text-based information in a document. It denotes certain text as links, headings, paragraphs, and lists. It also supplements text with embedded images and other objects. XHTML is a new, XML-based version of HTML.

An HTML document consists of elements. An HTML element always has a start tag (e.g. `<element-name>`) and may have an end tag (e.g. `</element-name>`) in opposite to XML or XHTML, where the end tag is required. Elements may have two basic properties: attributes, contained in the start tag (e.g. `<element-name attribute="value">`), and content, located between the tags (e.g. `<element-name>Content</element-name>`). If there is no content a start tag and an end tag can be presented in short form that is `<element-name/>`. Comments in HTML are delimited by `<!--` and `-->` sequences. Some HTML elements, for example `
`, do not have any content and must not have a closing tag.

The following example [10] contains document title (`<title>` element), heading (`<h1>` element), paragraphs (`<p>` elements) and links (`<a>` elements):

```
<html>
<head><title>About the Test Data</title></head>
<body>
<h1 align="center">About the Test Data</h1>
<p align="center">Matt Mahoney<br>
Last update: Dec. 17, 2006.
<a href="text.html#history">History</a>
<p>The test data for the <a
href="text.html">Large Text Compression
Benchmark</a> is the first 10<sup>9</sup> bytes
of the English Wikipedia dump on Mar. 3, 2006.
```

2.2 HTTP protocol

Hypertext Transfer Protocol (HTTP) is a communications protocol used to transfer information on World Wide Web. HTTP is a request/response protocol between a client and a server. The client is making an

HTTP request to the server, which delivers HTML files, images and other.

HTTP compression [13] is the technology used to compress contents from a web server (an HTTP server) and to decompress them in an user's browser. HTTP compression is a recommendation of the HTTP 1.1 protocol specification as it reduces network traffic and improves page download time on slow networks [15]. It is especially useful when size of the web pages is large.

The experiments conducted by Wan [21] showed that HTTP compression can be improved utilizing the previously requested files in a browsing session as a dictionary, but this idea was not embedded in HTTP protocol until today.

The popular LZ77-based gzip was intended to be the HTTP compression algorithm. Currently, HTTP servers and clients supports also LZ77-based deflate format. Lighttpd server supports also BWT-based bzip2 compression, but this format is only supported by lynx and some other console text-browsers. Deflate, gzip, and bzip2, however, are general-purpose compression algorithms and much better results can be achieved with a compression algorithm specialized for dealing with HTML documents.

2.3 Word-based compression

StarNT [20] is a dictionary-based scheme, which replaces natural language words with references to an external dictionary. A word in StarNT dictionary is a sequence of symbols over the alphabet $[a..z]$. There is no need to use uppercase letters in the dictionary, as there are two one-byte flags (reserved symbols), `fcl` and `fuw`, in the output alphabet to indicate that either a given word starts with a capital letter while the following letters are all lowercase, or a given word consists of capitals only. Another introduced flag, `for`, prepends an unknown word. Finally, there is yet a collision-handling flag, `fesc`, used for encoding occurrences of flags `fcl`, `fuw`, `for`, and `fesc` in the text.

The ordering of words in the dictionary D , as well as mapping the words to unique codewords, are important for the compression effectiveness. StarNT uses the following rules:

- The most popular words are stored at the beginning of the dictionary. This group has 312 words.
- The remaining words are stored in D according to their increasing lengths. Words of same length are sorted according to their frequency of occurrence in some training corpus.
- Only letters $[a..zA..Z]$ are used to represent the codeword (with the intention to achieve better compression performance with the backend compressor). Each word in D has assigned a corresponding codeword. Codewords' length varies from one to three bytes. As only the range $[a..zA..Z]$ for codeword bytes is used, there can be up to $52 + 52 \times 52 + 52 \times 52 \times 52 = 143,364$ entries in the dictionary. The first 52 words have codewords: `a`, `b`, ..., `z`, `A`, `B`, ..., `Z`. Words from the 53rd to the 2756th have codewords of length 2: `aa`, `ab`, ..., `ZY`, `ZZ`; and so on.

WRT [18] is an English text preprocessor, a successor of StarNT. WRT replaces words in input text file with shorter codewords and uses several other techniques to improve performance of latter compression

The dictionary is sorted according to the frequency of words as more frequent messages should be represented with shorter codes than less frequent messages. WRT English dictionary have 80,000 words. Each word in D has assigned a corresponding codeword. Codewords' length is variable and span from one to four symbols. Ordinary text files, at least English ones, consist solely of ASCII symbols not exceeding 127, so codewords' alphabet has 128 symbols (ASCII values from 128 to 255). If there is a symbol from codewords' alphabet in the input file, then WRT outputs token t_{esc} and this symbol. Codewords' alphabet (128 symbols) is divided into four separate parts. WRT uses the mapping $\langle 101, 9, 9, 9 \rangle$ for codewords, and thus there are $101 + 101 \cdot 9 + 101 \cdot 9 \cdot 9 + 101 \cdot 9 \cdot 9 \cdot 9 = 82,820$ distinct codewords available. It is enough for 80,000 words WRT dictionary. The codeword bytes are emitted in the reverse order, i.e., the range for the last codeword byte has always 101 values.

WRT uses several additional techniques to improve the compression performance. First is q -gram replacement, which is based on substituting frequent sequences of q consecutive characters, i.e., q -grams, with single symbols. The next technique that improves the compression performance is End-of-Line (EOL) coding. The general idea is to replace EOL symbols with spaces and to encode information enabling the reverse operation in a separate stream. The last technique used by WRT is surrounding words with spaces, which converts all words to be surrounded by space characters. This technique gives gain only if there are at least a few occurrences of the word, because it joins similar contexts in PPM compressor (helps in better prediction of the word's first symbol as well as the next symbol just after the word).

mPPM [1] is a text compressor, which is based on Shkarin's PPMd [16]. mPPM splits text into alternating sequences of words and non-words. Words and non-words use a common dynamic dictionary.

Each item has assigned a codeword, which always consists of two bytes. Therefore the dictionary may include up to 2^{16} items. If the dictionary is bigger least recently used (LRU) words are removed. Two codewords are reserved. The first is the End-Of-File flag and the second signals occurrence of a new item.

mPPM uses two separate PPM models. The first encodes only codewords. The second, auxiliary model encodes new items with a standard character-based PPM.

HufSyl [9] and LZWL [9] are the first text compressors that use syllables as units, instead of characters or words. Syllables are obtained by one of algorithms of decomposition into syllables. These algorithms use syllable-based compression in combination with respectively, adaptive Huffman and LZW coding.

These methods have their counterpart variants for whole words, which gave better results in our

experiments. We decided to include only results of word-based versions in our Table 2.

2.4 XML compression

Cheney's XMLPPM [4] is a streaming compressor which uses a technique named multiplexed hierarchical modeling (MHM). It switches between four models: one for element and attribute names, one for element structure, one for attributes, one for strings, and encodes them in one stream using PPMD+ or, in newer implementations, Shkarin's PPMd [16]. The tag and attribute names are replaced with shorter codes. An important idea in Cheney's algorithm is injecting the previous symbol from another model into the current symbol's context. Injecting means that both the encoder and decoder assume there is such a symbol in the context of the current symbol but do not explicitly encode nor decode it. The idea of symbol injection is to preserve (at least to some degree) contextual dependencies across different structural models.

SCMPPM [2] can be seen as an extreme case of XMLPPM. Instead of using only few models, it maintains a separate model for each element class. Every class contains elements having the same name and the same path from the document root. This technique, called Structure Context Modeling (SCM), wins over XMLPPM on large documents (tens of megabytes), but loses on smaller files. Also, SCMPPM requires lots of memory for maintaining multiple statistical models and under limited memory scenarios it may lose significantly, even compared to pure PPMd.

3 HTML Transform

In this section we present our two algorithms: Semi-Dynamic HTML Transform (SDHT) and Static HTML Transform (SHT). We introduce subsequent parts of our algorithms step by step.

3.1 End tag encoding

In the previous section we have described structure of HTML documents. In a well-formed HTML document, every end tag must match a corresponding start tag. This can hardly be exploited by general-purpose compression algorithms, as they maintain a linear, not stack-alike data model. The compression ratio can then be increased by replacing every matching end tag with merely an element closing flag.

Our transform puts elements on a stack when a start tag has appeared. The last inserted element is removed from a stack when an end tag has appeared. The problem with HTML is that not all elements must have a closing tag. It can be solved by ignoring elements that allow an end tag omission. The second problem with HTML is that some tags (e.g. $\langle p \rangle$) should have corresponding end tags, but human editors skip these closing tags. Moreover, web browsers do not report errors on documents of this kind. Therefore our transform allows

non-valid HTML documents. The above-mentioned problems do not occur in XHTML.

3.2 Quotes modeling

Attributes of HTML elements usually contain neighboring *equal* and *quotation mark* characters (e.g. `attribute="value"`). Sometimes attributes are encoded using *equal* and *apostrophe* characters (e.g. `attribute='value'`). We have found that replacing these two characters with a flag improves compression performance. We made the same with *quotation mark* and *angle right bracket (greater)* characters that closing start tags with attribute(s) (e.g. `<element-name attribute="value">`).

3.3 Spaces modeling

Layout of an HTML document (e.g., trailing spaces, tabulators, end of line symbols) is not relevant for web browsers, but it may be useful for human editors of a document. This kind of redundancy, typical to HTML documents created with editors caring about the output format, cannot be well exploited by general-purpose compression algorithms.

Our transform makes use of structural indentation by efficiently encoding the leading spaces in lines. For every line our transform counts number of occurrences for leading spaces with length from 1 up to 256 symbols. If number of occurrences for the certain length is higher than a predefined threshold our transform assigns a special codeword for leading spaces of this length.

3.4 Number encoding

Numbers appear very often in HTML documents. We found that storing numbers as text is ineffective. Numbers can be encoded more efficiently using a numerical system with base higher than 10.

In our transform every decimal integer number n is replaced with a single byte whose value is $\lceil \log_{256}(n+1) \rceil + 48$. The actual value of n is then encoded as a base-256 number. A special case is made for sequences of zeroes preceding a number – these are left intact.

Our transform encodes in a special way also other numerical data that represent specific information types. Currently our transform recognizes the following formats:

- dates between 1977-01-01 and 2153-02-26 in YYYY-MM-DD (e.g. “2007-03-31”, Y for year, M for month, D for day) and DD-MMM-YYYY (e.g. “31-MAR-2007”) formats;
- years from 1900 to 2155 (e.g. “1999”, “2008”)
- times in 24-hour (e.g., “22:15”) and 12-hour (e.g., “10:15pm”) formats;
- value ranges (e.g., “115-132”);
- decimal fractional numbers with one (e.g., “1.2”) or two (e.g., “1.22”) digits after decimal point.

Dates are replaced with a flag and encoded as a two bytes long integer whose value is the difference in days from

1977-01-01. To simplify the calculations we assume each month to have 31 days. If the difference with the previous date is smaller than 256, another flag is used and the date is encoded as a single byte whose value is the difference in days from the previous date.

Years are replaced with a sequence of two bytes representing respectively: the year flag and the difference between the actual year and 1900.

Times are replaced with a sequence of three bytes representing respectively: a flag signaling a time pattern (conforming to the presented notation), the hour in 24-hour convention, and minutes.

Value ranges in the format $x-y$ where $x < 65536$ and $0 < y-x < 256$ are encoded in four bytes: one for the range flag, two for the value of x , and one for the difference $y-x$.

Decimal fractional numbers with one digit after decimal point and value from 0.0 to 24.9 are replaced by two bytes: a flag and their value stored as fixed point integer. In case of those with two digits after decimal point, only their suffix, starting from the decimal point, is replaced with two bytes: a flag and the number’s fractional part stored as an integer.

3.5 Semi-dynamic dictionary

The backbone of the proposed transform is to replace the most frequent words with references to a dictionary. A semi-dynamic version of our transform constructs a separate dictionary for every processed document, but, once constructed, the dictionary is not changed during an HTML transform. The transform works in two passes. The dictionary is obtained in a preliminary pass over the data, and contains sequences of length at least l_{\min} characters that appear at least f_{\min} times in the document. The dictionary is sorted by word frequency and stored within the compressed file, thus making the reverse operation faster. Dictionary references are encoded using a byte-oriented prefix code, where the more frequent words have assigned shorter codewords (length varies from one to four bytes). The prefix code and the variables f_{\min} and l_{\min} depend on a back-end compression algorithm described in the next section. We have also tried a fully dynamic (one-pass) transform variant, but it gives much worse compression ratio as the same word can have assigned different codewords.

In the second pass of a semi-dynamic transform, the parsed data items are encoded in a byte-oriented manner (words, spaces and flags with a prefix code; numbers, dates, years, times, value ranges, and decimal fractional numbers with respective coding schemes), and then compressed with a compression algorithm and written to disk. We chose four algorithms of this kind: LZ77-based, LZMA/BWT-based, PPM-based, and PAQ-based, which are described in detail in the following section.

Our notion of a “word” is broader than its common meaning. Namely, semi-dynamic dictionary contains items from the following classes:

- ordinary words – sequences of lowercase and uppercase letters (a-z, A-Z) and 128-255 (which

supports, e.g., all languages with a Latin-based alphabet);

- start tags – sequences of characters that start with `<`, contain letters, digits, underscores, colons, dashes, or dots, and end with `>`. Start tags can also include one or more preceding spaces as HTML documents sometimes have regular arrangements of the lines in which individual tags very often begin in the same column, preceded with the same number of spaces,
- URL address prefixes – sequences of the form `http://domain/`, where `domain` is any combination of letters, digits, dots, and dashes,
- e-mails – patterns of the form `login@domain`, where `login` and `domain` are any combination of letters, digits, dots, and dashes,
- words in form `"&data;"`, where `data` is any combination of letters, representing HTML entities.

3.6 Matching shorter words

Our transform uses separate output alphabets for original words (not replaced with a reference to a dictionary) and codewords. Therefore it is easy to encode a part of a word, if the prefix matches some word in a dictionary but the whole word does not. Still, the gain we achieve in this way is insubstantial. Our algorithm with all above-mentioned ideas is called Semi-Dynamic HTML Transform (SDHT).

3.7 Static dictionary

Static HTML Transform (SHT) is similar to Semi-Dynamic HTML Transform (SDHT). The main difference is that a semi-dynamic dictionary is replaced with a static dictionary, which is embedded in the compressor and the decompressor.

There are two advantages of a static dictionary over a semi-dynamic dictionary: there is no need to make the first pass over the input data to create the semi-dynamic dictionary and there is no need to store the semi-dynamic dictionary within processed data to make decompression possible.

On the other hand a static dictionary is limited to some class of documents e.g. English language. The dictionary must be spread with the compressor and the decompressor. Moreover, a semi-dynamic dictionary contains words that are actually frequent in the document, not words that could potentially be frequent, as it is in the case of a static dictionary. Nevertheless for HTML documents a static English dictionary usually gives a better compression ratio than a semi-dynamic dictionary.

4 Back-end compression

Succinct word encoding appears to be the most important idea in Static HTML Transform (SHT) and Semi-Dynamic HTML Transform (SDHT). The dictionary references are encoded using symbols which are not existent in the input HTML document. If, however, one of reserved symbols occurs in the document, and is not a

part of an encoded word, the coder prepends it with a special escape symbol.

There are four modes of encoding, chosen depending on the attached back-end compression algorithm: LZ77-based [23], LZMA/BWT-based [3], PPM-based [5], and PAQ-based [11]. The encoding scheme, however, is the same for SHT and SDHT. In all cases, dictionary references are encoded using a byte-oriented prefix code, where the length varies from one to four bytes. Although it produces slightly longer output than, for instance, Huffman coding [8], the resulting data can be easily compressed further, which is not the case with the latter. Obviously, more frequent words have assigned shorter codewords.

4.1 LZ77-based compression

The LZ77 algorithm [23] finds duplicated sequences of bytes in the input data. The next occurrences of a sequence are replaced by a pointer to the previous occurrence. The pointer is encoded as a distance to the previous occurrence in a limited past buffer and a match length. Literals are encoded directly. Most LZ77 variants use Huffman coding for literals, match offsets, and match lengths. LZ77-based methods are the most widely-used compression algorithms. They are known for fast compression and very fast decompression, but limited effectiveness.

Gzip is a common LZ77-based compression algorithm. It uses a buffer (*sliding window*) for finding matches that has only 32 KB, which is mostly responsible for both very high compression speed and mediocre compression ratios. When a sequence of bytes does not occur anywhere in the previous 32 KB, it is emitted as a sequence of literal bytes. Match lengths are limited to 258 bytes. We used gzip 1.2.4 with default values in our experiments.

4.2 LZ77 optimized transform

In comparison to modern algorithms LZ77-based compressors are not complicated. For example, they do not predict characters on the basis of their context. The strength of LZ77 lies in succinct encoding of long matching sequences. In consequence, a transform optimized for LZ77 compression should attempt to:

- reduce the number of characters to encode;
- decrease the offset (in bytes) of the matching sequences;
- decrease the length (in bytes) of the matching sequence;
- virtually increase the sliding window, i.e., the past buffer in which matching sequences are looked for.

It appears that in case of LZ77 (but not necessarily LZMA, BWT, PPM, or PAQ), shortening the output of the transform improves the compression ratio. In accordance with this observation, we chose the biggest possible alphabet for codewords: byte values from 128 up to 255 and most values in range 0–31, plus a few more. These symbols are very rarely used in most HTML documents. If, however, one of these symbols occurs in

the document, and is not part of an encoded word, the coder marks it with a special escape symbol.

HTML elements contain usually textual content. Another important idea in LZ77 optimized transform is elimination of most spaces between words. Since usually a word is preceded by a single space, only exceptions from this rule require special treatment: only the positions where spaces should not be inserted are marked with a respective flag. Such an assumption is known as the spaceless word model [12].

Still, without spaces between words, there must be a way to detect codeword boundaries. In a LZ77 optimized transform dictionary references are encoded using a byte-oriented prefix code, where the length varies from one to three bytes. The first byte of the codeword can belong to one of three disjoint ranges:

- C_1 if it is a one-byte long codeword; there are $|C_1|$ such codewords available,
- C_2 if it is a prefix of two-bytes long codeword, followed by a single byte in the full possible value range; there are $|C_2| * 256$ such codewords available,
- C_3 if it is a prefix of three-bytes long codeword, followed by two bytes in the full possible value range; there are $|C_3| * 256 * 256$ such codewords available.

In this way, we obtain $|C_1| + |C_2| * 256 + |C_3| * 256 * 256$ codewords in total. As this is a kind of prefix code, all the codewords are immediately decodeable. The size of ranges C_1 , C_2 , and C_3 are set according to the size of the document to compress and the resulting dictionary size.

For SDHT a semi-dynamic dictionary contains sequences of length at least $l_{\min} = 2$ characters that appear at least $f_{\min} = 12$ times in the document. These values gave good results for most files used in the experiments.

4.3 LZMA-based compression

LZMA is a modern compression algorithm based on ideas from a LZ77 compression family. It also finds duplicated sequences of bytes in the input data, but it contains many improvements. Some of the major features of LZMA are sophisticated match parsing, working with large buffers (up to 1 GB), and low order contextual encoding of literals.

LZMA significantly improves compression ratio in comparison to LZ77-based algorithms at the cost of much slower compression (decompression speed is not much affected). LZMA is implemented in the well-known 7-zip [14] compression utility. We used LZMA 4.43 with default 8 MB dictionary in our experiments.

4.4 LZMA and BWT optimized transform

We found experimentally that a transform optimized for LZMA and BWT-based [3] (e.g. bzip2) compression should have similar characteristic and there is no need to create separate versions.

In the LZMA/BWT optimized transform the codeword alphabet consists of fewer symbols than LZ77 optimized transform. It uses only 128 symbols with byte values from 128 up to 255.

In the LZMA/BWT optimized transform dictionary references are encoded using a less dense variant of a byte-oriented prefix code, with non-intersecting ranges for different codeword bytes. We use only two disjoint ranges of bytes, C_1 and C_2 , but the codeword lengths still span from 1 to 3 bytes. Any codeword byte from the range C_1 is unambiguously recognized as the suffix byte. In this way, we have $|C_1|$ one-byte codewords, $|C_2| * |C_1|$ two-byte codewords, and $|C_2| * |C_2| * |C_1|$ three-byte codewords. Such a reversed byte order was found to improve a compression ratio.

As well as the LZ77 optimized transform the LZMA/BWT optimized transform uses spaceless word model. For SDHT a semi-dynamic dictionary contains sequences of length at least $l_{\min} = 2$ characters that appear at least $f_{\min} = 12$ times in the document. These values were found experimentally.

4.5 PPM-based compression

PPM [5] is an adaptive statistical compression method. A statistical model accumulates counts of symbols (usually 1-byte characters) seen so far in a given context. Thanks to that, an encoder can predict probability distribution for new symbols from the input data. The more skewed the probability distribution in contexts, the higher compression will result. Increasing the context length is beneficial for encoding symbols known in a given context, but amplifies the problem of efficient encoding of the symbols yet unseen in a given context (generally speaking, they are handled via an escape to a lower order model, but how to estimate the escape probability is a gross research topic).

HTML data might contain long repeated strings. These data are compressed with most PPM variants in a way far from optimal, as the highest order used by e.g. Shkarin's PPMd [16] is only 16. Skibiński and Grabowski [17] presented the PPMVC algorithm (PPM with variable-length contexts), a variant of PPM* [6] adapted to cooperate with modern PPM mechanisms. PPMVC extends the character-based PPM with string matching similar to the one used by the LZ77 algorithm.

The PPMVC mechanism works on maximum order contexts only; in shorter contexts the current symbol is encoded with an ordinary PPM model (namely, Sharkin's PPMd model was used).

In PPMVC (called PPMVC2 in [17]) each maximum order context holds a pointer to reference context (the previous occurrence of the context) and the minimum left match length. The left match length (LML) is the length of the common part of the active context and the reference context. LML, by definition, is always at least as large as the maximum PPM order. The right match length (RML) is defined as the length of the matching sequence between symbols to encode and symbols followed by the reference context.

When a character is encoded from the maximum order context, the longest LML is evaluated, using the last context's appearance. If it is below the minimal left match length ($minLML$), then the encoder uses ordinary PPM encoding (without emitting any escape symbol). In

the other case, the encoder uses this context to find the RML (zero or more) and encodes it using an additional global RML model.

There are two more ideas in PPMVC that improve the compression effectiveness. First is the minimum right match length (*minRML*). If the current right match length is below the *minRML* threshold, then PPMVC sets RML to 0. This assures that short matches are not used.

The second idea is to encode sequences of length being a multiple of the parameter d . For example, if there is a match of length 14, and d is 3, then only the first 12 characters of the match are encoded (the truncated characters might however be part of the next RML). In this way, matches are somewhat shorter than they could be, but their lengths are cheaper to encode. In the original PPMVC [17], RML was bounded by a constant, while in the current variant the maximum RML is automatically increased if very long matches are encountered.

PPMVC offers compression ratio higher than LZMA, and faster compression time. The PPMVC's drawback is that its decompression time is very close to its compression time, which means it is several times longer than gzip's or LZMA's decompression times. In our experiments we used PPMVC 1.2 with prediction model order 8 and 64 MB of model size.

4.6 PPM optimized transform

In the PPM optimized transform the codeword alphabet consists of the biggest possible alphabet for codewords: byte values from 128 up to 255 and most values in range 0–31, plus a few more.

In the PPM-friendly mode dictionary references are encoded using a prefix code, where the length varies from one to four bytes. The four disjoint ranges are of size $|C_1|$, $|C_2|$, $|C_3|$ and $|C_4|$, respectively. Namely, we have $|C_1|$ one-byte codewords, $|C_2| * |C_1|$ two-byte codewords, $|C_3| * |C_2| * |C_1|$ three-byte codewords, and $|C_4| * |C_3| * |C_2| * |C_1|$ four-byte codewords. The first byte of a codeword unambiguously defines its length. For instance, when encoding a two byte long codeword, a byte from the range of size $|C_2|$ will be followed by a byte from the range of size $|C_1|$. The parameters C_1 , C_2 , C_3 , C_4 are selected according to the size of the created dictionary, with the principle of maximizing the number of short codewords.

The PPM optimized transform does not use spaceless word model. For SDHT a semi-dynamic dictionary contains sequences of length at least $l_{\min} = 2$ characters that appear at least $f_{\min} = 64$ times in the document. These values gave good results for most files used in the experiments.

4.7 PAQ-based compression

PAQ [11] is a family of compressors, originally developed by Matthew Mahoney, based on context modeling. As opposed to most PPM variants, which use a character-based alphabet PAQ works on the bit level. In PPM a new symbol in a context must be encoded in lower orders using an escape mechanism. PAQ does not

use the escape symbol at all as in each step it must encode only 0 or 1.

The binary alphabet allows a new character in a context to be distinguished after first unseen bit, what is not possible in the case of PPM. This is the next improvement to the PPM algorithm. In the PAQ's coding stage a binary symbol is encoded with a predicted probability by an arithmetic encoder, like in the PPM algorithm.

Bit level coding in PAQ allows easy introduction of additional predicting models. PAQ8 uses several predicting models e.g., order- n models (n to 16), similar to the one used in PPM; a string matching model, similar to one used the LZ77 algorithm; and a number of text, multimedia, tabular, or binary data oriented models (e.g., for x86 executables or BMP images).

Mixing the prediction of individual models in PAQ8 is performed with several neural networks. The outputs of these networks are combined using a second-level neural network. Before submitted to an arithmetic coder, the outputs go through two stages of adaptive probability maps (APM). The APM mechanism is related to the secondary symbol probability estimation (SSE), known from the PPMII algorithm [16]. It updates the probability considering previous experience and the current context.

The main disadvantage of the PAQ8 algorithm are high memory requirements and low compression speed. It makes this algorithm unattractive from a practical point of view. This is why we prepared FastPAQ, stripped-down version of PAQ8, intended to improve compression and decompression speed. From PAQ8 we have left only the order- n models, and we have also simplified APM stages, in overall making it more similar to PPM. FastPAQ is still much slower than fast PPM variants, but achieves better compression ratios.

In our experiments we used FastPAQ8 with model size 140 MB.

4.8 PAQ optimized transform

In the PAQ optimized transform the codeword alphabet consists of fewer symbols than PPM optimized transform. It uses only 128 symbols with byte values from 128 up to 255.

In the PAQ-friendly mode dictionary references are encoded using the same prefix code as in the PPM optimized transform, where the length varies from one to four bytes. The four disjoint ranges are of size $|C_1|$, $|C_2|$, $|C_3|$ and $|C_4|$, respectively. Namely, we have $|C_1|$ one-byte codewords, $|C_2| * |C_1|$ two-byte codewords, $|C_3| * |C_2| * |C_1|$ three-byte codewords, and $|C_4| * |C_3| * |C_2| * |C_1|$ four-byte codewords. In the PAQ optimized transform, however, the parameters C_1 , C_2 , C_3 , C_4 are fixed and equal 64, 32, 16, and 16, respectively, what makes them better suitable for PAQ's bit-level predictors.

The PAQ optimized transform does not use spaceless word model. For SDHT a semi-dynamic dictionary contains sequences of length at least $l_{\min} = 2$ characters that appear at least $f_{\min} = 64$ times in the document. These values were found experimentally.

5 Experimental results

This section presents implementation details of the SDHT and SHT algorithms. It also contains description of files used for experiments and discussion on experimental results of the SDHT and SHT algorithms with four different back-end compression methods.

5.1 Implementation details

The SDHT and SHT implementation contains a fast and simple HTML parser built as a finite state automaton (FSA), which accepts proper words and numerical (including date and time) expressions. The parser does not build any trees, but treats an input HTML document as one-dimensional data. It has small memory requirements, as it only uses a stack to trace opening and closing tags. The parser supports the HTML 4.01 specification (e.g. allowed an end tag omission for some tags).

The SHT implementation uses a static English dictionary with about 80.000 words. In this dictionary, words are sorted with the relation to their frequency in a training corpus of more than 3 GB English text taken from the Project Gutenberg library. The words are stored in lower case as SHT implements the capital conversion method to convert the capital letter starting a word to its lowercase equivalent and denote the change with a flag. Additionally, SHT uses another flag to mark a conversion of a full uppercase word to its lowercase form.

SHT requires only one pass over the input data while SDHT works in two passes over the input data. In the first pass, a dictionary is formed and the frequency of each of its items is computed. For the semi-dynamic dictionary, we allocate 8 MB of memory. If the dictionary reaches that limit, it is frozen, i.e., the counters of already included words can be incremented but no new word can be added. Still, in practice we rarely get close to the assumed limit (which can also be changed with a program switch). The complete dictionary is stored within the compressed file, so this pass is unnecessary during decompression, making the reverse operation faster. The words selected for the dictionary are written explicitly, with separators, at the beginning of the output file. In the second pass, the actual transform takes place, data are parsed into proper words and numerical expressions and respectively encoded.

The crucial operation in the encoding is dictionary search. In SDHT a search function is called twice for each word in the document: first time during the semi-dynamic dictionary buildup, second time during the actual parsing and word encoding. The choice of a dictionary data structure can seriously affect the overall transform performance. We have decided to use a fixed-size (4 MB) array with chained hashing for search, which we previously tested in our work on a text transform [18]. Its advantages are simplicity, moderate memory usage, and $O(1)$ search time (assuming that a single word is read in constant time).

The reverse SDHT and SHT are simpler. Again we use an FSA, which now recognizes flags and codewords, and transforms them to the original form. Obviously, there is no real search in the dictionary, only lookups in $O(1)$ time per codeword.

Our implementation of SDHT and SHT has embedded four back-end compression algorithms: gzip, LZMA, PPMVC, and FastPAQ8. Of these, gzip is the fastest, but provides the lowest compression ratio. FastPAQ8 is the slowest, but gives the best compression effectiveness.

SDHT and SHT are truly lossless, i.e., they do not ignore the document layout (e.g., trailing spaces) and the decoded file is an exact copy of the encoded one. The transforms can handle any HTML documents with 8-bit (ISO-8859 and UTF-8) or 16-bit (Unicode) encodings. SDHT and SHT was implemented in C++ and compiled with MS Visual C++ 2008.

5.2 HTML corpus

In compression benchmarking, proper selection of documents used in experiments is essential. To the best of our knowledge, there is no publicly available and widely respected HTML corpus to this date. Therefore, we have based our test suite on entire common Internet web sites downloaded (without images, etc.) using WinHTTrack Website Copier. The resulting corpus represents a wide range of real-world HTML documents.

Detailed information for each group of the documents is presented in Table 1; it includes: URL address, number of files and total size of files. The size of a single file spans from 1 up to 296 KB.

<i>Name</i>	<i>URL address</i>	<i>no. files</i>	<i>Total size</i>
Hillman	hillmanwonders.com	781	34421 KB
Informatica	www.informatica.si	12	122 KB
Mahoney	www.cs.fit.edu/~mmahoney/	11	596 KB
MaxComp	maximumcompression.com	61	2557 KB
STL	www.sgi.com/tech/stl/	237	2551 KB
TightVNC	tightvnc.com	21	289 KB
Tortoise	tortoisesvn.net	393	5342 KB
Travel	travelindependent.info	69	3841 KB

Table 1: Basic characteristics for the HTML corpus used in the experiments

5.3 Compression ratio

The primary objective of experiments was to measure the performance of our implementation of the SDHT and SHT algorithms. For comparison purposes, we included in the tests general-purpose compression tools: gzip 1.2.4, LZMA 4.43, PPMVC 1.2, and FastPAQ8,

employing the same algorithms at the final stage of SDHT and SHT, to demonstrate the improvement from applying the HTML transform.

As we are not aware of any specialized algorithms for HTML compression we have compared our algorithms to well-known word-based text compression techniques: StarNT [20], WRT [18], HufSyl [9], LZWL [9], mPPM [1] and StarWE [22]. StarWE is based on WRT and gives almost identical results therefore its results were omitted. We have also tried to use XMLPPM [4] and SCMPPM [2], which work well with XHTML files, but it does not support HTML files. These algorithms are described in details in Section 2.

The first part of Table 2 contains results of word-based text compression algorithms. For each program and group of HTML documents a bitrate is given in output bits per input character, hence the smaller the values, the better. The last but one column includes an average bitrate computed for all the eight groups of documents. The last column presents the average improvement of preprocessors for all documents compared to the general purpose algorithms result.

The next parts of Table 2 contain compression results of the introduced HTML corpus using gzip, LZMA, PPMVC, FastPAQ, and our implementation of the SDHT and SHT algorithms combined with gzip, LZMA, PPMVC, and FastPAQ.

SHT with gzip achieves compression results better than all word-based text compression algorithms, including a PPM-based mPPM. Compared to the general-purpose compression tools, SDHT improves compression of the introduced HTML corpus on average by 4% in case of gzip, 0% for LZMA, almost 2% in case of PPMVC and about 1% for FastPAQ. SDHT is very fast and almost does not influence on compression and decompression speed of general-purpose compression algorithms. Moreover, it speeds up FastPAQ, because preprocessed data is smaller than original.

In the first section we were wondering if HTML is more similar to XML or to texts. Our experiments show that HTML is more similar to texts as Static HTML Transform (SHT) with a fixed English dictionary gives much better results than SDHT. SHT improves compression of the introduced HTML corpus on average by about 15% in case of gzip, 12% for LZMA, almost 8% in case of PPMVC and 10% for FastPAQ. Compression and decompression speed in comparison to SDHT is a little bit lower as there is a need to read a fixed English dictionary. SHT, however, allows to read the dictionary only once and processes all HTML documents in one run.

Concluding, SHT with gzip gives 15% improvement over gzip achieving comparable processing speed. Moreover, SHT with FastPAQ gives the best

compression effectiveness, which is 28% better than gzip without any transform.

To ease the comparison, Figure 1 shows size of compressed HTML corpus with all tested transforms and back-end compression algorithms.

6 Conclusions

HTML has many advantages, but its main disadvantage is verbosity, which can be coped with by applying data compression. HTML is usually used in combination with gzip compression, but gzip is a general-purpose compression algorithm and much better results can be achieved with a compression algorithm specialized for dealing with HTML documents.

In this paper we have presented the SDHT and SHT transform aiming to improve lossless HTML compression in combination with existing general purpose compressors. The main components of our algorithms are: a static dictionary or a semi-static dictionary of frequent alphanumeric phrases (not limited to “words” in a conventional sense), and binary encoding of popular patterns, like numbers and dates.

We have developed two versions of our transform: semi-dynamic (SDHT) and static (SHT). Both algorithms have some disadvantages. SDHT does not support streams as input (offline compression) as it requires two passes over an input file. SHT uses a fixed English dictionary required for compression and decompression. It might be the biggest obstacle for SHT to become standard.

Thanks to the SHT transform, however, compression ratio of the introduced HTML corpus was improved by as much as 15% in case of gzip, 12% for LZMA, 8% in case of PPMVC and almost 10% for FastPAQ.

SHT and SDHT have many nice practical properties. The transforms are completely reversible, i.e. the decoded document is an accurate copy of the input document. Moreover, SHT and SDHT are implemented as a stand-alone program, requiring no external compression utility, no HTML parser, thus avoiding any compatibility issues.

There is a way likely to increase the HTML compression further. Layout of an HTML document (e.g., trailing spaces, tabulators, end of line symbols) is not relevant for web browsers and can be transformed to a more compressible form. We expect that a lossy version of the SHT transform could produce a few percent better results for the price of further complication of the transform.

Acknowledgement

The author would like to thank Szymon Grabowski and Jakub Swacha for suggestions of possible improvements.

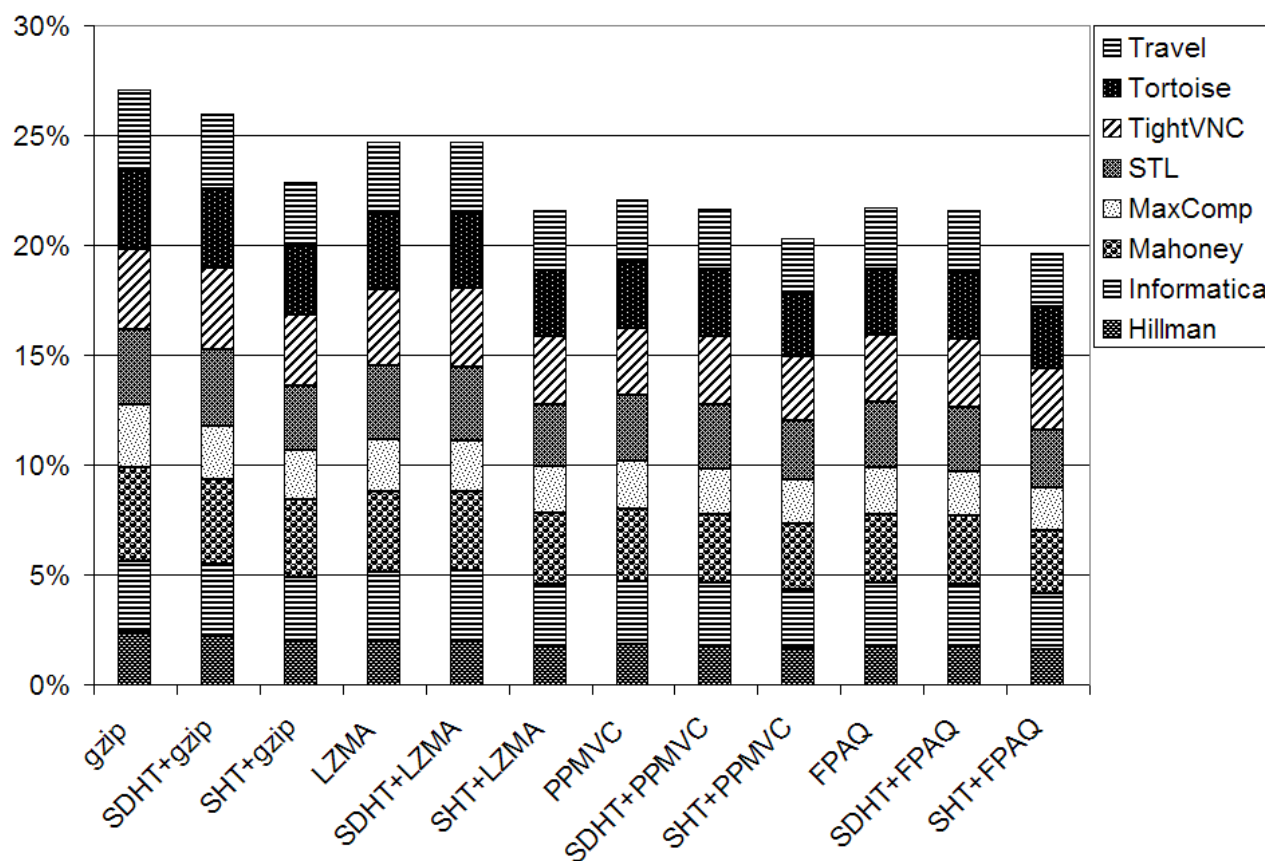


Figure 1: Size of compressed HTML corpus with different back-end compression algorithms

	Hillman	Informatica	Mahoney	MaxComp	STL	TightVNC	Tortoise	Travel	Average	Improvement
HufSyl	2.95	3.53	3.31	3.03	3.48	3.44	3.37	2.88	3.249	
LZWL	2.13	3.18	3.23	2.39	3.22	3.26	3.13	2.72	2.908	
gzip	1.51	2.08	2.72	1.86	2.19	2.34	2.27	2.34	2.164	
StarNT+gzip	1.42	1.94	2.54	1.79	1.97	2.17	2.08	2.06	1.996	7.74%
WRT+gzip	1.44	1.99	2.49	1.80	1.95	2.13	2.06	1.97	1.979	8.55%
mPPM	1.34	2.16	2.30	1.55	2.31	2.24	2.23	1.95	2.010	
gzip	1.51	2.08	2.72	1.86	2.19	2.34	2.27	2.34	2.164	
SDHT+gzip	1.40	2.13	2.45	1.56	2.20	2.39	2.31	2.19	2.079	3.93%
SHT+gzip	1.23	1.88	2.26	1.47	1.85	2.09	2.02	1.84	1.830	15.42%
LZMA	1.29	1.99	2.35	1.53	2.13	2.23	2.17	2.13	1.978	
SDHT+LZMA	1.29	2.04	2.30	1.46	2.16	2.29	2.21	2.07	1.978	0.00%
SHT+LZMA	1.13	1.78	2.08	1.38	1.79	1.99	1.92	1.74	1.726	12.71%
PPMVC	1.19	1.83	2.09	1.41	1.91	1.96	1.93	1.79	1.764	
SDHT+PPMVC	1.15	1.80	2.02	1.33	1.87	1.97	1.94	1.77	1.731	1.84%
SHT+PPMVC	1.06	1.71	1.92	1.30	1.71	1.86	1.83	1.60	1.624	7.94%
FPAQ	1.14	1.81	2.01	1.36	1.90	1.96	1.92	1.79	1.736	
SDHT+FPAQ	1.13	1.80	1.99	1.28	1.89	1.99	1.94	1.79	1.726	0.58%
SHT+FPAQ	1.01	1.65	1.83	1.24	1.67	1.82	1.77	1.56	1.569	9.65%

Table 2: Compression results for HTML datasets in output bits per input character.

References

- [1] Adiego, J., and de la Fuente, P.: *Mapping Words into Codewords on PPM*. String Processing and Information Retrieval, SPIRE, (2006), LNCS 4209, pp. 181–192.
- [2] Adiego, J., de la Fuente, P., and Navarro, G.: *Using Structural Contexts to Compress Semistructured Text Collections*. Information Processing and Management 43, 3 (May), (2007), pp. 769–790.
- [3] Burrows, M., Wheeler, D. J.: *A block-sorting data compression algorithm*. SRC Research Report 124. Digital Equipment Corporation, Palo Alto, CA, USA, (1994).
- [4] Cheney, J.: *Compressing XML with multiplexed hierarchical PPM models*. Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA, (2001), pp. 163–172.
- [5] Cleary, J. G., and Witten, I. H.: *Data compression using adaptive coding and partial string matching*. IEEE Trans. on Comm. 32, 4 (April), (1984), pp. 396–402.
- [6] Cleary, J. G., Teahan, W. J., and Witten, I. H.: *Unbounded Length Contexts for PPM*. Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA, (1995), pp. 52–61.
- [7] Deutsch, P.: *DEFLATE Compressed Data Format Specification version 1.3*. RFC1951, (1996), <http://www.ietf.org/rfc/rfc1951.txt>.
- [8] Huffman, D. A.: *A Method for the Construction of Minimum-Redundancy Codes*. Proc. IRE 40.9 (Sept.), (1952), pp. 1098–1101.
- [9] Lánský, J., Zemlička, M.: *Text Compression: Syllables*. Proceedings of the DATESO 2005 Annual International Workshop on Databases, TExts, Specifications and Objects. CEUR-WS, Vol. 129, pp. 32–45.
- [10] Mahoney, M.: *About the Test Data*, 2006, <http://cs.fit.edu/~mmahoney/compression/textdata.html>
- [11] Mahoney, M.: *Adaptive Weighing of Context Models for Lossless Data Compression*. Technical Report TR-CS-2005-16, Florida Tech., USA, 2005.
- [12] Moura E.S., Navarro G., Ziviani N.: *Indexing Compressed Text*. In Baeza-Yates R, editor, *Proceedings of the 4th South American Workshop on String Processing (WSP'97)*, Valparaiso, Carleton University Press, 1997; 95–111.
- [13] Nielsen H.F.: *HTTP Performance Overview*, 2003, <http://www.w3.org/Protocols/HTTP/Performance/>
- [14] Pavlov I.: *7-zip compression utility*. <http://www.7-zip.org>.
- [15] Radhakrishnan S.: *Speed Web delivery with HTTP compression*, 2003, <http://www-128.ibm.com/developerworks/web/library/wa-httpcomp/>
- [16] Shkarin, D.: *PPM: One Step to Practicality*. Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA, (2002), pp. 202–211.
- [17] Skibiński, P., and Grabowski, Sz.: *Variable-length contexts for PPM*. Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA, (2004), pp. 409–418.
- [18] Skibiński, P., Grabowski, Sz., and Deorowicz, S.: *Revisiting dictionary-based compression*. Software – Practice and Experience, 35(15), (2005), pp. 1455–1476.
- [19] Skibiński, P., Grabowski, Sz., and Swacha, J.: *Effective asymmetric XML compression*, Software – Practice and Experience, 38 (10), (2008), pp. 1027–1047.
- [20] Sun, W., Zhang, N., Mukherjee, A.: *Dictionary-based fast transform for text compression*. Proceedings of international conference on Information Technology: Coding and Computing, ITCC, (2003), pp. 176–182.
- [21] Wan, R.: *Browsing and Searching Compressed Documents*. PhD dissertation, University of Melbourne, 2003, http://www.bic.kyoto-u.ac.jp/proteome/rwan/docs/wan_phd_new.pdf
- [22] Yang, J., Savari, S.A.: *Dictionary-based English text compression using word endings*. Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA, (2007), pp. 410.
- [23] Ziv, J., and Lempel, A.: *A Universal Algorithm for Sequential Data Compression*. IEEE Trans. Inform. Theory 23, 3 (May), (1977), pp. 337–343.

Coordinated UAV Manoeuvring Flight Formation

Henry Hexmoor and Shahram Rahimi
 Department of Computer Science
 Southern Illinois University
 Carbondale, IL. 62901, USA
 E-mail: {hexmoor, rahimi}@cs.siu.edu

Jody T. Little
 Sierra Nevada Corporation
 4801 NW Loop, 410
 San Antonio, TX 78205, USA

Keywords: agents, collaboration, UAV, formation flight, affect

Received: May 28, 2007

A methodology is presented for real-time control of unmanned aerial vehicles (UAV) in the absence of a priori knowledge of location of sites in an inhospitable flight territory. Our proposed hostile control methodology generates a sequence of waypoints to be pursued on the way to the target. Waypoints are continually computed with new information about the nature of changing threat. The Dijkstra algorithm is used to account for a weighted combination of threat measures arising from the probability of encountering hostile ground to air fire as well as the internal urgency to complete the mission in the shortest time. UAVs broadcast latest sensed data to their counterparts. The sequence of waypoints defines the trajectory of the UAV to its target. By varying components of cost function, paths are altered to obtain a desired performance criterion. Validation of our methodology is offered by a series of agent-based simulations.

Povzetek: Predstavljena je metoda za upravljanje brezpilotnega letala ali helikopterja v sovražnem okolju.

1 Introduction

A powered, aerial vehicle that does not need a human pilot and uses aerodynamics forces to provide vehicle lift is called Unmanned Aerial Vehicle (UAV), also called a drone. UAVs are able to fly autonomously or can be remotely piloted. They might be expendable and can be equipped to carry lethal payloads. Ballistic missiles, cruise missiles, and artillery projectiles are not considered UAVs. Three varieties are shown in Figures 1-3.

UAVs were first conceived in the aftermath of the First World War and were used during the Second World War to train anti-aircraft gunners. Subsequently, early UAVs were large, remote-controlled planes until the late 20th century. Technological advances allowed the military to develop more capable fighting machines that could be used in performing dangerous missions without posing a significant risk to human life.

Military strategists have envisioned developing a wide array of roles performed by unmanned aircraft including deployment of ordinance [25]. UAVs are primarily used for intelligence, reconnaissance, surveillance. For example, the Global Hawk UAC (shown in Figure 1) provides the U.S. Air Force and joint battlefield commanders with near real-time, high resolution intelligence, surveillance, and reconnaissance

imagery. Predator (Figure 2) is a high endurance, medium altitude unmanned aircraft system used for surveillance and reconnaissance missions, as well as the refuelling of jets. Seiko Epson has developed the world's lightest and smallest UAV helicopter, the Micro Flying



Figure 1: A Global Hawk.



Figure 2: A Predator.



Figure 3: A Micro UAV.

Robot (shown in Figure 3). Sierra Nevada Corporation offers the most widely used UAV landing and takeoff software systems. Next, we outline the specific flight problem that is addressed.

Consider a ground to air protected hostile force zone that is of interest for surveillance by the friendly force. We need to compute the vulnerability in crossing the enemy region and find the safest trajectory across the region. Consider a swarm of coordinating reconnaissance UAVs, modelled as agents, which attempt to cross the hostile region. These UAV agents form a closely communicating friendly network and collectively produce estimates of threats needed to generate the safest possible flight trajectories.

The main objective of this work is to provide a novel approach for manoeuvring of a group of coordinating agents in a hostile environment. Due to the complexity of the problem involving formation and control of the UAVs in the real world, we are primarily concentrating on path generation [9].

2 Related work

The problem of multi-sensor surveillance involves detection of multiple intrusions and tracking of the intruders. Detection and target tracking have been explained from multiple viewpoints. To keep track of multiple moving objects researchers need to know all the joint probabilistic states of the all objects. These states grow exponentially with increase of number of moving objects. Sample-based variant of joint probability data association filters technique is used to track moving objects [22]. Mobile service agents are designed to operate in a dynamic environment by estimating the state of the dynamic object using probabilistic techniques [23]. The probabilistic estimates accurately identify the most likely state of each dynamic object. Multiple targets can be tracked by using dynamic time stamps. In order to detect, track, and avoid targets, researchers have used cluster-based approaches [12]. The focus of all these techniques has been building reliable estimators and trackers. These approaches do not use distributed sensors and are not directly useful for the problem of large area surveillance.

Within the context of distributed task allocation and sensor coordination [20] proposed a scheme for delegating and withdrawing agents to and from targets through the ALLIANCE architecture. The protocol for allocation was one based on the “impatience” of the

robot with respect to a target while withdrawal was based on “acquiescence.” [18] presented a strategy for tracking multiple intruders through a distributed mobile sensor network. Researchers have made significant advances in the areas of distributed sensor networks [7] and sensor management [18].

In [7], distributed robots cross a region using density estimates in a manner that facilitated maximal tracking of targets in that region. The decision for a robot to move to another region or to stay in its current region was based on certain heuristics. The method presented did not address collaborative or shared reasoning strategies for decision-making and action selection, such as the decision to move to a new area. Coordination between sensors was restricted to communicating their respective positions. Lesser’s group used sensor coordination for the purpose of tracking only one target [14].

In solving path planning and resource allocation problems researchers have considered UAVs to be independent of one another. Path planning methods focus on sequencing UAVs to arrive at specified locations or target locations [15]. Resource allocation methods have concentrated on target assignment, [16] and classification [5]. In the research of Ken Nygard [19], a scheme based on hierarchical decomposition was presented to assign a sub-team of UAVs to a particular task. McLain used a methodology that allowed the UAVs to reach the target simultaneously by avoiding pop-up threats whose locations were known a-priori [17].

This article presents a strategy that is a variant of methods described in [17] in that the UAVs can be coordinated to arrive at a desired target location or respective target locations within a specified time period. However, unlike McLain’s [17], locations of targets are not known beforehand. A set of waypoints are generated through a reinterpreted Dijkstra algorithm, and the computation is updated every time a *UAV agent* discovers new information. Our contribution is largely in our unique interpretation of path costs. There are numerous applications of this algorithm in robotic motion planning [8]. However, there are no known adaptations of this algorithm that suggest an emotional interpretation of path costs. This is a novel contribution in our approach. We have redefined the interned distance function from the standard algorithm, e.g., given in [6].

Simulations can be greatly simplified when the agent’s mission is executed using formation control strategy [1]. Formation control strategy is used to study aerodynamic effects involving multiple-aircraft [15] and to explore large areas with the aid of UAVs whose sensor capabilities are limited [19]. The aspect ratio¹ of the formation can be increased by using a formation control strategy. Two or more closely spaced agents can be treated as a single unit, which will decrease the workload of a remote operator and the cost of communication with

¹ The aspect ratio of two-dimensional shape is the ratio of its longest dimension to its shortest dimension.

other aircraft because the distance between aircrafts is decreased.

In many missions, UAVs fly in formations that are thought to increase chances of success. There are three different formation control strategies: leader following approach, behavioral approach, and virtual structures approach.

In the leader following approach, one of the agents acts like a leader, while the rest of the agents act like followers. The follower agents track the position and orientation of the leader. The problem with this approach is that the motion of the agent is controlled by the motion of the leader; if the leader agent fails in its motion then the following agent will not make the best move. Researchers have proposed variations of the *leader following approach* in order to prevent the centralized control of the agents' motion. In these approaches, agents designate multiple leaders, form chains, and create tree topologies. These approaches only solve the centralized motion problem partially, because subsets of the agents follow the designated leader agent. There have been a number of studies of leader-following techniques in the mobile robotics community. Using leader-following technique [5], robots can cooperatively move a box.

The goals of the behavioral approach are to prescribe several desired behaviors for each agent [1] and to make the control action of each agent a weighted average of the behaviors. Possible behaviors include collision avoidance, obstacle avoidance, goal-seeking and formation-keeping. There are also numerous variations on the behavioral approach to multi-agent coordination, most of which are derived by novel weight of the behaviors. In the behavioral approach agents use the decentralized motion method, which is desired in Multi Agent Systems. Botelho and Alami (1999) applied the behavioral approach to the problem of satellites in an equally distributed ring formation [2].

In the virtual structure approach, the entire agent formation is treated as a single structure. This approach defines the desired motion of the structure. Next, the motion of the virtual structure is translated into the desired motion for each agent. Finally, tracking controls for each agent are derived. It is very easy to implement this approach, which is its strength. However, applications in which we can use this approach are very limited.

Agent interaction in multiagent systems is associated with some form of communication [3]. Researchers rely on agent communication to solve standard multi-agent problems, like coordination and negotiation. The communication primitives that are exchanged among agents are typically referred to as communicative acts or speech acts. Some communicative acts are informing about the environment, querying about environment, telling about an action, advising other agents, and directing agents. These communicative acts help agents to make decisions. Agents have to use a standardized language format for exchanging information so that agents can easily understand each other. Several agent communication languages have been proposed by researchers aiming at standardizing the multi-agent

communication process. The two most notable ones are KQML and FIPA ACL [24], each using a slightly different syntax and set of communicative acts.

Intelligent agents, having incomplete knowledge of the operating environment, must learn the structure of the environment to better accomplish their tasks. This exploration may be performed in a different phase, or it may be combined with the task at hand. Algorithms which guide agents in unknown physical environments can be classified as full exploration algorithms or navigation algorithms [26]. Full exploration algorithms are used when the entire environment is mapped out a-priori. Navigation algorithms are used when a specific target location has to be reached. Modified search algorithms are widely used in solving these types of problems.

Hill climbing is a heuristic method of searching for solutions to problems that have huge solution spaces [21]. In the hill climbing approach, the problem space is converted to a graph, where each solution corresponds to a node associated with a value. The current path is extended with a successor node, which is closer to the solution than the end of the current path. In the simple hill climbing approach, the node closest to the present node is added to the solution space. In steepest ascent hill climbing, all successors are compared, and the one closest to the solution is selected. A fitness function is defined for evaluating the effectiveness of the node newly added to the solution. Unless the fitness function is smooth and effective, these two approaches will fail to reach global maximum. This happens when there are local maxima in the search space, which are not solutions. This can be partially overcome using varied hill climbing methods such as iterated hill climbing, stochastic hill climbing, random walks, and simulated annealing [26].

To avoid getting stuck on the first local maximum, several hill climbs are repeated, each time starting from different randomly chosen points. This method is known as iterated hill climbing. This approach increases the probability of reaching the global maximum value by detecting different local maximum points. If there are several local maximum points in problem space, it is not a good method to implement. This approach surprisingly gives optimal results in many applications.

In the stochastic hill climbing approach, a randomly chosen neighbor node is evaluated using the fitness function. This node is only retained if it increases the value of the fitness function; otherwise, another neighbor node is selected randomly for evaluation of the fitness function. Stochastic hill climbing usually starts from a random point. By combining iterated hill climbing with stochastic hill climbing, solution paths avoid getting stuck on local optimum values.

Simulated annealing is an optimization technique proposed by [13] by extending the Monte Carlo method to determine the equilibrium state of a collection of atoms at any given temperature T . Simulated annealing is inspired by a technique involving the heating and controlled cooling of a material that increases the size of its crystals and reduces their defect. The heat makes

atoms move randomly due to the high energy state. The slow cooling gives them more chance of finding more stable configuration than the initial one. Simulated annealing consists of randomly choosing a solution from the neighboring nodes of the current node. If the value of the fitness function increases the current one, the new node is accepted as the new current node. If the fitness function is not improved, the new solution is retained with probability. Simulated annealing differs from hill climbing algorithms in that it allows the possibility of going downhill if the temperature is high enough. When the temperature is high, the motion is more random. Because the probability of going downhill is inversely related to temperature, when temperature decreases, it is hard to go downhill. Simulated annealing generally starts with high temperature values. The slower the cooling, the easier it becomes to find the global solution. Infinitely slow cooling certainly produces a global optimum solution, although it might take infinite time. The main difficulty of simulated annealing is to find an appropriate temperature decrease rate.

Thus far, we have discussed the research related to coordination and communication in multiagent systems. In the following sections, we will present details of our approach.

3 Implementation

We have made certain abstractions while transforming the real world situation to a simulated environment. We also make assumptions about the modelling of inter-agent communication. We define our simulation system by describing agents present in the system, their capabilities, and the missions assigned to the agents. There are two different types of agents present in our system: friendly UAV agents and hostile, ground-based agents. UAV agents are autonomous and compute their trajectory dynamically. They perform surveillance over a square-shaped surveillance zone. The main objective for UAV agents is to move from a location outside the surveillance area to a designated target zone. These agents use a variety of algorithms to achieve their objective. They continually compose and update a trajectory to reach the target location consistent with perception of their surroundings. Our simulation allows for user-defined number of UAV agents to be modelled. These agents communicate with one another to effectively analyze the surveillance zone.

Hostile agents protect the surveillance area from intrusion. From the perspective of UAV agents, hostile agents are considered to be threatening. Hostile agents are static and reactive. Hostile agents operate independently of one another. The objective of these agents is to harm unknown vehicles within their perceptual range. Although they are an integral component of our system, they are not considered to be part of the multi-agent component as they do not communicate or coordinate among themselves. They remain in the same location throughout the simulation. Therefore, once UAV agents identify the location of a hostile agent, that hostile agent need not be sensed again.

Hostile agents are capable of firing anti-aircraft missiles. The success of hitting UAV agents depends on the distance of the UAV agent from the hostile agent. In our simulation, UAV agents cannot be completely destroyed by anti-aircraft missiles. Each UAV agent will continue to travel the surveillance area even after it has sustained a strike by an anti-aircraft missile. This assumption is made so that we can compute the effectiveness of the path algorithm used for the reconnaissance mission.

We introduce a number of assumptions in order to define our model. Hostile agents are guarding the surveillance area in which UAV agents are attempting to cross. UAV agents divide the surveillance zone into number of square cells for the trajectory computation. The firing strength of the hostile agent is the probability that it can shoot down a UAV agent when the UAV agent flies over the cell occupied by it. The probability of hitting a UAV agent decreases as the distance of the UAV agent from the cell occupied by hostile agent increases. As shown in Figure 4, in an eight-connected sense, the occupied cell by a hostile agent is denoted by O, the four nearest neighbors are denoted by N1 and the four diagonal neighbors by N2. If the distance between the centers of two four-connected cells is termed a unit, cells that are two units away are denoted by N2. Figure 4 is used to compute actual firing strength of hostile forces. In contrast, Figure 8 will be used to reflect the perspective of a friendly UAV flying over the hostile territory and how they perceive threats from hostile agents. Since UAVs share information, subsequent figures show cumulative perceptions of threat.

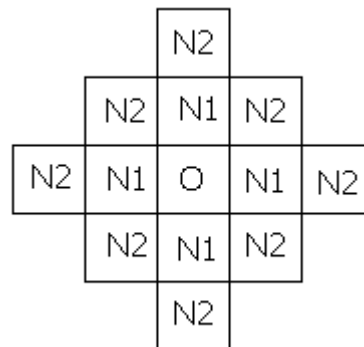


Figure 4: Labeling of cells with respect to a central hostile cell denoted by O; i.e., the cell that is occupied by a hostile agent.

Direct firing strength is defined as the probability of hitting a UAV agent when the UAV agent is flying over cell O. First neighboring firing strength is defined as the probability of hitting UAV agent from the cells labelled N1. Second neighbor firing strength is defined as the probability of hitting UAV agent from the cells labelled N2.

For a UAV agent that flies over a cell i , the probability that it is hit by a hostile agent is computed by Equation 1.

$$p_{fi} = \begin{cases} fs_o; i \in O \\ fs_{N1}; i \in N1 \\ fs_{N2}; i \in N2 \\ 0; elsewhere \end{cases} \quad fs_o > fs_{N1} > fs_{N2}$$

(Equation 1)

fs_o is direct firing strength, fs_{N1} is firing strength at the first neighbor's level, and fs_{N2} is firing strength at the second neighbor's level.

If a UAV agent flies directly over an occupied cell, the UAV agent has the maximum probability of being hit, but its probability of being struck decreases if it flies over cell N1 and further decreases if it flies over cell N2.

p_{fj} is the probability that a hostile agent situated at cell j can fire successfully at the UAV agent flying over cell i .

Thus, for a UAV agent flying over cell i , the cumulative probability of being shot is the sum of the individual probabilities of being successfully fired at by the hostile agents positioned at the cell beneath or any of the N1 or N2 cells. The cumulative probability of being shot is given by P_{si} , i.e.,

$$P_{si} = \sum_j p_{fj}, j \in O \text{ or } j \in N1 \text{ or } j \in N2$$

(Equation 2)

In the absence of prior information regarding the occupancy of the cells in the region, the prior probability of cell j being occupied by a hostile agent is denoted by p_{Oj} . The cumulative probability of being shot when flying over cell i and lacking such information regarding the occupancy of the cells is given by Equation 3.

$$P_{si} = \sum_j p_{Oj} p_{fj}, j \in O \text{ or } j \in N1 \text{ or } j \in N2$$

(Equation 3)

It is assumed that the a-priori information regarding the number of hostile agents, n_{HA} , is known while their coordinates or the cells that they occupy are unknown. If the total number of cells is n_C , then $p_{Oj} = \frac{n_{HA}}{n_C}$. As the

UAV agents move over the habitat they obtain information regarding the occupancy or non-occupancy of a cell. This information is then broadcast to the other UAV agents. The occupancy probability of a cell for which there is no information yet is recomputed as in Equation 4.

$$p_{Oj} = \frac{n_{HA} - n_V}{n_C - v} \tag{Equation 4}$$

In Equation 4, n_V represents the number of cells that are occupied by hostile agents. Here, v represents the total number of cells for which the UAV agent has obtained information about.

Distance anxiety denotes the anxiety experienced by the UAV agents to travel unit distance over the hostile area. We use w to denote distance anxiety. When $w=0$ the UAV agent has the maximum innate distance anxiety. This encourages UAV agent to seek the path with the shortest possible distance. At $w=1$ the UAV agent will not have any distance anxiety. At this value, UAV agent will seek paths of least probabilistic resistance that need not be optimal in terms of distance. It is to be noted, however that increasing w does not always imply increasing path lengths – it only implies paths with reduced chances of being fired at or a search that is biased towards least probabilistic paths. Decreasing w does not always imply paths along which there are increased chances of getting fired at – it only implies a search that is biased towards shortest distance paths. In other words, a path of least probabilistic resistance obtained for $w=1$ could well be a path that is shortest in terms of distance.

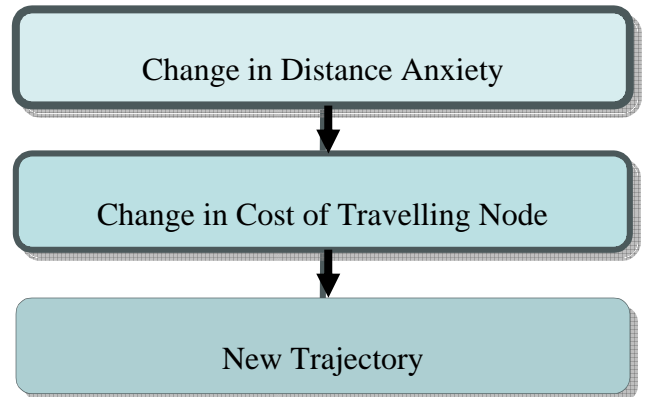


Figure 5: The effect of distance anxiety on trajectory computation.

We used parameters to control the communication behavior in the system. Communication range denotes the number of UAV agents to whom UAV agent can send messages. Communication penalty is the cost of communicating in the system for a single unit of distance. It is directly related to the distance that the UAV agent travels. If UAV agents travel a long distance the cost of communication will increase.

3.1 Path Generation

In this model, UAV agents have to traverse a square hostile surveillance area. While crossing this area, UAV agents compute trajectory by considering the hostile elements on the ground. This trajectory is also influenced by the UAV agent's distance anxiety value, which can cause the computed trajectory to have fewer steps. We used the standard Dijkstra algorithm to create our path generation algorithm, for algorithm consult (Knuth, 1997). The nodes in Dijkstra's algorithm

correspond to our UAV agent surveillance cells. Similarly, the edges correspond to adjacency among our surveillance cells. For an edge directed from cell m to cell n , the cost of traversing the edge C_{mn} is given by the weighted combination given in Equation 5.

$$C_{mn} = (1 - w)\psi d_{mn} + wP_{sn} \quad (\text{Equation 5})$$

P_{sn} is the same as the left-hand side of Equation 3; i.e., the cumulative probability of being shot at cell n . The distance between the cells in the Euclidean distance denoted by d_{mn} , while ψd_{mn} represents the fatigue accrued by the UAV agent after traveling a distance of d_{mn} . ψ is a normalization constant that allows d_{mn} to be scaled to similar values as P_{sn} . w is the weighing factor in the range of 0.0 to 1.0. To recapitulate, Equation 5 suggests that the cost of traveling from cell m to cell n is the sum of the overall probability of being fired at while over cell n as well as the fatigue developed with distance.

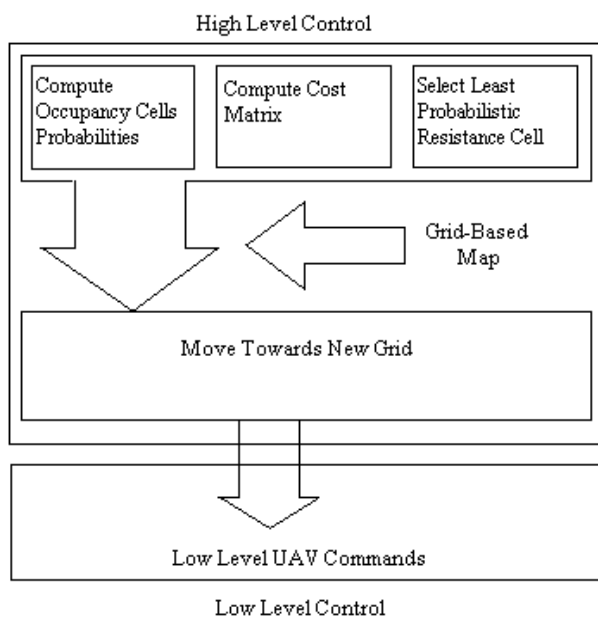


Figure 6: High level flow of algorithm.

Figure 6 depicts the high level flow of the algorithm. Paths are updated whenever information regarding a new cell in the form of either presence or the absence of a hostile agent occupying it is discovered. The occupancy probabilities of unobserved cells are recomputed by using Equation 4, and the cost matrix is updated by using Equation 5. Change in the distance anxiety value triggers the change in the cost of traversing two nodes as shown in Figure 5. This change will affect the waypoints generated by the path generation algorithm. The algorithm we devised for computation of the waypoints for reaching target locations uses a hill climbing technique. The waypoints are always computed from the current position to the target position. The neighbor node, which produces the optimal solution from the current node, is always added to solution space. This is similar to the steepest ascent hill-climbing search.

The overall algorithm given a specific parametric value w is given in Figure 7.

- For all UAV agents alive in the system, do steps 2 and 3 until the last waypoint or target is reached.
- If new information is available about the presence or absence of a hostile agent occupying a cell, continue with the following:
 - Update occupancy probabilities at all cells.
 - Compute the new cost matrix based on the most recent probabilities.
 - Search in the space of w for all paths that satisfy the time upper bound.
 - Select the path with the least probabilistic resistance or distance metric.
 - Move towards the next waypoint.

Figure 7: Overview of system architecture

4 Simulation details

A piece of surveillance area is shown in Figure 8. We divided the surveillance area into cells. By doing so, we can observe the enemy location and UAV agent motion. We also can understand the reason for UAV agent’s motion in the surveillance area. In contrast to Figure 4, Figure 8 depicts the UAV perspective.

As mentioned earlier, the hitting probability for a cell can be computed by adding the likelihood of hostile agent being present in the vicinity. Cells that are located either in the corners or at the sides of the surveillance area have fewer neighbors. Consequently, these cells have the least probabilistic values, so the path computed by UAV agents using the algorithm will be always along these cells. We assigned shading for cells to depict their perception of levels of threat.

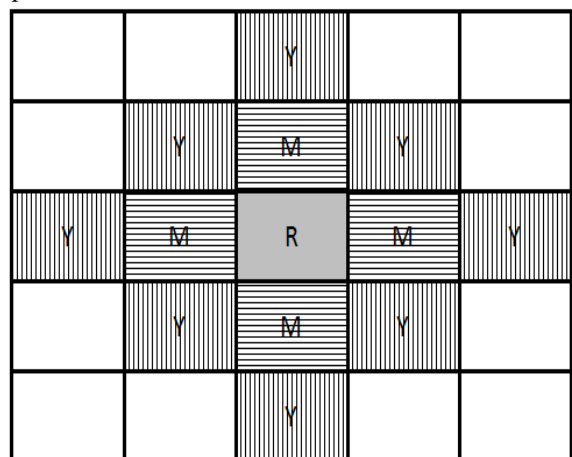


Figure 8: Shading assignment for UAV perception of the surveillance zone: white shade denotes no threat, solid gray denotes maximum threat, horizontal shade denotes moderate threat, vertical shade denotes minimal threat.

We identified the perception of severity of threats for each cell by using a specific shade. White shaded cells are neutral areas, i.e., these are perceived to be free of hostile agents. These cells are not used for computing paths. Perception of a hostile agent at a cell is denoted by

solid gray shade. The first neighboring cells of a hostile agent are shaded horizontally. The second neighboring cells of a hostile agent are shaded vertically. Figure 9 shows our GUI where we can assign different parameter values for the simulation system. These values can be set using the sliders shown in Figure 9. We can change *firing strength* values and *distance anxiety* values for the simulation. These values can be changed at run time.

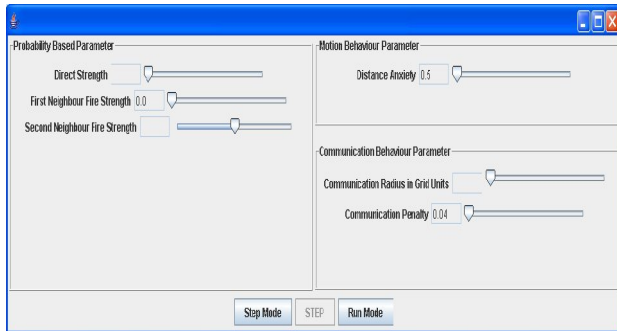


Figure 9: Simulation system controls: The GUI.

5 Results

We simulated the environment of a UAV surveillance *Multi Agent System* and carried out experiments. All of the assumptions that we mentioned in the beginning of Section 5 were made for this simulation.

The figures 10, 11 and 12 show paths of *UAV agents* for $w = 1, 0.5, \text{ and } 0$ respectively. *UAV agents* themselves are not aware of their own locations beforehand until one of them identifies a *hostile agent* during flight. Since UAVs share information, subsequent figures show cumulative perceptions of threat. A *UAV agent's* sensing range at any instant is defined as the area covered by 9 surrounding cells, three along the length and three along the breadth. In Figure 10, the cells are shaded.

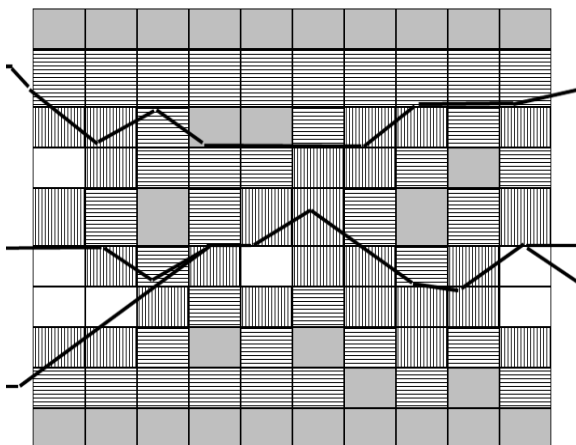


Figure 10: The paths traced by the *UAV agents* for $w=1$. Threat perceptions for cells are indicated by shadings: gray for most threat, vertical for the least threat, and horizontal for moderate threat.

As expected, despite their lack of apriori knowledge regarding the locations of *hostile agents*, for $w = 1$, the

UAV agents manage to find their paths through the cells where the probability of getting fired upon is minimal. On the contrary, for $w = 0$, the *UAV agents* move bravely through hostile cells to minimize their distance. The shortest distance is not a straight line between start and target locations since the graph search is through an eight-connected lattice. For $w = 0.5$, the paths turn out to be neither the shortest distance paths nor the shortest probability paths but paths that minimize $C_{mn} = 0.5(\psi d_{mn} + P_{sn})$. The sum of probabilities can exceed 1.0 since it represents the total chances of being fired upon whenever the *UAV agent* visits the cell and based upon the path computed at that instant. While it is this sum that gets minimized during the Dijkstra search for paths of least probability, this is different from the computation that evaluates the probability that a *UAV agent* gets past λ cells safely. That computation is given

$$\text{by } \prod_{i=1}^{\lambda} (1 - P_{si}) \text{ and is the safety factor of a path that}$$

traverses λ cells. However, it can be shown that at any given instant based on the knowledge of the environment, the path that minimizes the sum of probabilities would also be the one that gives the maximum value of the safety factor.

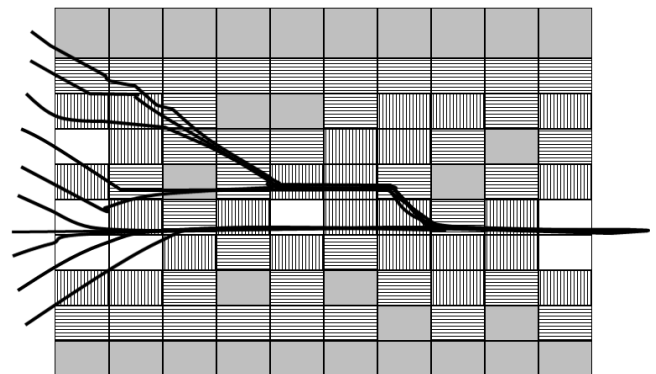


Figure 11: The paths traced by the *UAV agents* for $w=0.5$

It is seen that for $w = 0$ the paths with least prior and posterior probability sums, but with maximum distances, are obtained. This is due to the fact that the paths of least probabilities go through the center of the environment and the *UAV agents* that enter the habitat at its top and bottom search their way to the center due to lack of prior information about the *hostile agents* locations. For $w=1$, *UAV agents* take the shortest path with the least amount of fatigue. While computing paths with $w=1$, *UAV agents* will travel in an area with a high density of *hostile agents* to reduce the distance. This will increase their chances of being fired upon by *hostile agents*.

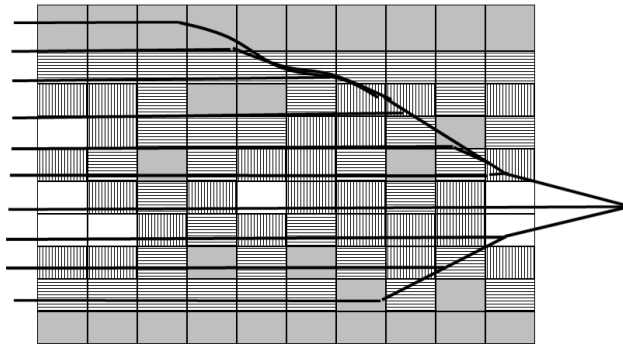


Figure 12: The paths traced by the *UAV agents* for $w=0$.

6 Conclusion

A methodology for parametric flight control of multiple *UAV agents* is presented so that a desired criterion is met. The desired criterion can be one that minimizes distance fatigue or the path that minimizes the chances of being fired upon. The method works for situations where a-priori knowledge of the hostile habitat is not available. It also lends itself to situations where the trajectories of *UAV agents* can be modified dynamically by adjusting a weighing factor w , since the paths are recomputed every time new information about the location of hostile agents is discovered. The effect of communication on the simulation system is observed. Communication in the simulated *multiagent system* improved the performance of the system by rapidly increasing the number of known cells to the agents at any time. The path computed by the algorithm chose waypoints towards the cells whose information is known. If *UAV agents* start at different intervals of time, the computed trajectory moves towards the known cells whose information is gathered by the *UAV agent*, which started the mission earlier. Due to this, *UAV agents* may not explore the surveillance area in which they are moving. The main objective was to find the safest path based on a given set of constraints.

Future scope of this work includes incorporating communication constraints such as latency, minimum distance to be maintained between *UAV agents* for information exchange, and an investigation into the role for embedding the *UAV agents* with social notions like autonomy and benevolence that yielded useful results in [13].

Acknowledgements

This research is sponsored by the Air Force Research Laboratory (AFRL) under contract FA8750-06-C-0138. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of AFRL or the United States Government. Special thanks go to the Net Centric SIGINT Focused Information Enterprise (NCSFIE) at AFRL for its support.

References

- [1] Balch, T., & Arkin, R.C. (1998). Behavior-based formation control for multi-robot teams. *IEEE Transactions on Robotics and Automation*, 14(6), pp 926-939.
- [2] Botelho, S., & Alami, R. (1999). A scheme for multi-robot cooperation through negotiated task allocation and achievement. In *Proceedings of the International Conference on Robotics and Automation*, pp 1234-1239.
- [3] Cannan, J.W. (1999). Seeing more, and risking less, with UAVs. *Aerospace America*, 37(10).
- [4] Chandler, P. R., Rasmussen, S. J. (2002). Military applications: Unmanned aerial vehicles: Multi UAV: a multiple UAV simulation for investigation of cooperative control. In *Proceedings of the 34th Conference on Winter Simulation: Exploring New Frontiers*.
- [5] Chandler, P., Rasmussen, S., & Pachter, M. (2000). UAV cooperative path planning. In *Proceedings of the AIAA Guidance Navigation, and Control Conference*, Denver, CO.
- [6] Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C. (2001). *Introduction to Algorithms*, Second Edition. MIT Press and McGraw-Hill. ISBN 0-262-03293-7. Section 24.3: Dijkstra's algorithm, pp.595–601.
- [7] Das, A.K., Fierro, R., Kumar, V., Ostrowski, J.P. (2002). A vision-based formation control framework. *IEEE Trans. Robot. and Automat.*, 18(5), 813-- 825.
- [8] Dudek, G., (2005). *Computational Principles of Mobile Robotics*, Cambridge University press.
- [9] Hexmoor, H., & Pasupuleti, S. (2003). Institutional versus interpersonal influences on role adoption. *International workshop on autonomous agents (AAMAS-03)*, Australia.
- [10] Kirkpatrick, S., Gelatt Jr., C.D., & Vecchi, M.P. (1983). *Optimization by simulated annealing*. *Science*, 220(4598), pp. 670-680.
- [11] Knuth, D.E. (1997). *The art of computer programming: Fundamental algorithms*, (Vol. 1, 3rd ed.). Addison-Wesley.
- [12] Krishna, M., Hexmoor, H., & Pasupuleti, S. (2003). Avoiding collision logjams through cooperation and conflict propagation. *International Conference on Integration of Knowledge Intensive Multi Agent Systems (KIMAS-03)*, Boston, Massachusetts.
- [13] Krishna, M., Hexmoor, H., & Pasupuleti, S. (2004). Role of autonomy in a distributed sensor network for surveillance. In *proceedings of the International conference on artificial intelligence (ICAI)*, Las Vegas, Nevada.
- [14] Lesser, V., Horling, B., Vincent, R., Miller, R., Shen, J., Becker, R., & Rawlins, K. (2001). Distributed Sensor Network for Real Time Tracking. In *Proceedings of the 5th International Conference on Autonomous Agents*. pp 417- 424. Montreal, Canada.

- [15] McLain, T., Chandler, P., Ramussen, S., & Pachter, M. (2001). Cooperative control of UAV rendezvous. *In Proceedings of the American Control Conference (ACC)*, pp. 2309-2314, Arlington, VA.
- [16] McLain, T., Beard, R., & Kelsey, J. (2002). Experimental demonstration of multiple robot cooperative target intercept. *In Proceedings of the AIAA Guidance and Control Conference and Exhibit*, Monterey, CA.
- [17] McLain, T., Beard, R., Goodrich, M., & Anderson, E. (2002). Coordinated target assignment and intercept for unmanned air vehicles. *IEEE Transactions on Robotics and Automation*.
- [18] Mataric, M., Sukhatme, G., & Ostergaard, E. (2003). Multi-robot task allocation in uncertain environments. *Autonomous Robots*, 14(2-3), pp. 255-263.
- [19] Nygard, K., Chandler, P., & Pachter, M. (2001). Dynamic network optimization models for air vehicle resource allocation. *In Proceedings of the American Control Conference (ACC)*, pp. 1853-1856.
- [20] Parker, L. (1993). Designing control laws for cooperative agent teams. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pp 582-587, Atlanta, GA.
- [21] Pierce, D., & Kuipers, B. (1991). Learning Hill-Climbing Functions as a Strategy for Generating Behaviours in a Mobile Robot. From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior, pp 327-336, Cambridge, MA: The MIT Press/Bradford Books.
- [22] Schulz, D., Burgard, W., Fox, D., & Cremers, A.B., (2001). Tracking Multiple Moving Targets with a Mobile Robot using Particle Filters and Statistical Data Association.. *In Proceedings of IEEE International Conference on Robotics and Automation*. Seoul Korea, pages 1665-1670.
- [23] Schulz, D. & Burgard, W. (2001). Probabilistic State Estimation of Dynamic Objects with a Moving Mobile Robot. *Robotics and Autonomous Systems*, 34(2-3).
- [24] Silva, V.T., Carlos, J. P., & Lucena, D. (2003). Doctoral papers: MAS-ML: a multi-agent system modeling language. *In Companion of the 18th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications*, Anaheim, CA.
- [25] Williams, W., Harris, M. (2002). The Challenges of Flight-Testing Unmanned Air Vehicles, in Systems Engineering, Test & Evaluation Conference, Sydney, Australia.
- [26] Yuret, D., and Maza M. (1993). Dynamic Hill-Climbing: Overcoming the limitations of optimization techniques. The Second Turkish Symposium on Artificial Intelligence and Neural Networks, pp. 208-221.

Robustness and Visualization of Decision Models

Andrej Bregar
 Informatika d.d.
 Vetrinjska ulica 2, SI-2000 Maribor, Slovenia
 E-mail: andrej.bregar@informatika.si

József Györkös and Matjaž B. Jurič
 University of Maribor, Faculty of Electrical Engineering and Computer Science, Institute of Informatics
 Smetanova ulica 17, SI-2000 Maribor, Slovenia
 E-mail: jozsef.gyorkos@uni-mb.si, matjaz.juric@uni-mb.si

Keywords: decision support, multi-criteria decision analysis, robustness metrics, mathematical optimization, principal components analysis, utility theory, promethee, electre

Received: June 20, 2008

Robustness analysis and visualization are two of key concepts of multi-criteria decision support. They enable the decision-maker to improve his understanding of both the model and the problem domain. A class of original mathematical optimization based robustness metrics is hence defined in this paper. In addition, several efficient existing techniques that have been successfully used in various ICT projects are presented. They include the stability intervals/regions and the principal components analysis. All approaches are applied to the multi-attribute utility function, and to the PROMETHEE II and ELECTRE TRI methods. Their benefits are discussed and demonstrated on real life cases.

Povzetek: Vpeljane so izvirne, na matematični optimizaciji temelječe metrike robustnosti večkriterijskih odločitvenih modelov ter predstavljeni učinkoviti pristopi k analizi občutljivosti in vizualizaciji, ki so bili uspešno uporabljeni na projektih iz področja informacijsko-komunikacijskih tehnologij.

1 Introduction

The decision model is a formal, simplified representation of the problem domain. It transforms input parameters, which are set by the decision-maker, into numerical or qualitative assessments, also called model assumptions (Power, 2002). These assessments should, however, not directly influence the implemented decision; they should rather be further analysed because they are often derived from data that are subject to uncertainty, imprecision and indetermination (Roy, 1996). These phenomena are the consequence of:

- incomplete domain knowledge or information;
- high domain complexity and high cognitive load of the decision-maker;
- insufficient insight into relations between model parameters;
- nonsystematic subjective assessments of criteria weights and evaluations of alternatives.

It is thus necessary to thoroughly and systematically test the inferred model assumptions. Preference aggregation, which is performed in order to assess alternatives, must represent merely the first phase of the decision-making process since the aim of decision analysis is not only to deal with the common problematics of selecting, ranking or classifying alternatives (Roy, 1996), but primarily to provide the decision-maker with a deep understanding of the problem domain, and to clearly expose the influence

of preferential parameters and relations between them on the derived results. For this reason, a technique called the sensitivity analysis is used. It enables the decision-maker to judge in a formal and structured manner (Turban and Aronson, 2001):

- the influence of changes in input data – decision and uncontrollable variables – on the proposed solution that is expressed by the values of output variables;
- the influence of uncertainty on output variables;
- the effects of interactions between variables;
- minimal changes of preferential parameters that are required to obtain (un)desirable results;
- the robustness of both the decision model and the suggested decision in dynamically changing conditions.

Sensitivity/robustness analysis is one of key concepts in the field of multi-criteria decision aiding (Saltelli et al., 1999). It helps the decision-maker to prepare for the uncertain and potentially extreme future, and to improve his understanding of the problem domain by reflecting back on his judgements, synthesising preferences and observing changes. Yet, experiences of researchers and practitioners show that multi-dimensional complexity of the problem domain poses great challenges with regard to the sensitivity analysis as extensive tasks are difficult

to communicate (Hodgkin et al., 2005). On the contrary, visual displays are a powerful means of communication for the majority of people. It is therefore recommended to implement and use interactive visual tools, in order to considerably improve the problem solving process.

Several approaches to sensitivity analysis exist that have been defined in conjunction with various decision-making methods (Frey and Patil, 2002; Vincke, 1999b). Because they are designed for specific types of decision models, they do not cover all relevant aspects of problem solving. Especially the following deficiencies should be taken into consideration:

- Existing L_p -metric based optimization methods and algorithms address sensitivity analysis only partially. They eliminate some dilemmas, but to systematically verify robustness it is necessary to simultaneously measure:
 1. the minimal modification of parameters according to which the best alternative loses its priority over any suboptimal alternative;
 2. the smallest modification that suffices for a selected suboptimal alternative to become the best one;
 3. the largest deviation that preserves the preferential relation of two alternatives.
- In the case of outranking methods ELECTRE and PROMETHEE, the robustness is measured only with regard to criteria weights, aggregated credibility degrees or inferred net flows. Other preferential parameters, such as thresholds, are not analysed.

The purpose of this paper is therefore (1.) to introduce a class of original L_p -metric optimization algorithms and programs that can be applied to holistically measure the robustness of decision models in conjunction with both the utility function and the outranking methods, (2.) to extend the concept of robustness analysis in the context of the ELECTRE TRI method to pseudo-criterion related thresholds, (3.) to formally present fundamental existing sensitivity analysis and visualization techniques that the authors have successfully used within the scope of their project work, and (4.) to discuss the benefits of these techniques. It should be noted that the utility function based approaches are adapted solely to determining the influence of criteria weights. This is a common practice because weight derivation is generally more subjective than specification of criterion-wise values of alternatives.

The rest of the paper is organized as follows. Section 2 provides a brief description of three decision methods – utility function, PROMETHEE II and ELECTRE TRI – to which the techniques of robustness measurement are applied. More detailed explanations can be found in the literature (Figueira et al., 2005). Section 3 gives a review of related work. Section 4 formally presents the stability intervals/regions based automatic sensitivity analysis. In Section 5, several new approaches to multi-dimensional robustness analysis are defined, which utilize (non)linear mathematical programming. This Section represents the original contribution of the paper. In Section 6, practical

examples are provided. They demonstrate the strengths and benefits of the described techniques, and correspond to the results of projects. Section 7 concludes the paper by giving a resume and directions for further research.

2 Theoretical foundations of decision methods

2.1 Multi-attribute utility function

Since the utility theory was axiomatized by Keeney and Raiffa (1993), it has become the most widespread and probably the most relevant approach to decision analysis. Its foundations lay in the dogma of rational behaviour, so it is based on five axioms that provide a framework for a generic strategy that people should adopt when making reasonable decisions. The central concept of all axioms is the lottery, which is a space of outcomes that occur with certain probabilities. If preferences of the decision-maker satisfy these axioms, a real-valued function exists, which is called the utility function and correlates outcomes with a scale that expresses judgements on the $[0, 1]$ interval.

It is uncomplicated to model the utility function for a single attribute (Zeleny, 1982). However, in practice an alternative is generally chosen by expressing preferences on a set of attributes or criteria $\{x_1, \dots, x_n\}$. In this case, the alternative a_i is represented with a vector of values $a_i = (x_1(a_i), \dots, x_n(a_i))$. Its utility is determined by assigning the vector a real value between 0 and 1. It is difficult to directly assess alternatives with the multi-attribute utility function, so this problem is reduced by defining a partial (one-dimensional) utility function for each attribute:

$$u_j(a_i) : x_j(a_i) \rightarrow [0, 1].$$

Partial utilities are aggregated with a decomposition rule. It can have several forms of which the most widely used is the weighted additive decomposition:

$$u(a_i) = \sum_{j=1..n} w_j \cdot u_j(a_i).$$

2.2 PROMETHEE I and II methods

PROMETHEE is a family of methods that are based on the concepts of pseudo-criterion, outranking relation and pairwise comparisons (Brans and Vincke, 1985). For a pair of alternatives a_i and a_j , and for each criterion x_k , the preference function $P_k(a_i, a_j)$ is defined on the interval $[0, 1]$ according to criterion-wise values $g_k(a_i)$ and $g_k(a_j)$, and according to the chosen indifference (q_j), preference (p_j) or Gauss (s_j) thresholds. This function expresses the degree to which a_i outranks (outperforms) a_j . It can have one of six possible shapes of which the linear is the most widely used:

$$P_k(a_i, a_j) = \begin{cases} 0 & , d_k(a_i, a_j) \leq q_k , \\ \frac{d_k(a_i, a_j) - q_k}{p_k - q_k} & , q_k < d_k(a_i, a_j) < p_k , \\ 1 & , d_k(a_i, a_j) \geq p_k , \end{cases}$$

where $d_k(a_i, a_j) = g_k(a_i) - g_k(a_j)$. The outranking degrees are calculated for both “directions”, so that the $P_k(a_i, a_j)$

and $P_k(a_j, a_i)$ values are obtained. Criterion-wise indices are aggregated by taking criteria weights into account:

$$\pi(a_i, a_j) = \sum_{k=1..n} w_k \cdot P_k(a_i, a_j).$$

In the next step, the positive and negative ranking flows $\phi^+(a_i)$ and $\phi^-(a_i)$ are computed for every alternative a_i . They indicate the average degree to which a_i performs better respectively worse than all other alternatives:

$$\begin{aligned} \phi^+(a_i) &= \frac{1}{n-1} \cdot \sum_{a_j \in A} \pi(a_i, a_j) \text{ and} \\ \phi^-(a_i) &= \frac{1}{n-1} \cdot \sum_{a_j \in A} \pi(a_j, a_i). \end{aligned}$$

The inferred flows can be interpreted in two ways. The PROMETHEE I method considers them simultaneously. A partial rank-order of alternatives is thereby derived, in which the incomparability relation may exist in addition to the preference and indifference relations. More often, a weak rank-order is obtained with the PROMETHEE II method. For this purpose, alternatives are evaluated with the net flow:

$$\begin{aligned} \phi(a_i) &= \phi^+(a_i) - \phi^-(a_i) = \\ &= \frac{1}{n-1} \cdot \sum_{a_j \in A} \sum_{k=1..n} w_k \cdot (P_k(a_i, a_j) - P_k(a_j, a_i)). \end{aligned}$$

2.3 Dichotomic ELECTRE TRI method

The above described PROMETHEE I and II methods are designed to rank-order alternatives. Yet, the concepts of pseudo-criterion and outranking relation enable sorting as well. Two variants of PROMETHEE dealing with the sorting problematic have been recently introduced (Araz and Ozkarahan, 2007; Doumpos and Zopounidis, 2004), while the most widespread outranking method for sorting is ELECTRE TRI (Mousseau et al., 2000; Roy, 1991). The latter has been slightly modified within the scope of our research work by following the localization principle and preventing the incomparability relation, in order to allow for group consensus seeking and automated multi-agent negotiation (Bregar et al., 2008).

The dichotomic ELECTRE TRI method compares all alternatives with the profile b . Acceptable choices belong to the positive category C^+ , while unsatisfactory ones are members of the negative category C^- . Let $s_j(a_i, b)$ express the degree to which the option a_i outperforms the profile b according to the criterion x_j . Its calculation is based on the indifference and preference thresholds q_j and p_j :

$$s_j(a_i, b) = \max \left(\min \left(\frac{g_j(a_i) - g_j(b) - q_j}{p_j - q_j}, 1 \right), 0 \right).$$

Analogously, $s_j(b, a_i)$ represents the valued outranking of a_i by b . To express the degree of concordance with the assertion “the alternative a_i belongs to the class C^+ ”, the indices $s_j(a_i, b)$ and $s_j(b, a_i)$ are aggregated with a fuzzy averaging operator:

$$c_j(a_i) = \frac{1}{2} \cdot (s_j(a_i, b) + (1 - s_j(b, a_i))).$$

For the sake of compensation of small weaknesses, the indices $c_j(a_i)$ are combined so that each is scaled by the weight w_j which represents the voting power of the j -th criterion and determines its contribution to the decision:

$$c(a_i) = \sum_{j=1..n} w_j \cdot c_j(a_i).$$

For each criterion, the discordance index is also defined based on the discordance and veto thresholds u_j and v_j . It reflects the partially noncompensatory degree of veto on the assertion “ a_i belongs to C^+ ”:

$$d_j(a_i) = \max \left(\min \left(\frac{g_j(b) - g_j(a_i) - u_j}{v_j - u_j}, 1 \right), 0 \right).$$

The overall nondiscordance relation is grounded in two ways:

$$\tilde{d}'(a_i) = \prod_{j=1..n} (1 - d_j(a_i)) \text{ or}$$

$$\tilde{d}''(a_i) = 1 - d(a_i), \text{ where } d(a_i) = \max_{j=1..n} d_j(a_i).$$

Because of its absolute and noncompensatory nature, the nondiscordance index does not need to be combined with the concordance index. However, the valued outranking relation is usually obtained as a result of the following multiplication:

$$\sigma(a_i) = c(a_i) \cdot \tilde{d}(a_i),$$

$$\text{so that } \tilde{d}(a_i) = \tilde{d}'(a_i) \text{ or } \tilde{d}(a_i) = \tilde{d}''(a_i).$$

As $\sigma(a_i) = 0.5$ denotes strict equality among the profile and the alternative, an appropriate λ -cut should be used to determine the “crisp” membership of the alternative:

$$a_i \in C^+ \Leftrightarrow \sigma(a_i) \geq \lambda, \text{ where } \lambda \in [0.5, 1].$$

3 Existing approaches to sensitivity analysis and visualization

3.1 Techniques and studies

Hites et al. (2006) have explored the applicability of multi-criteria decision-making concepts to the robustness framework by observing the similarities and differences between multi-criteria and robustness problems. In their opinion, a conclusion is called robust if it is true for all or almost all scenarios, where a scenario is a plausible set of parameter values used to solve the problem. In a similar manner, Vincke (1999a) has provided the definition of a robust preference aggregation method. He has analyzed the robustness of eleven methods for the construction of an outranking relation.

Several researchers have investigated the L_p -metric sensitivity analysis of additive multiple attribute value models. Barron and Schmidt (1988) have introduced a procedure for the computation of weights that make the utility of one alternative exceed the utility of a compared alternative by the amount of δ . They have measured the closeness of derived and original weights by the squared deviation. Wolters and Mareschal (1995) have presented a similar method for determining the modification of a given set of weights, which sums up absolute deviations.

In addition to the closeness of weights, Ringuest (1997) has developed a second measure of sensitivity: a decision is considered insensitive if the rank order of weights that led to the original best solution must be altered for the optimal solution to change. A method has been defined which applies both criteria simultaneously by searching for solutions that minimize the L_1 and L_∞ distance metrics subject to a set of linear constraints. Jansen et al. (1997) have described the problems that may occur when using standard software for linear programming. Accordingly, they have proposed a framework for performing efficient sensitivity analysis.

Zopounidis and Doumpos (2002) discuss optimality measures for classification and sorting with respect to the assignment of alternatives in the reference set. Two L_1 -norm distance metrics determine the classification error and the satisfaction of classification rules, respectively. Mousseau et al. (2001) measure the minimal difference α between the credibilities of alternatives and the cutting level that determines to which classes alternatives should be sorted. The larger is the value of α , the more stable are the assignments. Dias et al. (2002) do not approach the measurement of robustness numerically. Instead, their aim is to identify unrobust alternatives that have a wide range of classes to which they may be sorted, since they are strongly affected by the imprecision of data.

Hodgkin et al. (2005) argue that systematic multi-dimensional sensitivity analysis is not well supported by available facilities. Their review of existing techniques for the display of multi-dimensional data reveals many approaches which may be grouped in three categories:

1. approaches that try to retain all information and display it in some manner;
2. reduction of the dimensionality by applying the multi-variate statistical analysis;
3. displays of sensitivity analysis which focus on the outcomes rather than the input data, such as stability intervals, triangles of the weight space, etc.

Hodgkin et al. describe two softwares for the robustness analysis and visual interactive modelling – the triangle plot and the principal components analysis plot. The first reveals three-dimensional stability regions of the weight space, while the latter reduces dimensionality. Both plots have been evaluated from the perspective of end users. The triangle plot is found to be intuitive and easy to use. It exposes robustness and serves as an analytical device with which users can quickly deduce whether the results are as expected. The principal components analysis plot, on the contrary, is rather a heuristic device that exposes comparisons and directs users to further investigations.

3.2 Variance based methods

It has been established that people have difficulties with interpreting and visualizing information in four or more dimensions. An approach that confronts this problem is the principal components analysis (Jolliffe, 2002), which has already been applied in many fields of science for the purpose of reducing dimensionality and providing a good insight into correlations between variables by preserving

a high degree of variance in data. It is often possible to identify a few groups of variables that capture the same key principles, and are hence strongly correlated. Linear combinations of these original variables define a set of principal components forming the unique non-redundant orthogonal basis of a new space of data. Each component corresponds to an axis of the new space. It is selected in such a way that its variance is the highest of all possible choices for this axis. The set of principal components has equal power to the set of original variables, however the sum of variances for only the first two or three principal components generally exceeds 80 percent of variance in original data. For this reason, it is sufficient to consider a small subset of principal components in order to preserve the majority of information. Because of the most simple and understandable interpretation and visualization, the projection on a two-dimensional plane, which is defined by the 1st and the 2nd component, is usually performed.

The principal components analysis may be applied in combination with nearly all multi-criteria decision-aiding methods. Probably the first method that has used it under the name GAIA for almost two decades is PROMETHEE II (Brans and Mareschal, 1994). It takes criteria-wise net ranking flows as the basis for visualization:

$$\phi_k(a_i) = \frac{1}{n-1} \cdot \sum_{a_j \in A} P_k(a_i, a_j) - P_k(a_j, a_i).$$

Espinasse et al. (1997) have applied GAIA planes in a multi-agent negotiation framework. They have developed several levels of group planes, which represent decision-makers, coalitions, criteria and weights with the purpose of assisting the mediator during the negotiation process. Radojević and Petrović (1997) have used GAIA within the scope of fuzzy multi-criteria ranking. They have thus extended the applicability of PROMETHEE methods to the cases when criteria values are fuzzy variables.

Saltelli (2001) has studied the properties of variance based methods in the context of importance assessment. He has considered two settings. In the first, the objective has been to identify the most important factor that would lead to the greatest reduction of variance. In the second, the required target variance has been obtained by fixing simultaneously the smallest possible number of factors.

3.3 Integration in decision support systems

In order to make the process of preference assessment interactive, Mustajoki et al. (2005) have developed and described the WINPRE software, which seeks for three-dimensional stability regions in the weight space, ranges of allowed imprecise weights and partial utility intervals. Another decision support system that visualizes utilities of alternatives in the context of group decision-making is RINGS (Kim and Choi, 2001). By observing overlapping of utility ranges for individual decision-makers and the whole group, consensus can be reached. Moreno-Jimenez et al. (2005) have implemented a spreadsheet module for consensus building, which is able to visualize preference structures with radial graphic representation maps. Each structure is mapped to a planar polygon whose vertices are placed at the end of rays cast from a central point.

Bana e Costa et al. (1999) have integrated several decision support systems which implement visualization and sensitivity analysis techniques. EQUITY provides graphical cost-benefit efficiency analysis, MACBETH depicts value functions, while V.I.S.A. visualizes partial utilities of alternatives and computes stability intervals. Siskos et al. (1999) have embedded visual components into the MIIDAS system. The decision-maker can shape the value function in terms of its curveness and turning point, graphically perform trade-offs, observe the ordinal regression curve and view the net graph coming from the cluster analysis. Jimenez et al. (2003) have introduced a system that allows for imprecise assignments of weights and utilities, whereby inputs can be subjected to different sensitivity analyses and visualization aids, including:

- pie charts of certainties and probabilities;
- bar charts of weights and utilities;
- graphical representations of utility functions;
- stability intervals of weights;
- several types of simulation techniques designed to randomly modify weights by preserving their rank order or numerical intervals.

4 Stability intervals and regions

4.1 Stability intervals

The inference of stability intervals represents the most basic form of sensitivity analysis, next to the “what-if” analysis which is, in connection with interactive graphic tools, used primarily in the phases of criteria structuring and preference elicitation. It is implemented by many decision support systems that help companies and large corporations make important organizational and business decisions (Forman and Selly, 2001). The purpose of this technique is to determine for what intervals of values of a single parameter (for example, a criterion weight), the rank-order of alternatives is preserved. Its main strength is the ability to identify boundaries of stability intervals automatically, without any manual intervention. It is thus appropriate for robustness checking after the preference aggregation phase completes.

To determine the influence of the criterion $x_i \in X$ on the rank-order of alternatives, its weight w_i continuously increases on the interval from 0 to 1. The weights of all other criteria $x_j \in X \setminus \{x_i\}$ decrease inversely proportioned according to their relative portions d_j that exclude w_i :

$$d_j = \frac{w_j}{s_i}, \text{ where } s_i = \sum_{k \neq i}^{k=1..n} w_k.$$

If the normalization of weights is required, such that their sum equals to 1, it becomes clear that the weight of the x_j criterion decreases by $\Delta w_j = d_j \cdot \Delta w_i$ when the weight of the observed criterion x_i is increased by Δw_i . Thereby, the theoretical foundations for the graphical representation of stability intervals are laid. Complementary, the analytical computation of all possible weights w_i at which the rank-order changes is also useful. The utilities of alternatives must be compared in this case for all pairs of a_k and a_l , so that $k, l = 1, \dots, m$ and $k \neq l$. This requires $(m \cdot (m - 1)) / 2$

pairwise comparisons. Since the weighted additive utility function is applied, the point of indifference between two alternatives can be expressed with a linear equation:

$$w_i \cdot u_i(a_k) + \sum_{j \neq i} d_j \cdot (1 - w_i) \cdot u_j(a_k) = w_i \cdot u_i(a_l) + \sum_{j \neq i} d_j \cdot (1 - w_i) \cdot u_j(a_l).$$

The weight w_i is easily derived:

$$\frac{w_i}{1 - w_i} = \frac{\sum_{j \neq i} d_j \cdot (u_j(a_l) - u_j(a_k))}{u_i(a_k) - u_i(a_l)}.$$

Analogously, one-dimensional stability intervals can be found for the PROMETHEE II method, which is based on additive aggregation as well:

$$\frac{w_i}{1 - w_i} = \frac{\sum_{b \in A} \sum_{j \neq i} d_j \cdot (P_j(a_l, b) - P_j(b, a_l) - P_j(a_k, b) + P_j(b, a_k))}{\sum_{b \in A} P_i(a_k, b) - P_i(b, a_k) - P_i(a_l, b) - P_i(b, a_l)}.$$

4.2 Two-dimensional stability regions

It is possible to generalize the stability regions analysis to two or more dimensions. This subsection discusses the interaction of two criteria weights because otherwise the reduction of dimensionality or (non)linear programming must be performed. The latter approach is addressed in the next section. The first is realized by the principal components analysis and is applied by the visual GAIA analysis (Brans and Mareschal, 1994), which projects the multi-dimensional criteria space on a plane, and thereby loses some preferential information.

The two-dimensional sensitivity analysis considers each pair of weights that belong to criteria of the same hierarchical group (let these be the w_i and w_j weights). For a pair of alternatives a_k and a_l , it is determined for which values of w_i and w_j the indifference relation holds. In general, a single point (meaning that alternatives are equivalent for unique weights w_i and w_j), a straight line (implying indifference for an infinite space of weights), or an empty set (meaning that one alternative is preferred to the other for all values of w_i and w_j) is obtained. Lines and points delimit regions within which the rank-order of alternatives remains constant. The stability regions are additionally delimited with borderlines $w_i = 0$, $w_j = 0$ and $w_i + w_j = 1$. It is clear that the new model has one degree of freedom more than the model of stability intervals:

$$w_i \cdot u_i(a_k) + w_j \cdot u_j(a_k) + (1 - w_i - w_j) \cdot \bar{u}(a_k) = w_i \cdot u_i(a_l) + w_j \cdot u_j(a_l) + (1 - w_i - w_j) \cdot \bar{u}(a_l),$$

where $\bar{u}(a_k)$ respectively $\bar{u}(a_l)$ is a constant utility of $n - 2$ criteria that do not change during analysis:

$$\bar{u}(a_k) = \sum_{h \neq i, j} \frac{w_h}{W} \cdot u_h(a_k), \text{ where } W = \sum_{h \neq i, j} w_h.$$

The correlation between weights is now obtained:

$$w_i = \frac{\bar{u}(a_l) - \bar{u}(a_k) - w_j \cdot (u_j(a_k) - u_j(a_l) - \bar{u}(a_k) + \bar{u}(a_l))}{u_i(a_k) - u_i(a_l) - \bar{u}(a_k) + \bar{u}(a_l)}.$$

By setting $w_j = 0$ and $w_j = 1 - w_i$ it can be seen when two alternatives a_k and a_l become equivalent. Analogous two-dimensional sensitivity analysis has been implemented

for the PROMETHEE II method as a functionality of the PROMCALC decision support system.

5 Multi-dimensional robustness analysis

Mathematical programming can be applied to judge the influence of arbitrary many simultaneously changing parameters. The motivation for its use lies in the fact that multi-dimensional information is totally preserved, while in the case of visualisation it gets partially lost because of the projection on a plane. For this reason, several original robustness metrics are proposed. They are implemented with optimization algorithms.

5.1 Optimization approaches for the multi-attribute utility function

The goal of the approaches is to test how robust the rank-order of alternatives is with regard to the weights of all criteria that are structured into a common hierarchical group. Thereby, a comprehensive insight into the model and its robustness must be assured with as few metrics as possible. Four mathematical optimization programs are hence defined. The first exposes the minimal change of the weight vector that causes the best ranked alternative to lose its priority over any other, originally less optimal solution, which means that the best ranked alternative changes. This measurement is of essential importance, since a rational decision is to choose an alternative with the highest utility/value. The robustness of such a choice is obtained with the following program:

$$\Delta_w = \min \frac{\left[\sum_{j=1..n} (w_j - \tilde{w}_j)^P \right]^{1/P}}{\Delta_w^{\max}}$$

subject to

$$\begin{aligned} u(a_k) &= \sum_{j=1..n} w_j \cdot u_j(a_k) \leq \\ &\leq \max_{l \neq k} \left(u(a_l) = \sum_{j=1..n} w_j \cdot u_j(a_l) \right), \\ \sum_{j=1..n} w_j &= 1, \\ 0 \leq w_j \leq 1, \forall j &= 1, \dots, n, \end{aligned}$$

where \tilde{w}_j are current and w_j newly derived weights, and where it holds:

$$\begin{aligned} \tilde{u}(a_k) &= \sum_{j=1..n} \tilde{w}_j \cdot u_j(a_k) = \\ &= \max_{l=1..m} \left(\tilde{u}(a_l) = \sum_{j=1..n} \tilde{w}_j \cdot u_j(a_l) \right). \end{aligned}$$

The parameter P , $1 \leq P \leq \infty$, determines which one of the L_P distance metrics is used. Usually, the Manhattan norm (L_1), which returns the rectangular distance between two vectors, or the Euclidean norm (L_2), which takes the hypotenuse of a square triangle as the distance, are used because of the simplest interpretation. The distance has to be normalized by division with the largest possible change of the weight vector Δ_w^{\max} . For the case when all

criteria weights are allowed to have any value from the $[0, 1]$ interval ($\forall j : dw_j = uw_j - lw_j = 1$), the vector changes maximally when exactly two of its components move from one extreme to the other:

$$\begin{aligned} w_i &= 1, \forall k \neq i : w_k = 0 \rightarrow \\ &\rightarrow w_j = 1, i \neq j, \forall k \neq j : w_k = 0. \end{aligned}$$

In this special situation, Δ_w^{\max} equals to 2. However, for arbitrary differences dw_j , such that $\neg \forall j : dw_j = 1$ holds, the following mathematical program is solved:

$$\Delta_w^{\max} = \max \left[\sum_{j=1..n} (w_j^E - w_j^S)^P \right]^{1/P}$$

by deriving

$$w_j^S, w_j^E, \forall j = 1, \dots, n$$

subject to

$$\begin{aligned} \sum_{j=1..n} w_j^S &= 1, \sum_{j=1..n} w_j^E = 1, \forall j = 1, \dots, n, \\ lw_j \leq w_j^S &\leq uw_j, lw_j \leq w_j^E \leq uw_j, \forall j = 1, \dots, n. \end{aligned}$$

S and E denote the starting respectively ending weights, and also the initial respectively final utilities in the next two programs. To find the largest allowed deviation of the weight vector, such that the preferential relation is preserved for a pair of selected alternatives a_1 and a_2 , the below optimization problem must be dealt with:

$$\text{maximize } \Delta_l(w^S, w^E)$$

subject to

$$\begin{aligned} u^S(a_1) &= \sum_{j=1..n} w_j^S \cdot u_j^S(a_1) \neq \\ &\neq u^S(a_2) = \sum_{j=1..n} w_j^S \cdot u_j^S(a_2), \\ u^E(a_1) &= \sum_{j=1..n} w_j^E \cdot u_j^E(a_1) = \\ &= u^E(a_2) = \sum_{j=1..n} w_j^E \cdot u_j^E(a_2), \\ \sum_{j=1..n} w_j^S &= \sum_{j=1..n} w_j^E = 1, \\ w_j^S &\in [0, 1], w_j^E \in [0, 1], \forall j = 1, \dots, n. \end{aligned}$$

The last addressed problem is to find the smallest change of the weight vector for which any initially suboptimal alternative becomes the best ranked one. As it is similar to the previous optimization problem, the mathematical program is slightly modified:

$$\text{minimize } \Delta_l(w^S, w^E)$$

subject to

$$\begin{aligned} u^S(a_1) &< u^S(a_i), \exists i = 2, \dots, m, \\ u^E(a_1) &> u^E(a_i), \forall i = 2, \dots, m, \\ u^S(a_i) &= \sum_{j=1..n} w_j^S \cdot u_j^S(a_i), \forall i = 1, \dots, m, \\ u^E(a_i) &= \sum_{j=1..n} w_j^E \cdot u_j^E(a_i), \forall i = 1, \dots, m, \\ \sum_{j=1..n} w_j^S &= \sum_{j=1..n} w_j^E = 1, \\ w_j^S &\in [0, 1], w_j^E \in [0, 1], \forall j = 1, \dots, n. \end{aligned}$$

It is presupposed that the alternative selected to become optimal for the final inferred distribution of weights is denoted with a_1 , and that there exists at least one initially superior alternative.

5.2 Optimization approaches for the ELECTRE TRI method

Three types of distance metrics are defined. They reflect the minimum deviations of weight, veto and preference vectors that cause the reassignment of an alternative to the other category. When, considering the alternative a_i , any of these measures is low, the membership of a_i is not sufficiently robust because only a slight modification of preferences may result in a different decision. The most simple task is to find the smallest change of the weight vector so that the reassignment of a_i to the other class occurs: $a_i \in C^+ \rightarrow a_i \in \tilde{C}^-$ or $a_i \in C^- \rightarrow a_i \in \tilde{C}^+$. The problem is solved with a linear optimization program, for which all used symbols have already been defined:

$$\Delta_w(a_i) = \min \left[\frac{\sum_{j=1..n} (w_j - \tilde{w}_j)^P}{\Delta_w^{\max}} \right]^{1/P}$$

by deriving

$$w_j, \forall j = 1, \dots, n$$

subject to

$$\sigma(a_i) = d(a_i) \cdot \left(\sum_{j=1..n} w_j \cdot c_j(a_i) \right) = \lambda,$$

$$\sum_{j=1..n} w_j = 1, lw_j \leq w_j \leq uw_j, \forall j = 1, \dots, n.$$

A harder problem is to measure the robustness of veto and discordance thresholds v_j and u_j . An advanced metric is needed that allows for the aggregation of discordance indices, and indicates the minimal threshold deviations that would cause the observed alternative to reassign:

$$\Delta_v(a_i) = \min \left[\frac{\sum_{j=1..n} (\delta_j)^P}{\sum_{j=1..n} (2 \cdot (g_j(b) - p_j - D_j^-))^P} \right]^{1/P}$$

by deriving

$$u_j \text{ and } v_j, \forall j = 1, \dots, n$$

subject to

$$\sigma(a_i) = c(a_i) \cdot \prod_{j=1..n} (1 - d_j(a_i)) = \lambda,$$

$$d_j(a_i) = \max \left(\min \left(\frac{g_j(b) - g_j(a_i) - u_j}{v_j - u_j}, 1 \right), 0 \right),$$

$$\delta_j = |u_j - \tilde{u}_j| + |v_j - \tilde{v}_j| + \left| (v_j - u_j) - (\tilde{v}_j - \tilde{u}_j) \right|, \forall j = 1, \dots, n,$$

$$p_j \leq u_j \leq v_j \leq b_j, \forall j = 1, \dots, n.$$

The program minimizes the distances between previous and new values of discordance and veto thresholds. In addition, it pays regard to the distances between different

thresholds ($|v_j - u_j|$), to prevent anomalies that can occur if thresholds converge towards the same value. It clearly demonstrates the problematic of finding the smallest change of u_j and v_j thresholds that causes reclassification. Yet, it has to deal with piecewise linear functions with unknown segments. For this reason, it is substituted with a different optimization program. For each value $g_j(a_i)$, an appropriate partial discordance degree is found so that the product of these degrees equals the required overall discordance $\tilde{d}(a_i)$ calculated by dividing the fixed cut level λ with the fixed concordance index $c(a_i)$. Then, the criterion-wise coefficient k_j of a linear function is derived according to $g_j(a_i)$ (x -axis) and $\tilde{d}_j(a_i)$ (y -axis), for each index $j = 1, \dots, n$. The induced function determines the u_j and v_j thresholds (at $y = 0$ and $y = 1$), and minimizes the distance metric:

$$\Delta_v(a_i) = \min \left[\frac{\sum_{j \in F} (\delta_j)^P}{\sum_{j \in E} (2 \cdot (b_j - p_j - D_j^-))^P} \right]^{1/P}$$

by deriving

$$\tilde{d}_j(a_i) \text{ and } k_j, \forall j \in F$$

subject to

$$E = \{1, \dots, n\}, F \subseteq E,$$

$$\prod_{j \in F} (1 - \tilde{d}_j(a_i)) \cdot \prod_{j \in E \setminus F} (1 - d_j(a_i)) = \tilde{d}(a_i),$$

$$0 \leq \tilde{d}_j(a_i) \leq 1, \forall j \in F,$$

$$\delta_j = \delta_j^u + \delta_j^v + \delta_j^{uv}, \forall j \in F,$$

$$\delta_j^u = u_j - g_j(a_i) + \frac{\tilde{d}_j(a_i)}{k_j}, \forall j \in F,$$

$$\delta_j^v = v_j - g_j(a_i) - \frac{1 - \tilde{d}_j(a_i)}{k_j}, \forall j \in F,$$

$$\delta_j^{uv} = v_j - u_j - \frac{1}{k_j}, \forall j \in F,$$

$$\frac{1 - \tilde{d}_j(a_i)}{D_j^+ - D_j^- - g_j(a_i)} \leq k_j \leq \frac{\tilde{d}_j(a_i)}{g_j(a_i) - p_j}, \forall j \in F.$$

Figure 1 gives the graphical interpretation on how the new u_j and v_j thresholds are inferred by inducing the k_j coefficient. The thresholds may be modified either with a parallel shift of the function or by changing its slope with the increase/decrease of the k_j coefficient. Consequently, their absolute difference or the initial value of u_j must be preserved. The third possibility also exists: by combining the shift and the angle adjustment, all differences Δu , Δv and Δuv become positive.

On Figure 1, k_0 and k_1 depict the initial respectively the extreme possible induced angle of the linear function. Similarly, y_0 denotes the initial partial discordance degree and y_1 represents the required adjusted degree. Finally, x_0 corresponds to the criterion-wise value of the alternative $g_j(a_i)$. If a_i is the member of the positive category C^+ , the discordance degree must increase in order to cause the

reassignment, which is a prerequisite to properly measure robustness. Then, $y_1 > y_0$; otherwise $y_1 < y_0$.

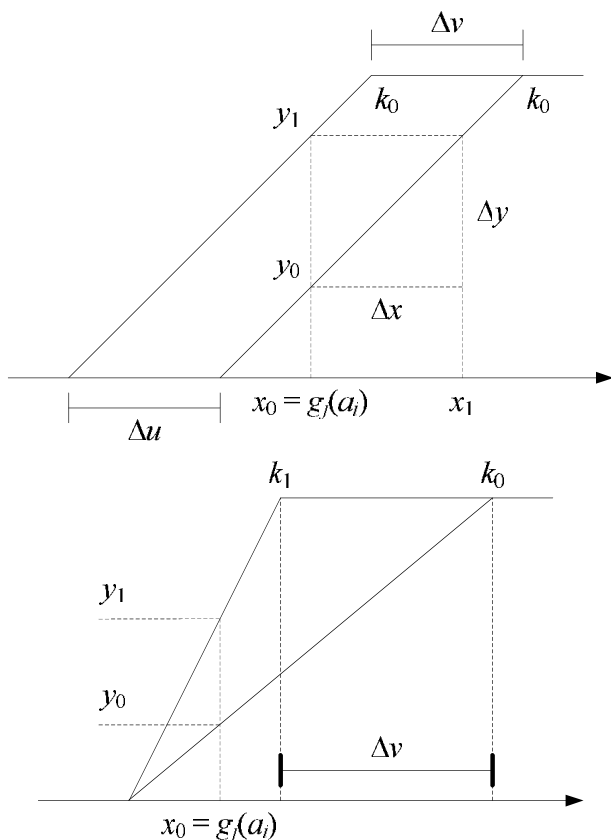


Figure 1: Inference of discordance and veto thresholds with the parallel shift and with the slope adjustment.

The problem of finding the deviations of indifference and preference thresholds that would cause the classification of an alternative into a different category is very similar to the one described above. The optimization is slightly more demanding because it has to deal with symmetry of partial concordance indices. This difficulty is overcome by multiplying each newly derived index with a sign that is determined by comparing the $g_j(a_i)$ and $g_j(b)$ values.

6 Practical examples

All examples described in this Section are based on the utility theory. Partial utility functions are not presented as it is not necessary to be acquainted with them in order to comprehend the discussed use of robustness techniques. Partial utilities are aggregated with the weighted additive decomposition rule, which is defined in Subsection 2.1. Methodological details on the optimization programs and on the computation of stability regions are omitted, since they are thoroughly introduced in Sections 4 and 5. In their original forms, all decision models are extensive. Hence, a subset of the most relevant criteria is treated for the demonstrative purposes. Similarly, the application of robustness algorithms for the ELECTRE TRI method requires a complex example that exceeds the scope of the paper. It can be found in the literature (Bregar, 2009).

Figure 2 shows two examples of stability regions. In the first case, the decision is robust because a substantial

modification of the observed weights w_1 and w_2 is needed for the alternative a_1 to gain a higher utility than the best ranked alternative a_2 . On the contrary, the decision is not robust in the second example. A small change of current weights suffices for a_1 to be selected as the best available option instead of a_3 . In this way, a thorough insight into the decision model is provided in addition to the derived rank-order and assessments of alternatives. The examples are based on the analysis which has been performed for the purpose of toll systems evaluation (Jurič et al., 2005). Since project data are not public, alternatives and criteria are not explicitly named.

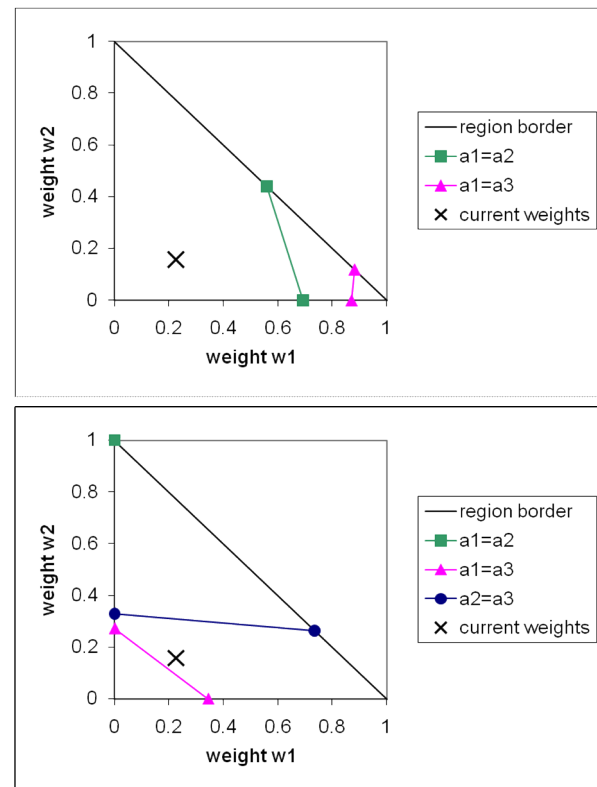


Figure 2: Examples of stability regions.

In order to measure robustness with regard to arbitrary many criteria, mathematical programming has been used for the purpose of above described evaluation, as well as to select the best service-oriented architecture. Because this paper focuses on the formal definition of several new and several established decision analysis techniques, and not on the assessment of service-oriented architectures, any prioritization of the latter is avoided. Hence, the evaluated BEA WebLogic/AquaLogic, IBM SOA, JBoss, Microsoft SOA and Oracle SOA Suite architectures are simply denoted with symbols a_1 to a_5 , so that the order is randomly mixed. Although over 100 criteria have been specified, only five are considered here:

- x_1 – service-oriented architecture (global goal),
- x_2 – functionality,
- x_3 – support for business rules,
- x_4 – administrative tools,
- x_5 – business intelligence.

In this example, the criteria x_1 to x_5 are not dealt with in a hierarchically structured manner, yet in practice, x_2 is a

subcriterion/descendant of x_1 and x_3 to x_5 are descendants of x_2 . To clearly demonstrate the strengths and benefits of the proposed class of robustness analysis techniques, a mathematical optimization program is applied to solve the problem of finding the minimal required modification of the weight vector, such that the best ranked alternative changes. This is the first program from Subsection 5.1. It is operationalized to measure the Euclidean distance and to allow all weights to be between 0 and 1. The obtained results are organized in Table 1.

Table 1: Utilities of alternatives and robustness degrees.

Alternative	x_1	x_2	x_3	x_4	x_5
a_1	0.85	0.89	0.55	0.79	0.82
a_2	0.65	0.72	0.60	0.72	0.89
a_3	0.63	0.69	0.30	0.73	0.55
a_4	0.55	0.78	0.42	0.69	0.78
a_5	0.55	0.42	0.61	0.38	0.25
Robustness	0.40	0.62	1.00	0.91	0.15

For each alternative, its criteria-wise utilities are written. The last line contains the measured robustness degrees, which represent the distance between the original and the derived weight vector. The minimal possible robustness degree is 0, while the maximal is 1. It can be observed that these degrees provide far richer information than the computed utilities:

- According to criteria x_2 and x_4 , there is almost no difference between the best and the second best alternative. The increase in utility is 0.11 and 0.06, respectively, on the scale from 0 to 1. This does not suffice for the decision-maker to be confident in the proposed decision. However, the degree of robustness is very high (0.62 and 0.91), which means that preferences are firmly stated. Consequently, the reliability of the model drastically improves.
- According to the third criterion, the best and the second best alternative are almost indifferent, as their utilities are 0.61 and 0.60, respectively. It is hence virtually impossible for the decision-maker to rationally choose between them solely on the basis of utilities. However, the robustness index has the highest value of 1, which means that no combination of weights can be found to change the preferential relation $a_5 P a_2$. In this way, it becomes obvious that a_5 represents the only reasonable solution.
- With regard to x_5 , the robustness degree gives a conformation to the fact that the decision-maker should be extremely cautious when choosing a_2 over a_1 or a_4 . This should be a clear sign for him to properly revise the decision model.

In the cases when both the difference in utilities of two best ranked alternatives and the degree of robustness are moderate, the proposed technique may be useful as well. Table 2 shows how the weights of subcriteria should be adjusted in order to change the best ranked alternative with respect to the criterion x_1 . The weight of the costs subcriterion increases to such an extent (from 0.28 to 0.60) that the derived value is unacceptable.

Table 2: Required adjustments of the weight vector.

Criteria	Original weights	Derived weights
Functionality	0.32	0.15
Impact on investments	0.40	0.25
Costs	0.28	0.60

Figure 3 depicts the results of the principal components analysis for the fictitious case of selecting an Eastern European country for cooperation on a multilateral ICT project. Criteria are shown as vectors and alternatives as points. It can be clearly seen which alternatives perform well with respect to which criteria. The GAIA analysis additionally includes the so called decision stick on the plane. It is obtained by projecting the weight vector onto the two-dimensional coordinate system, and points in the direction of the best possible alternative.

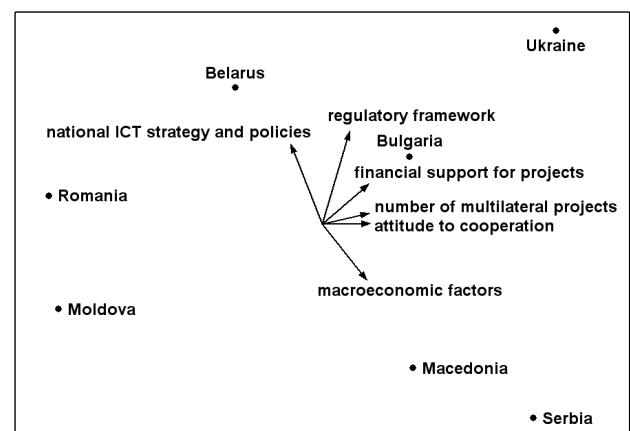


Figure 3: Visualization on the basis of principal components analysis.

Criteria and alternatives (countries) are adopted from the GREAT-IST questionnaire based survey (Györkös et al., 2006), however both the scope of the decision and the data are deliberately as well as significantly modified for the purpose of this example. Randomly generated data in the form of utilities are presented in Table 3. With x_1 to x_6 , the following criteria are denoted:

- x_1 – number of multilateral projects,
- x_2 – attitude to international cooperation,
- x_3 – financial support for projects,
- x_4 – national ICT strategy and policies,
- x_5 – regulatory framework,
- x_6 – macroeconomic factors.

Table 3: Fictitious randomly sampled utilities.

Country	x_1	x_2	x_3	x_4	x_5	x_6
Belarus	0.4	0.3	0.6	0.8	0.7	0.3
Bulgaria	0.5	0.5	0.7	0.4	0.9	0.6
Macedonia	0.7	0.6	0.6	0.5	0.5	0.9
Moldova	0.2	0.1	0.2	0.7	0.4	0.2
Romania	0.2	0.1	0.3	1.0	0.4	0.2
Serbia	0.8	0.9	0.8	0.3	0.5	1.0
Ukraine	1.0	1.0	1.0	0.8	0.8	0.7

As is evident from Table 4, most preferential information are preserved on the two-dimensional plane. Nearly 90

percent of cumulative variance is covered by the first two principal components.

Table 4: Variance of principal components.

Principal component	Percentage of variance	Cumulative variance
1	70.65 %	70.65 %
2	17.65 %	88.30 %
3	10.64 %	98.94 %
4	0.74 %	99.68 %
5	0.27 %	99.95 %
6	0.06 %	100.00 %

7 Conclusion

Robustness analysis and visualization provide for several benefits. They:

1. help the decision-maker in achieving flexibility and adaptability to quickly changing conditions and characteristics of the observed situation or domain;
2. enable better understanding of the problem dealt with and the decision suggested/made;
3. increase confidence in the decision model, which can be gained through the structured process of subjectively expressing preferential information.

Therefore, several techniques for measuring robustness and for visualizing multiple criteria decision models of various types have been defined. Most of them represent novel approaches to sensitivity analysis, while some are already established, but have been successfully applied on projects. Additional algorithms will be introduced in the scope of future research work, in order to determine:

- for what convex polyhedron of parameter values the observed alternative is selected as the best one, identified as the only acceptable choice, or classified/sorted into the appropriate category;
- for what convex intersections of polyhedrons available alternatives become indifferent or get classified/sorted into the same category.

References

- [1] Araz, C., Ozkarahan, I. (2007). Supplier evaluation and management system for strategic sourcing based on new multicriteria sorting procedure. *International Journal of Production Economics*, vol. 106, no. 2, pp. 585–606.
- [2] Bana e Costa, C. A., Ensslin, L., Correa, E. C., Vansnick, J.-C. (1999). Decision support systems in action: Integrated application in a multicriteria decision aid process. *European Journal of Operational Research*, vol. 113, no. 2, pp. 315–335.
- [3] Barron, H., Schmidt, C. P. (1988). Sensitivity analysis of additive multiattribute value models. *Operations Research*, vol. 36, no. 1, pp. 122–127.
- [4] Brans, J., Mareschal, B. (1994). The PROMCALC & GAIA decision support system for multicriteria decision aid. *Decision Support Systems*, vol. 12, no. 4–5, pp. 297–310.
- [5] Brans, J., Vincke, Ph. (1985). A preference ranking organisation method: The PROMETHEE method for multi-criteria decision-making. *Management Science*, vol. 31, no. 6, pp. 647–656.
- [6] Bregar, A. (2009). *An Aggregation/Disaggregation Approach to Integrative Negotiations*. PhD thesis. University of Maribor, Faculty of Electrical Engineering and Computer Science. URLs: <http://dkum.uni-mb.si/Dokument.php?id=7309> and <http://slais.ijs.si/theses/2009-02-02-Bregar.pdf>.
- [7] Bregar, A., Györkös, J., Jurič, M. B. (2008). Interactive aggregation/disaggregation dichotomic sorting procedure for group decision analysis based on the threshold model. *Informatica (Vilnius)*, vol. 19, no. 2, pp. 161–190.
- [8] Dias, L., Mousseau, V., Figueira, J., Climaco, J. (2002). An aggregation/disaggregation approach to obtain robust conclusions with ELECTRE TRI. *European Journal of Operational Research*, vol. 138, no. 2, pp. 332–348.
- [9] Doumpos, M., Zopounidis, C. (2004). Multicriteria classification approach based on pairwise comparisons. *European Journal of Operational Research*, vol. 158, no. 2, pp. 378–389.
- [10] Espinasse, B., Picolet, G., Chouraqui, E. (1997). Negotiation support systems: A multi-criteria and multi-agent approach. *European Journal of Operational Research*, vol. 103, no. 2, pp. 389–409.
- [11] Figueira, J., Greco, S., Ehrgott, M. (2005). *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer.
- [12] Forman, E. H., Selly, M. A. (2001). *Decision by Objectives*. World Scientific.
- [13] Frey, H. C., Patil, S. M. (2002). Identification and review of sensitivity analysis methods. *Risk Analysis*, vol. 22, no. 3, pp. 553–578.
- [14] Györkös, J., Bregar, A., Cizelj, B. (2006). *Report on GREAT-IST WG3: Questionnaire analysis*. University of Maribor, Faculty of Electrical Engineering and Computer Science.
- [15] Hites, R., De Smet, Y., Risse, N., Salazar-Neumann, M., Vincke, Ph. (2006). About the applicability of MCDA to some robustness problems. *European Journal of Operational Research*, vol. 174, no. 1, pp. 322–332.
- [16] Hodgkin, J., Belton, V., Koulouri, A. (2005). Supporting the intelligent MCDA user: A case study in multi-person multi-criteria decision support. *European Journal of Operational Research*, vol. 160, no. 1, pp. 172–189.
- [17] Jansen, B., De Jong, J. J., Roos, C., Terlaky, T. (1997). Sensitivity analysis in linear programming: just be careful! *European Journal of Operational Research*, vol. 101, no. 1, pp. 15–28.
- [18] Jimenez, A., Rios-Insua, S., Mateos, A. (2003). A decision support system for multiattribute utility evaluation based on imprecise assignments. *Decision Support Systems*, vol. 36, no. 1, pp. 65–79.
- [19] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

- [20] Jurič, M. B., Bregar, A., Rozman, I. (2005). *Evalvacija sistemov cestninjenja: Odločitveni model cestninjenja z vrednotenjem*. University of Maribor, Faculty of Electrical Engineering and Computer Science.
- [21] Keeney, R. L., Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press.
- [22] Kim, J. K., Choi, S. H. (2001). A utility range-based interactive group support system for multiattribute decision making. *Computers & Operations Research*, vol. 28, no. 5, pp. 485–503.
- [23] Moreno-Jimenez, J. M., Joven, J. A., Pirla, A. R., Lanuza, A. T. (2005). A spreadsheet module for consistent consensus building in AHP-group decision making. *Group Decision and Negotiation*, vol. 14, no. 2, pp. 89–108.
- [24] Mousseau, V., Figueira, J., Naux, J.-Ph. (2001). Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. *European Journal of Operational Research*, vol. 130, no. 2, pp. 263–275.
- [25] Mousseau, V., Slowinski, R., Zielniewicz, P. (2000). A user-oriented implementation of the ELECTRE TRI method integrating preference elicitation support. *Computers & Operations Research*, vol. 27, no. 7, pp. 757–777.
- [26] Mustajoki, J., Hämäläinen, R. P., Salo, A. (2005). Decision support by interval SMART/SWING – Incorporating imprecision in the SMART and SWING methods. *Decision Sciences*, vol. 36, no. 2, pp. 317–339.
- [27] Power D. J. (2002). *Decision Support Systems: Concepts and Resources for Managers*. Greenwood Publishing.
- [28] Radojević, D., Petrović, S. (1997). A fuzzy approach to preference structure in multicriteria ranking. *International Transactions in Operational Research*, vol. 4, no. 5–6, pp. 419–430.
- [29] Ringuest, J. L. (1997). L_p -metric sensitivity analysis for single and multi-attribute decision analysis. *European Journal of Operational Research*, vol. 98, no. 3, pp. 563–570.
- [30] Roy, B. (1991). The outranking approach and the foundation of ELECTRE methods. *Theory and Decision*, vol. 31, no. 1, pp. 49–73.
- [31] Roy, B. (1996). *Multicriteria Methodology for Decision Aiding*. Kluwer Academic Publishers.
- [32] Saltelli, A. (2001). Sensitivity analysis for importance assessment. *Proceedings of the 3rd International Symposium on Sensitivity Analysis of Model Output*, pp. 3–18.
- [33] Saltelli, A., Tarantola, S., Chan, K. (1999). A role for sensitivity analysis in presenting the results from MCDA studies to decision makers. *Journal of Multi-Criteria Decision Analysis*, vol. 8, no. 3, pp. 139–145.
- [34] Siskos, Y., Spyridakos, A., Yannacopoulos, D. (1999). Using artificial intelligence and visual techniques into preference disaggregation analysis: The MIIDAS system. *European Journal of Operational Research*, vol. 113, no. 2, pp. 281–299.
- [35] Turban, E., Aronson, J. E. (2001). *Decision Support Systems and Intelligent Systems*. Prentice-Hall.
- [36] Vincke, Ph. (1999). Robust and neutral methods for aggregating preferences into an outranking relation. *European Journal of Operational Research*, vol. 112, no. 2, pp. 405–412.
- [37] Vincke, Ph. (1999). Robust solutions and methods in decision aid. *Journal of Multi-Criteria Decision Analysis*, vol. 8, no. 3, pp. 181–187.
- [38] Wolters, W. T. M., Mareschal, B. (1995). Novel types of sensitivity analysis for additive MCDM methods. *European Journal of Operational Research*, vol. 81, no. 2, pp. 281–290.
- [39] Zeleny, M. (1982). *Multiple Criteria Decision Making*. McGraw-Hill.
- [40] Zopounidis, C., Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, vol. 138, no. 2, pp. 229–246.

Run-time Manipulation of Programs in a Statically-Typed Language

Sašo Greiner

University of Maribor, Faculty of Electrical Engineering and Computer Science,
Smetanova 17, 2000 Maribor, Slovenia

E-mail: saso.greiner@uni-mb.si,

<http://labraj.uni-mb.si/disertacija.pdf>

Thesis Summary

Keywords: programming languages, metaprogramming, reflection

Received: March 31, 2009

This article is an extended abstract of a doctoral dissertation on metaprogramming and programming language design. A metaprogramming model is studied and implemented in a statically-typed pure object-oriented programming language Zero. The object model of language is based on closures which enables metaprogramming model to achieve a high degree of dynamic manipulation, normally only found in dynamically-typed languages. Metaprogramming in Zero is based on safely-typed structural and behavioural reflection.

Povzetek: Prispevek predstavlja doktorsko disertacijo s področja načrtovanja programskih jezikov in metaprogramiranja.

1 Introduction

Metaprogramming [6] is a key programming language feature in implementation of today's rapidly growing enterprise software systems. Metaprogramming allows manipulation of program behaviour and structure during program execution. This is vital for software that requires a high degree of availability and scalability. The concept is usually found in dynamic languages, such as Smalltalk [3], Lisp [7, 1], and Self [8]. In the dissertation we designed a metaprogramming model applicable for a statically-typed language. We developed the language Zero [4] which is a pure object-oriented programming language allowing structural and behavioural manipulation of programs at run-time.

2 The language Zero

The language Zero is built on top of language Z_0 [5]. Zero is a statically-typed language which makes execution more efficient than with dynamic languages and less fallible in terms of typing. Zero enables a high degree of application manipulation at run-time as it supports both behavioural and structural metaprogramming. In other words, changing the functionality of an application may be addressed by modifying its behaviour and structure when application is already running. The metaprogramming model of language Zero is based on pure object-orientation. That is, all values including control structures and methods in a program, are objects. Such pure object-oriented model enables efficient implementation of the metaprogramming

core in the language. The most important aspect of such a representation is that all dynamic changes can be type-checked at run-time. This is vital as maintaining a program in a type-safe state is mandatory for statically-typed languages. The metaprogramming model of Zero is based on metaclasses. Metaclasses provide introspective features, such as obtaining information about classes, methods, and parameter types, as well as dynamic features for changing structural and behavioural properties of a running program. There are 3 main metaclasses in Zero: `Class`, `Method`, and `Closure`. Metaclass `Class` represents run-time class objects. A run-time method is represented by metaclass `Method`. Methods themselves are based on metaclass `Closure` which serves as the fundamental class of all control structures.

The metaprogramming model allows inspection of running programs as well as their manipulation. The latter includes decomposition of existing functionality and construction of a new one. By allowing changes to programs at run-time it becomes unnecessary for the programs to shut down and recompile. Behavioural reflection in Zero is realised by handlers, which are in fact method objects that may be attached to closures. Handlers resemble aspect-oriented programming (AOP) [2], where attached method objects may be viewed as advices. A join-point, a spot where program behaviour may be extended, is always a closure in Zero. We demonstrate practical cases where manipulating program structure and behaviour may be used to achieve that running programs meet the new requirements. Often enough, certain parameters only become available at run-time. The metaprogramming model of Zero allows such programs to be dynamically restructured and reconfigured

taking these new parameters into account.

The Zero metaprogramming model works on instance and class levels. This means program structure and behaviour may address only a particular instance or a class and consequently all instances of this class.

Metaprogramming in Zero is used for fine-grain manipulation as well as for modifying large structures. Fine-grain manipulation works with closures which are basic building blocks of control structures, such as loops and selection statements, and method bodies. Modifying large structures such as replacing entire methods and superclasses is based on signature compatibility.

3 Conclusion

We have designed and developed a statically-typed object-oriented programming language Zero which allows dynamic changes of program structure and behaviour. Run-time changes of running programs are addressed with a metaprogramming model based on metaclasses which ensure that changes applied do not cause typing errors. Safe method and class replacements are based on signature compatibility. The metaprogramming model of language Zero allows fine-grain tuning of programs with the use of closures as building blocks of methods and all control structures. A more rigid tuning is achieved by replacing entire methods and classes in class hierarchies. As closures are basic blocks of programs, all parts of a program may be modified either by changing the behaviour or their entire structure.

Acknowledgement

The work on dissertation was supervised by prof. dr. Janez Brest and prof. dr. Viljem Žumer.

References

- [1] Stanley Jefferson and Daniel P. Friedman. A simple reflective interpreter. *Lisp and Symbolic Computation*, 9(2-3):181–202, 1996.
- [2] Gregor Kiczales, Erik Hilsdale, Jim Hugunin, Mik Kersten, Jeffrey Palm, and William G. Griswold. An overview of AspectJ. *Lecture Notes in Computer Science*, 2072:327–355, 2001.
- [3] Wilf R. LaLonde and John R. Pugh. *Inside Smalltalk Volume I*. Prentice-Hall International, Inc., 1990.
- [4] Sašo Greiner, Viljem Žumer, Janez Brest. Zero – a blend of static typing and dynamic metaprogramming. *Comput. Lang. Syst. Struct.*, 35(3):241–251, 2009.
- [5] Sašo Greiner, Damijan Rebernak, Janez Brest, and Viljem Žumer. Z_0 – a tiny experimental language. *SIG-PLAN Not.*, 40(8):19–28, 2005.
- [6] Diomidis Spinellis. Rational metaprogramming. *IEEE Softw.*, 25(1):78–79, 2008.
- [7] Guy Steele. Common lisp: The language. *Digital Equipment Corporation*, 1984.
- [8] David Ungar and Randall B. Smith. Self: The power of simplicity. *OOPSLA'87*, 4(8):227–242, 1987.

First Call for Papers: IFIP AI 2010

IFIP AI 2010: The Third IFIP International Conference on Artificial Intelligence in Theory and Practice

Brisbane, Australia, September 20th-23rd 2010

<http://www.ifiptc12.org/ifipai2010>

IFIP AI 2010 is one of the constituent parts of the IFIP World Computer Congress. The conference will follow the same format as the highly successful IFIP AI 2008 at the IFIP WCC 2008 in Milan, Italy. For details see <http://www.wcc2010.org>.

Papers describing original work are invited in the areas listed below or other areas of Artificial Intelligence for oral presentation at the conference. Acceptance will be based on quality, relevance and originality and the practical value of the work.

Papers of no more than 10 pages should be submitted following the publisher's prescribed format (for details see the conference web site). Final versions may be submitted in either Word or Latex format but at this stage all papers should be submitted as PDF files. The proceedings will be published by Springer in the IFIP AICT (Advances in Information and Communication Technology) Series. Papers should be uploaded via the conference website at <http://www.ifiptc12.org/ifipai2010> from which further information is available.

Important dates

Submission deadline: January 31st, 2010

Notification of acceptance: April 23rd, 2010

Camera-ready deadline: May 15th, 2010

Conference: September 20th-23rd, 2010

Conference chair

Tharam Dillon, Curtin University, Australia
tharam.dillon@cbs.curtin.edu.au

Program committee chair

Max Bramer, University of Portsmouth, United Kingdom
max.bramer@port.ac.uk

Program committee members

A full list of members is available on the conference website.

Topics:

- Industrial Applications of Artificial Intelligence
- Intelligent Decision Support Systems
- Integration of AI with other Technologies
- Evaluation of AI Systems
- AI Languages, Programming Techniques and Tools
- Knowledge Acquisition
- Expert and Knowledge-based Systems
- Fuzzy Logic and Plausible Inference
- Neural Networks
- Evolutionary Computation and Algorithms
- Speech and Natural Language Interfaces
- Machine Vision
- Intelligent Information Retrieval
- Genetic Algorithms
- Planning and Scheduling
- Bayesian Networks and Stochastic Reasoning
- Evolutionary Programming
- Learning and Adaptive Systems
- Intelligent Agents
- Distributed AI Algorithms, Techniques and Applications
- Distributed AI Systems and Architectures
- Intelligent Tutoring Systems
- Structured and Unstructured Data Mining
- Case-Based Reasoning Systems
- Hardware and Robotics
- Social Impact, Acceptance and Implications of AI
- Intelligent Systems Engineering and Design Methodologies
- Ontologies and Semantic Web
- Knowledge and Information Management
- Business Process Management and Enterprise Portals
- Organisational Memory Knowledge Systems
- Inter-organisational KM portals
- Knowledge Management, E-Learning and Enterprise Portals

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan-Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 800 staff, has 600 researchers, about 250 of whom are postgraduates, nearly 400 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of Slovenia (or S^onia). The capital today is considered a crossroad between East, West and Mediter-

anean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.: +386 1 4773 900, Fax.: +386 1 251 93 85
WWW: <http://www.ijs.si>
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA
AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS
INVITATION, COOPERATION

Submissions and Refereeing

Please submit an email with the manuscript to one of the editors from the Editorial Board or to the Managing Editor. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica L^AT_EX format and figures in .eps format. Style and examples of papers can be obtained from <http://www.informatica.si>. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

- Send Informatica free of charge
- Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than sixteen years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering the European computer science and informatics community - scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name:	Office Address and Telephone (optional):
Title and Profession (optional):
.....	E-mail Address (optional):
Home Address and Telephone (optional):
.....	Signature and Date:

Informatica WWW:

<http://www.informatica.si/>

Referees:

Witold Abramowicz, David Abramson, Adel Adi, Kenneth Aizawa, Suad Alagić, Mohamad Alam, Dia Ali, Alan Aliu, Richard Amoroso, John Anderson, Hans-Jurgen Appelrath, Iván Araujo, Vladimir Bajič, Michel Barbeau, Grzegorz Bartoszewicz, Catriel Beerli, Daniel Beech, Fevzi Belli, Simon Beloglavec, Sondes Bennisri, Francesco Bergadano, Istvan Berkeley, Azer Bestavros, Andraž Bežek, Balaji Bharadwaj, Ralph Bisland, Jacek Blazewicz, Laszlo Boeszoermenyi, Damjan Bojadžijev, Jeff Bone, Ivan Bratko, Pavel Brazdil, Bostjan Brumen, Jerzy Brzezinski, Marian Bubak, Davide Bugali, Troy Bull, Sabin Corneliu Buraga, Leslie Burkholder, Frada Burstein, Wojciech Buszkowski, Rajkumar Bvyya, Giacomo Cabri, Netiva Caftori, Patricia Carando, Robert Cattral, Jason Ceddia, Ryszard Choras, Wojciech Cellary, Wojciech Chybowski, Andrzej Ciepiewski, Vic Ciesielski, Mel Ó Cinnéide, David Cliff, Maria Cobb, Jean-Pierre Corriveau, Travis Craig, Noel Craske, Matthew Crocker, Tadeusz Czachorski, Milan Češka, Honghua Dai, Bart de Decker, Deborah Dent, Andrej Dobnikar, Sait Dogru, Peter Dolog, Georg Dorfner, Ludoslaw Drellichowski, Matija Drobnič, Maciej Drozdowski, Marek Druzdzel, Marjan Družovec, Jozo Dujmović, Pavol Ďuriš, Amnon Eden, Johann Eder, Hesham El-Rewini, Darrell Ferguson, Warren Fergusson, David Flater, Pierre Flener, Wojciech Fliegner, Vladimir A. Fomichov, Terrence Forgarty, Hans Fraaije, Stan Franklin, Violetta Galant, Hugo de Garis, Eugeniusz Gatnar, Grant Gayed, James Geller, Michael Georgiopolus, Michael Gertz, Jan Goliński, Janusz Gorski, Georg Gottlob, David Green, Herbert Groiss, Jozsef Gyorkos, Marten Haglind, Abdelwahab Hamou-Lhadj, Inman Harvey, Jaak Henno, Marjan Hericko, Henry Hexmoor, Elke Hochmueller, Jack Hodges, John-Paul Hosom, Doug Howe, Rod Howell, Tomáš Hruška, Don Huch, Simone Fischer-Huebner, Zbigniew Huzar, Alexey Ippa, Hannu Jaakkola, Sushil Jajodia, Ryszard Jakubowski, Piotr Jedrzejowicz, A. Milton Jenkins, Eric Johnson, Polina Jordanova, Djani Juričič, Marko Juvancic, Sabhash Kak, Li-Shan Kang, Ivan Kapustok, Orlando Karam, Roland Kaschek, Jacek Kierzenka, Jan Kniat, Stavros Kokkotos, Fabio Kon, Kevin Korb, Gilad Koren, Andrej Krajnc, Henryk Krawczyk, Ben Kroese, Zbyszko Krolikowski, Benjamin Kuipers, Matjaž Kukar, Aarre Laakso, Sofiane Labidi, Les Labuschagne, Ivan Lah, Phil Laplante, Bud Lawson, Herbert Leitold, Ulrike Leopold-Wildburger, Timothy C. Lethbridge, Joseph Y-T. Leung, Barry Levine, Xuefeng Li, Alexander Linkevich, Raymond Lister, Doug Locke, Peter Lockeman, Vincenzo Loia, Matija Lokar, Jason Lowder, Kim Teng Lua, Ann Macintosh, Bernardo Magnini, Andrzej Małachowski, Peter Marcer, Andrzej Marciniak, Witold Marciszewski, Vladimir Marik, Jacek Martinek, Tomasz Maruszewski, Florian Matthes, Daniel Memmi, Timothy Menzies, Dieter Merkl, Zbigniew Michalewicz, Armin R. Mikler, Gautam Mitra, Roland Mittermeir, Madhav Moganti, Reinhard Moller, Tadeusz Morzy, Daniel Mossé, John Mueller, Jari Multisilta, Hari Narayanan, Jerzy Nawrocki, Rance Necaie, Elzbieta Niedzielska, Marian Niedq' zwiędziński, Jaroslav Nieplocha, Oscar Nierstrasz, Roumen Nikolov, Mark Nissen, Jerzy Nogiec, Stefano Nolfi, Franc Novak, Antoni Nowakowski, Adam Nowicki, Tadeusz Nowicki, Daniel Olejar, Hubert Österle, Wojciech Olejniczak, Jerzy Olszewski, Cherry Owen, Mieczyslaw Owoc, Tadeusz Pankowski, Jens Penberg, William C. Perkins, Warren Persons, Mitja Peruš, Fred Petry, Stephen Pike, Niki Pissinou, Aleksander Pivk, Ullin Place, Peter Planinšec, Gabika Polčicová, Gustav Pomberger, James Pomykalski, Tomas E. Potok, Dimithu Prasanna, Gary Preckshot, Dejan Rakovič, Cveta Razdevšek Pučko, Ke Qiu, Michael Quinn, Gerald Quirchmayer, Vojislav D. Radonjic, Luc de Raedt, Ewaryst Rafajlowicz, Sita Ramakrishnan, Kai Rannenber, Wolf Rauch, Peter Rechenber, Felix Redmill, James Edward Ries, David Robertson, Marko Robnik, Colette Rolland, Wilhelm Rossak, Ingrid Russel, A.S.M. Sajeev, Kimmo Salmenjoki, Pierangela Samarati, Bo Sanden, P. G. Sarang, Vivek Sarin, Iztok Savnik, Ichiro Satoh, Walter Schempp, Wolfgang Schreiner, Guenter Schmidt, Heinz Schmidt, Dennis Sewer, Zhongzhi Shi, Mária Smolárová, Carine Souveyet, William Spears, Hartmut Stadtler, Stanislaw Stanek, Olivero Stock, Janusz Stokłosa, Przemysław Stpicyński, Andrej Stritar, Maciej Stroinski, Leon Strous, Ron Sun, Tomasz Szmuc, Zdzislaw Szyjewski, Jure Šilc, Metod Škarja, Jiří Šlechta, Chew Lim Tan, Zahir Tari, Jurij Tasič, Gheorge Tecuci, Piotr Teczynski, Stephanie Teufel, Ken Tindell, A Min Tjoa, Drago Torkar, Vladimir Tomic, Wieslaw Traczyk, Denis Trček, Roman Trobec, Marek Tudruj, Andrej Ule, Amjad Umar, Andrzej Urbanski, Marko Uršič, Tadeusz Usowicz, Romana Vajde Horvat, Elisabeth Valentine, Kanonkluk Vanapipat, Alexander P. Vazhenin, Jan Verschuren, Zygmunt Vetulani, Olivier de Vel, Didier Vojtisek, Valentino Vranić, Jozef Vyskoc, Eugene Wallingford, Matthew Warren, John Weckert, Michael Weiss, Tatjana Welzer, Lee White, Gerhard Widmer, Stefan Wrobel, Stanislaw Wrycza, Tatyana Yakhno, Janusz Zalewski, Damir Zazula, Yanchun Zhang, Ales Zivkovic, Zonling Zhou, Robert Zorc, Anton P. Železnikar

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: <http://www.informatica.si>.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Vožarski pot 12, 1000 Ljubljana, Slovenia.

The subscription rate for 2009 (Volume 33) is

- 60 EUR for institutions,
- 30 EUR for individuals, and
- 15 EUR for students

Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.

Printing: Dikplast Kregar Ivan s.p., Kotna ulica 5, 3000 Celje.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):

Robotics Society of Slovenia (Jadran Lenarčič)

Slovene Society for Pattern Recognition (Franjo Pernuš)

Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)

Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)

Automatic Control Society of Slovenia (Borut Zupančič)

Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Igor Grabec)

ACM Slovenia (Dunja Mladenič)

Informatica is surveyed by: Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math
--

The issuing of the Informatica journal is financially supported by the Ministry of Higher Education, Science and Technology, Trg OF 13, 1000 Ljubljana, Slovenia.

Informatica

An International Journal of Computing and Informatics

Special Issue on Information Reuse and Integration	R. Alhadj, K. Zhang	243
Identifying Learners Robust to Low Quality Data	A.A. Folleco, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano	245
A General Purpose Software Evaluation System	B.H. Far, V. Mudigonda, A.-H. Elamy	261
Online WordNet Based Tagging System for Social Sharing and Retrieval of Images on Visited Pages	T. Özyer	271
Dialog System for Open-Ended Conversation Using Web Documents	M. Shibata, T. Nishiguchi, Y. Tomiura	277
Optimization of Actions in Activation Timed Influence Nets	M.F. Rafi, A.K. Zaidi, A.H. Levis, P. Papantoni-Kazakos	285
Two-Way Mapping between Object-Oriented Databases and XML	T. Naser, R. Alhadj, M.J. Ridley	297
<hr/> <i>End of Special Issue / Start of normal papers</i>		
Rate Distortion Manifolds as Model Spaces for Cognitive Information	J.F. Glazebrook, R. Wallace	309
Comparative Study of Tripartite Identity-Based Authenticated Key Agreement Protocols	M. Hölbl, T. Welzer, B. Brumen	347
Semantic Graphs Derived From Triplets with Application in Document Summarization	D. Rusu, B. Fortuna, M. Grobelnik, D. Mladenić	357
Improving HTML Compression	P. Skibiński	363
Coordinated UAV Manoeuvring Flight Formation	H. Hexmoor, S. Rahimi, J.T. Little	375
Robustness and Visualization of Decision Models	A. Bregar, J. Györkös, M.B. Jurič	385
Run-time Manipulation of Programs in a Statically-Typed Language	S. Greiner	397

