Volume 34 Number 4 December 2010 ISSN 0350-5596

# Informatica

An International Journal of Computing and Informatics

Special Issue:

**E-Service Intelligence**

Guest Editor:

**Costin Badica**

# EDITORIAL BOARDS, PUBLISHING COUNCIL

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 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. Informatica 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 Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

# Editorial

# E-Service Intelligence

Service Science is an emergent interdisciplinary field aiming at the study of complex systems comprising humans, organizations and technologies engaged in value-added interactive processes. Electronic services, also known as e-services, involve provision of services using digital technologies. Several application areas of e-services have emerged during the last decade: e-business, e-commerce, e-government, e-science, e-learning, and e-health. The next paradigm shift in the digital era proposes the application of state-of-the-art intelligent digital technologies for increasing the various qualities of e-services, like adaptation, personalization, trust, and decision support.

The articles of this special issue address a highly relevant group of topics for the advancement of the field of e-service intelligence: (i) trust, reputation, and expertise modeling in e-services; (ii) intelligence in e-learning; and (iii) Earth observation and collaborative information processing e-services.

The following three papers show how explicit modeling and evaluation of trust, reputation, and expertise can increase credibility of e-services.

E-services for provisioning of human recommendations are usually based on social network environments. The article "Context-based Global Expertise in Recommendation Systems" by Vincenza Carchiolo, Alessandro Longheu, Michele Malgeri, and Giuseppe Mangioni introduces a model and method for context-based assessment of people expertise in a social network. Their model is experimentally evaluated and validated on the Epinions[1] data set.

The blogosphere is a rich and interactive virtual community environment maintained by bloggers that allows tracking of interconnected comments, events, and opinions. The article "BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis" by Yu Weng, Changjun Hu, and Xuechun Zhang proposes a new model for blogger reputation evaluation in distributed environments by mining the opinion relations between bloggers.

Intelligence of e-services can be enhanced by employing rule-based reasoning processes wrapped as software agents. The article "Trusted Reasoning Services for Semantic Web Agents" by Kalliopi Kravari, Efstratios Kontopoulos, and Nick Bassiliades presents the EMERALD multi-agent framework that integrates various trusted reasoning services via software agents' interoperability.

The next two papers address the use of intelligent software technologies for enhancing e-learning services.

The article "A Software System for Viewing and Querying Automatically Generated Topic Maps in the E-Learning Domain" by Liana Stanescu, Gabriel Mihai, Dumitru Burdescu, Marius Brezovan, and Cosmin Stoica Spahiu shows how Topic Maps allow students to semantically browse and query learning resources in a "subject-centric" e-learning system.

Adaptation and personalization are a recent trend for enhancing intelligence of e-learning services. The article "Accommodating Learning Styles in an Adaptive Educational System" by Elvira Popescu, Costin Badica, and Lucian Moraret proposes an innovative learning-style based educational system called WELSA.

The last two papers address specific problems of e-service intelligence in the areas of e-science and decision support in complex situations.

Grid infrastructures have a lot of potential for enhancing Earth observation services. The article "Earth Observation Data Processing in Distributed Systems" by Dana Petcu, Silviu Panica, Marian Neagu, Marc Frincu, Daniela Zaharie, Radu Ciorba, and Adrian Dinis introduces a grid-enabled service-oriented distributed architecture, as well as its proof of concept implementation as a training platform at the West University of Timisoara, Romania.

Decision support in complex applications like crisis management requires efficient collaborative processing of large quantities of heterogeneous information. The article "Dynamic Process Integration Framework: Toward Efficient Information Processing in Complex Distributed Systems" by Gregor Pavlin, Michiel Kamermans, and Mihnea Scafes introduces a flexible service oriented architecture that supports dynamic creation of distributed workflows for collaborative reasoning in knowledge rich domains with minimal ontological commitments.

## Acknowledgements

This special issue is based on selected papers that were presented at IDC'2009, the 3rd International Symposium on Intelligent Distributed Computing held in Cyprus during October 13-14, 2009. The papers were substantionally upgraded and extended. We would like to thank George Angelos Papadopoulos and Konstantinos Kakousis from Department of Computer Science, University of Cyprus, Cyprus who helped us with hosting and organizing IDC'2009. In addition, we would like to thank all the reviewers for their restless reviewing effort and valuable feedback and all the authors who submitted papers to this special issue.

Costin Badica
Craiova, November 2010

[1] http://www.epinion.com

# Context-based Global Expertise in Recommendation Systems

Vincenza Carchiolo, Alessandro Longheu, Michele Malgeri and Giuseppe Mangioni
Dip. Ingegneria Informatica e delle Telecomunicazioni
Facoltà di Ingegneria - Università degli Studi di Catania – Catania - Italy

**Keywords:** expertise, recommendation systems, collaborative filtering, social networks

**Received:** December 8, 2009

*Expertise assessment is frequently required in social networks. This work proposes a method to globally rank people according to their expertise according to a set of topics We also introduce contexts similarity to allow related contexts to be exploited in expertise assessment. These ideas are applied to the Epinions.com recommender systems, showing that expertise in recommendation matters.*

*Povzetek: Opisana je metoda za rangiranje posameznikov glede na socialno omrežje.*

## 1 Introduction and related work

The concept of expertise can be simply defined as the skill or knowledge that a person has in a particular field; more formally, expertise is *"the ability to discriminate meaningful classes of domain features and patterns, and to take decisions or actions that are appropriate to the class at hand"* [9].

Apart from definitions, less trivial matters are how the expertise can be evaluated, and which effective applications it can have. The expertise assessment can be hard, expecially in nowadays virtual social networks (e.g. Facebook, or even e-commerce oriented, as eBay), due to the the lack of real person to person interactions used in real world to judge someone's expertise level. Several approaches to this issue have been proposed [6, 12, 31, 32].

In particular, in [6] the authors aim at ranking the expert candidates in a given topic based on a data collection, hence they locate three components, i.e. a supporting document collection, a list of expert candidates, and a set of expertise topics. This work (as others) shows that an expertise rank is strictly related to a topic, so the question of which set of topics as well as their relationship (for instance, an arrangement into an ontology) should be addressed. That paper also highlights that expertise is commonly inferred from a set of documents (personal profiles, web pages, forum messages etc.) that represent the *use case* where expertise is applied, and they are the virtual counterpart of real world interactions between persons usually used to assess the each other's expertise. The evaluation of an expertise rank by exploiting some data is not a new idea[13], and it has been successfully applied in other scenarios, e.g. Usenet news messages [26] or computer supported cooperative work (CSCW) [16].

In [12], topic expertise is ascertained by exploiting collaborative tagging mechanisms that enable the formation of social networks around tags or topics. Authors state that inferring expertise from data as personal profiles is problematic since users should keep them updated, and they also debate about the granularity in skill levels that should not be either too coarse or too fine (in the former case, automated systems have a difficult time selecting the right people, whereas in the latter users can hardly determine their levels in relation to others).

The work presented in [31] is a propagation-based approach for expertise assessment that takes into account both person local information and relationships between persons; this raises the question of *local* vs *global* approaches that is frequent whenever complex networks are considered, hence not only in expert finding scenarios but also others (e.g. trust, recommendation systems etc.)

In [32] the question of expertise within online communities is addressed, and network structure as well as algorithms are tailored to the case of Java Forum; this suggests that a fundamental role is played by the specific (possibily complex) network being considered and by its properties [18, 2].

In this work we present a method to rank people according to their expertise in a set of topics. We perform this assessment in an *expertise network*, i.e. where the relationship among nodes is the expertise rank assigned in a given context. In particular, we aim at evaluating the *global* expertise a node $v$ has in a network within a specific context based on *local* expertise ranks assigned by $v$'s neighbor nodes. The placement of our work in comparison with issues highlighted in works cited so far can be schematized as follows:

- we infer expertise from data, in particular we exploit the Epinions [25] dataset, where people provide reviews about products and rating about reviews; these ratings are used (see section 3) to infer expertise about people that provided reviews
- expertise is associated to a context (products categories), thus an user can be assigned with several expertise ratings, one for each context; moreover, products categories are arranged into a hierarchy (from general to specific category), hence we leverage this

ontology to manage the granularity of expertise skill levels, the coarser granularity is needed, the more general categories are considered

– since users provide reviews for products in a specific category, we do not need to infer the association between an user and topic expertise (categories), hence we just focus on the expertise ratings assessment

– we aim at taking into account both global and local information, indeed we want to predict the review a specific user is likely to assign to an unknown product based on his neighbours' review about the same item (local information) weighting such reviews with global expertise ranks, as illustrated in sec. 4

The last point reveals the scenario where we intend to apply expertise, i.e. recommendation systems. The concept of expertise indeed is useful in several real applications, e.g. trust and reputation management [11], the assignment of task in an enterprise, or paper reviewers in a conference [12].

Recommender systems are *"a specific type of information filtering technique that attempts to present information items that are likely of interest to the user"*. Such systems gained more and more attention since 1990s, when collaborative filtering approaches were developed [24], since the increasing amount of products available, the markets expansions due to e-commerce, and the diversity of customers also supported and enhanced by social networks, all endorse the need for effective recommender systems [1], a goal that not only relies on algorithms, but also imposes to consider several factors [15]. Recommender systems can adopt the *content-based* or *collaborative filtering* [3] approach; in the former case the system recommends an user about items similar to those he chose in the past, whereas in the latter the system suggests items chosen by similar users. The concept of *similar* users is often based on user profiles, however others ([19]) argue that trustworthiness among users might be also considered. We adopt the expertise as a discriminant factor when considering users' opinions (reviews) to get an effective recommendation; this approach is considered in several works, e.g. [12, 27]. Moreover, we also introduce the contexts *similarity* to allow related contexts to be exploited during ranks assessment, in order to provide an effective recommendation even when the experts' context is not exactly the same the recommended item falls within.

In the rest of paper, section 2 introduces the formalization of our approach, while in section 3 we apply the proposed model to the expertise network built from Epinions.com data set, defining and exploiting contexts similarity, showing results and application to recommender systems in section 4. Section 5 presents our conclusion and future works.

# 2 Expertise evaluation model

The expertise network we refer to in the following is modeled as $G(V, L, lab)$, i.e. a labeled multi-digraph where the set of nodes $V$ represent users[1], $L$ is a set of oriented edges, i.e. an arc $(v, w)$ means that $v$ assigned $w$ at least an expertise rank in a context [2], and the labeling function $lab : L \rightarrow \{(C \times [0, 1])\}$ associates to each arc $(v, w)$ the set of pairs $\{(c_i, r_i)\}$ being $r_i \in [0, 1]$ the expertise rank $v$ assigned to $w$ within context $c_i \in C$ ($C$ is the set of all contexts). Note that for a given arc, the rank is at most one for each context. In the following we indicate $l^{\overline{c}}_{v,w}$ as the rank $r$ associated to the arc $(v, w)$ within context $\overline{c}$, assuming that $l^{\overline{c}}_{v,w} = 0$ for any context $\overline{c}$ when arc $(v, w)$ does not exist, and $L^{\overline{c}} = [l^{\overline{c}}_{v,w}]$ as the weighted adjacency matrix. A transition matrix $P^{\overline{c}}$ is then defined starting from $L^{\overline{c}}$, as detailed later. Each element $p^{\overline{c}}_{vw}$ of $P^{\overline{c}}$ represents a normalized expertise rank $v$ assigned to $w$ in context $\overline{c}$.

To illustrate the mechanism we adopt to assign a global expertise context-specific rank we initially focus on two generic users $v$ and $w$, evaluating how $v$ can assign rank to $w$ in a given context $\overline{c}$. In the real world, if $w$ is one of $v$'s neighbours, it is reasonable to use $p^{\overline{c}}_{vw}$, otherwise $v$ can ask to his neighbours whether they know $w$ to get an opinion about him in $\overline{c}$. In this case, if each of $v$'s neighbours (denoted as $j$) directly knows $w$ he can provide $p^{\overline{c}}_{jw}$, and it is reasonable that $v$ weights these values with ranks he assigned to his neighbours within the same context, thus having $r^{\overline{c}}_{vw} = \sum_j p^{\overline{c}}_{vj} \cdot p^{\overline{c}}_{jw}$.

This one-step neighbours ranking can be written into matrix form as $(\mathbf{r}^{\overline{c}}_v)_{(1)} = (P^{\overline{c}})^T \cdot \mathbf{p}^{\overline{c}}_v$, where $\mathbf{r}^{\overline{c}}_v$ and $\mathbf{p}^{\overline{c}}_v$ are the vectors built from $r^{\overline{c}}_{vw}$ and $p^{\overline{c}}_{vj}$ respectively.

If neighbours $j$ do not directly know the target $w$, $v$ can further extend its requests to two, three, ..., k-steps neighbours, hence at step $(k + 1)$ the ranking assessment is expressed by eq. (1). If $P^{\overline{c}}$ is neither reducible nor periodic [30], $\mathbf{r}^{\overline{c}}_v$ will converge to the same vector for every $v$, specifically to $P^{\overline{c}}$ eigenvector associated with the principal eigenvalue $\lambda_1 = 1$, leading to a *global* expertise rank for $w$ in $\overline{c}$.

$$(\mathbf{r}^{\overline{c}}_v)_{(k+1)} = ((P^{\overline{c}})^T)^{k+1} \cdot (\mathbf{r}^{\overline{c}}_v) \quad (1)$$

This approach is similar to the one proposed by EigenTrust [23] (where trust is replaced by expertise rank), and is frequently adopted also in other well-known works [20], [14]. They also offer a probabilistic interpretation of their method derived from the random walker graph model [17], a widely accepted mathematical formalization of a trajectory. We interpret the random walk as follow: *if an agent $v$ is searching for an expert within a given context $\overline{c}$, he can move along the network choosing the further node $w$ with probability $p^{\overline{c}}_{vw} \in P^{\overline{c}}$*; crawling with this method until a stable state is achieved, the agent is more likely to be at an expert node than at an unqualified node. From the random walker point of view, the first principal eigenvector

[1] In the following, we will use the terms *user* and *node* interchangeably.
[2] in the following, we will use "context" or "topic" indifferently

of $P^{\overline{c}}$ correspond to the standing probability distribution of the Markov chain defined by $P^{\overline{c}}$, and network nodes are its states; thus, we define the expertise vector as the stationary point of the transformation given in (1) with non-negative components.

So far, the context $\overline{c}$ was fixed, but we also want to study contexts influence, i.e. even if the walker is biased by the context $\overline{c}$, it can walk towards an user in a context *similar* to $\overline{c}$, thus we have to choose how the walker moves along the network. To this purpose, we introduce two different walking models, described in the following, both taking into account that an user that has a large number of incoming links (within context $\overline{c}$) is considered an *expert* within $\overline{c}$, hence a walker that moves along a path connecting experts should enforce this quality.

### Strong biased model

Given a topic (or context) $\overline{c} \in C$, a walker standing in a node $v$ at step $k$ moves to one of his outgoing neighbours $w$ at step $k+1$ *if and only if* $l^{\overline{c}}_{vw} > 0$ (i.e. one of $v$'s neighbours is somehow expert within $\overline{c}$).

According to the above definition, we define the transition matrix $P^{\overline{c}}$ as eq. (2), where $outdeg(v)^{\overline{c}}$ is the number of arcs for which $l^{\overline{c}}_{vw} > 0$ (hence $\sum_w p_{vw}$ is always 0 or 1).

$$p^{\overline{c}}_{vw} = \begin{cases} 1/outdeg(v)^{\overline{c}} & \text{if} \quad l^{\overline{c}}_{vw} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

### Smooth biased model

Given a topic $\overline{c} \in C$, a walker standing in a node $v$ at step $k$ moves to one of his outgoing neighbours $w$ at step $k+1$ according to a probability distribution depending on similarity (relatedness) between $\overline{c}$ and all couples $(c_i, r_i)$ labelling $(v, w)$. The relatedness in smooth biased approach $p^{\overline{c}}_{vw}$ is defined accordingly to relatedness between topics pairs labelling the link $(v, w) \in L$ and $\overline{c} \in C$ (eq. (3)). In particular given a topic $\overline{c}$, we define the relatedness function as $d : V \times V \times C \to [0, 1]$ (denoted as $d(v, w, c)|(v, w) \in L, c \in C$). Accordingly, the probability is defined in eq. (3)

$$p^{\overline{c}}_{vw} = d(v, w, \overline{c}) / \sum_j d(v, j, \overline{c}) \tag{3}$$

Note that both strong and smooth biased models may lead to a transition matrix $P^{\overline{c}}$ where all elements of some rows and/or some columns are 0 (therefore $P^{\overline{c}}$ is not irreducible), and sometimes the associated graph might be disconnected.

To find stationary vector the transition matrix is required to be irreducible (equivalent to state that associated digraph is strongly connected [30]) and aperiodic: the first condition implies that exists a directed path from each node to any other, whereas the second implies that for any users $v$ and $w$, there are paths from $v$ to $w$ of any length except for a finite set of lengths.

Both strong and smooth biased models do not work with dangling user and disconnected graphs; dangling users are those with no outgoing link that can be present in any real network. Moreover, in the strong biased case, users that have no outgoing links labelled by topic $\overline{c}$ also became dangling.

Among several solutions for dangling users that have been proposed [20, 5], we choose that a walker in a sink moves to any user according to a given probability distribution. We then define a new transition matrix $(P^{\overline{c}})'$ as $(P^{\overline{c}})' = P^{\overline{c}} + \delta \cdot \alpha^T$, where $\alpha = (1/n, \ldots, 1/n)$ and $\delta = [\delta_i]$ where $\delta_i = 1$ if $i$ is a dangling user and 0 otherwise; this guarantee that $\sum_w p^{\overline{c}}_{vw} = 1, \forall v, w \in V$. The same trick is used to avoid users without ingoing links (that violates the aperiodic property), so achieving the following formula[3]:

$$\begin{aligned} (P^{\overline{c}})'' &= q \cdot (P^{\overline{c}})' + (1-q) \cdot A, \\ where \quad A &= (1, \ldots, 1) \cdot \alpha^T, \quad q \in [0, 1]. \end{aligned} \tag{4}$$

Thus from a non dangling user a walker follows one of the local outgoing links with probability $q$ and jumps to some $w \in V$ with probability $(1-q)$; a common value for $q$ is 0.05 [5].

## 3 Epinions.com: a case study

Epinions (http://www.epinions.com) is a recommendation system that *"helps people make informed buying decisions"*[25]. This goal is achieved through unbiased advice, personalized recommendations, and comparative shopping. Epinions allows registered *user*s to rate *products* writing a *review* in order to provide visitors with *opinions*; a review can be represented as a numeric value plus a text comment about the product. Registered users could also rate the reviews, actually providing an expertise rank about other users. We use the Epinions dataset to validate our approach because it is a large and real dataset and although it is mainly a recommendation network, the reviews voting actually implements an author's reviews reputation mechanism based on products categories (i.e. authors are assigned an expertise rank within contexts).

We however still need to investigate the raw dataset about the assessment of (1) expertise and (2) contexts similarity.

To address the former issue, we consider an user $w$ writing a review on a product (belonging to a given category, e.g. *electronics*), and another user $v$ that can provide a rank to $w$'s review, considering it *useful* or not; $w$ can provide several reviews on products belonging to different categories, and $v$ can rate all of them. Based on such information, we then build the arc $(v,w)$ and label it with a set of pairs $\{(c_i, r_i)\}$, where we associate each context to exactly one products category, and the expertise rank with the

[3] in literature the formula is referred as *teleportation vector*

rate $v$ provided about $w$'s review for the product belonging to that category; note that in the case $w$ reviewed more products belonging to the same category, we evaluate the normalized average rate provided by $v$ over all these products, so that $r_i$ is within the $[0, 1]$ range. Of course, we discard all users that did not provide any review.

Another issue is to define a metric to evaluate context similarity, (e.g. *TVs* category is intuitively much more related to *electronics* than *wellness and beauty*); this is needed by the random walker to exploit different yet related contexts. This semantic distance is a function we name $sim(c_h, c_k) \in [0, 1]$ where 0 means no similarity and 1 means that contexts $c_h$ and $c_k$ are identical terms. Measuring the semantic distance between terms (contexts) has been extensively considered in literature (Vector Space Model [22], Resnik [21], Lesk similarity [4]). Since Epinions provides a hierarchical arrangement of contexts, e.g. *electronics* includes sub-contexts as *cameras & accessories, Home audio*, we can exploit this to provide a simple semantic distance evaluation. In particular, to find the semantic distance between $c_1$ and $c_2$ we search for *"the concept $c_3$ which generalizes $c_1$ and $c_2$ with type $T_3$ such that $T_3$ is the most specific type which subsumes $T_1$ and $T_2$; the semantic distance between $c_1$ and $c_2$ is the sum of the distances from $c_1$ to $c_3$ and $c_2$ to $c_3$"*.

This metric is described in [10, 8] and it satisfies reflexivity, symmetry and triangle inequality properties. Moreover topics types are always the same, therefore our metric can be stated as the "sum of the distance between two concepts and first common ancestor" (along the hierarchical classification provided by Epinions). Finally, we normalize the semantic distance between contexts in order to have values into the $[0, 1]$ range as shown in eq. (5), where $max\_dist$ is the length of the longest path between contexts and $a \prec b$ means that $a$ is an ancestor of $b$ in the Epinion hierarchy.

$$sim(c_i, c_j) = \frac{min(d(c_i, c_k) + d(c_j, c_k))}{max\_dist} \qquad (5)$$

$$\forall c_k \prec c_i, c_j$$

Therefore, the similarity function defined in eq. (5) is used in smooth biased approach to calculate the relatedness between users (see eq. (6)).

$$d(v, w, \overline{c}) = \begin{cases} 1 & \text{if } l^{\overline{c}}_{vw} > 0 \\ \sum_k sim(\overline{c}, c_k) * r_k / \sum_k r_k & \text{otherwise} \end{cases} \qquad (6)$$

Table 3 shows the characteristics of the dataset extracted from Epinions website we used in our first set of experiments.

## 3.1 Results

The expertise network built from Epinions dataset is used to validate the proposed expertise rank assessment model, in particular we evaluate the stationary point of transformation of the transition matrix in eq. (4) using $P^{\overline{c}}$ as defined for strong and smooth biased models, comparing them with an *unbiased* case (i.e. context independent) defined as follows:

| Dataset extracted from www.epinions.com | |
|---|---|
| # nodes | 37 321 |
| # sink (out-degree = 0) | 1 538 |
| # source (in-degree = 0) | 0 |
| # link | 460 504 |
| # totanl n. of topics | 791 |
| average in-degree | 12.13 |
| average out-degree | 8.81 |
| average topics per node | 17.71 |

Table 1: Characteristics of the dataset extracted from Epinions website

$$p_{vw} = \begin{cases} 1/outdeg(v) & if\ outdeg(v) > 0 \\ 0 & otherwise \end{cases} \qquad (7)$$

In the unbiased case (eq. (7)), the transition probability from a node $v$ to a node $w$ is independent from the context $\overline{c}$ hence the steady state probability vector depends only on the structure of the expertise network, i.e. the more links a node has, the more often it will be visited by a random walker. This also means that using the unbiased random walker model an user that has a low number of links will receive a low expertise rank value, even if he is the only one labelled as expert on a given topic $\overline{c}$.

In real life expertise is always assigned within a given context $\overline{c}$ and our idea is to capture this behaviour using a random walker biased by context, as explained in the previous sections. In order to validate the strong and smooth biased random walker models presented in section 2, we will show that the probability of reaching nodes with expertise identical or similar to the target $\overline{c}$ grows with respect to the unbiased case.

In the following we report the results of a set of experiments performed using the network we extracted from Epinions. For each experiment we set a specific topic $\overline{c}$ and we evaluate the expertise vector for the unbiased random walker and for both the strong and the smooth biased random walker models. Therefore, for each topic $c_i$ we sum all the expertise ranks (or steady state probability) of those users labelled with $c_i$ obtaining the so-called *cumulative expertise* of topic $c_i$. It corresponds to the steady state probability that a random walker visits a node belonging to the topic $c_i$. For the sake of simplicity, in the following all the Epinions' topics are indicated by a number instead of their names.

Figures 1,2,3,4 show the comparison of percentage increment of biased models cumulative expertise with respect to unbiased (eq. (7)) considering a common and a rare topic with respect to the unbiased case. In particular, we focused on topic #30 which is very common (i.e. 14995 links associated to it over 460504 total, 14995 source nodes over 35783 and 5134 targets over 26000), and on topic #536

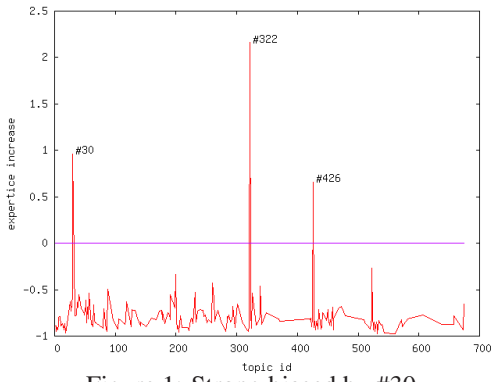



Figure 1: Strong biased by #30

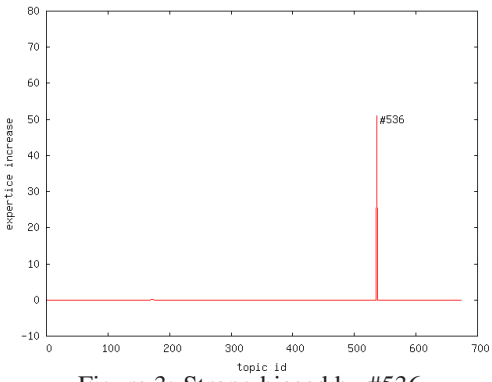



Figure 3: Strong biased by #536

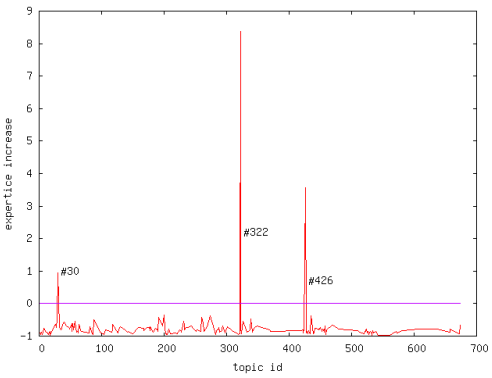



Figure 2: Smooth biased by #30

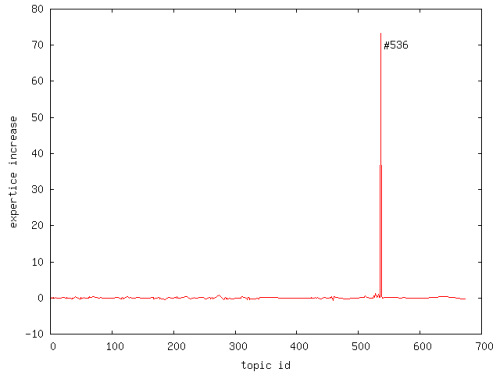



Figure 4: Smooth biased by #536

which is quite rare (5 links, 5 sources and 1 target, respectively).

Results highlight that cumulative expertise on $\overline{c}$ always grows with respect to the unbiased case. Let us note that when expertise is biased by the common topic #30, the cumulative expertise related to some other topics (namely #322 and #426) also increase, whereas when the rare topic #536 is used only nodes labelled with such a topic are affected by biasing (figs. 1,2). The fact that biasing on topic #30 also affects the cumulative expertise of other topics is mainly due to the structure of Epinions network, indeed being topic #30 very common means that a large amount of nodes are somehow expert in that topic. Some of these nodes are also expert in topics #322 and #426 and a certain number of them have a high network degree, so there is a high probability that they are involved in most of paths followed by the random walker, hence the side effect of cumulative expertise increasing for topics #322 and #426 occurs.

Also note that expertise in smooth biased model increases much more for both rare and common topics with respect to the strong biased model, confirming the advantage in exploiting similarity between topics during expertise rank assessment (see fig. 3,4).

Another experiment focuses on Epinions' users, showing their expertise in the rare topic #536, where just one target node $w$ is considered expert by just five other nodes.

Figure 5 highlights each user's expertise on topic #536, evaluated using unbiased, strong and smooth biased models. In particular, we focus on users #3442 and #577, where the former is the only user labelled as expert in topic #536. Expertise evaluated in unbiased and strong-biased case slightly differs for all nodes but #3442 as expected. Indeed the unbiased case for node #3442 shows an expertise value that is nearly zero due to the low number of links such a node has. This confirms that our biased models are able to capture the expertise of a user on a given topic even if the topic is rare and also the node has few links with respect the average nodes degree in the network. The diagram also shows that user #577's expertise for unbiased case is the same as strong biased case since it has no in-links labelled with the topic #536.

The comparison of the smooth biased case with others is more interesting, indeed:

1. node #3442's expertise increases much more than the corresponding strong biased model
2. node #577's expertise increases also!

Item 1 is the expected behavior and confirms our hypothesis that the expertise of a node depends on the opinions of his/her neighbours. Item 2 instead puts in evidence

the influence of highly connected nodes on the expertise evaluation. Specifically, node #577 is much more connected than the average node's connectivity having an out-

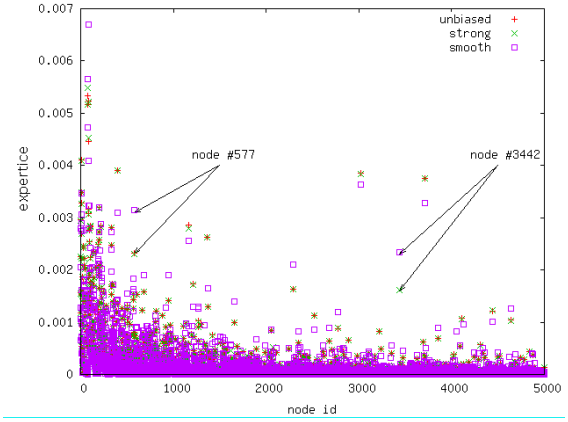



Figure 5: User's expertise assessment

degree of 1333 (versus a network average of 8.81), and in-degree of 241 (versus a network average of 12.13). This means that a random walker tends to visit such a node more frequently than the other nodes of the network, since it is included in many paths. In conclusion, the increasing of expertise not only trivially depends on the expertise in a given topic, but it is also affected by the structure of the network, i.e. the presence of hubs that can be somehow considered expert for (almost) any topic.

# 4 Application of expertise

The assessment of expertise within a social network described so far allows to globally establish how much a user is expert and in which topic; this is useful for several applications (as described in sec. 1), in particular we want to exploit expertise in recommender systems to predict the rating an user will assign to a given item (belonging to a specific category) based on other users reviews' on the same item, mediating such reviews with the reviewers expertise.

We first introduce a measure of the utility an user receives by a specific product, then we consider the Epinion social network again as the application scenario, showing how to predict product ratings using the global expertise ratings.

## 4.1 Getting Utility

In economics, the *utility* is a measure of the relative satisfaction deriving from the consumption of goods or services [7], hence we define the utility function $u(o)$ that allows us to rank the user's preferences on consumed products, that is $u(o_1) > u(o_2)$ means that user strictly prefers $o_1$ instead of $o_2$. Based on the expertise assessment presented in previous sections we define the utility $u(o)$ as follows:

$$u(o) = \frac{\sum_{i \in R^o} r_i \cdot r_i^o}{|R^o|} \tag{8}$$

where $R^o$ is the set of nodes that provided a review about product $o$, $r_i$ is the global expertise rank associated to node $i$ according to the eq. (1) and $r_i^o$ is the numeric score $i$ assigned to $o$ . Note that the real effectiveness of $u(o)$ strictly depends on the number of existing rewiews, indeed the more reviews of distinct users are available, the more affordable $u(o)$ will be.

It is reasonable however that $u(o)$ actually depends on the specific user being considered, i.e. it can happen that $u(o_1) > u(o_2)$ for an user $v$ but not for $w$, for instance due to personal preferences or depending on the categories $o_1$ and $o_2$ belong to. Based on this consideration, we introduce the utility function for a generic user $v$ as $u_v(o) : I \rightarrow [0, 1]$, where $I$ is the set of items (products). The $u_v(o)$ is simply defined as $r_v^o$ if $v$ rated $o$ and zero otherwise.

In addition to the utilities functions, we also consider the *expected* utilities, used to predict how much an user likes an *unknown* product; in recommender systems users usually leverage the others' experience to predict such values. We then introduce the function $e(o) : I \rightarrow [0, 1]$ (and similarly $e_v(o)$) as the expected preference on product $o$ (personal expected preference of $v$ on $o$, respectively); in the following we exploit the global expertise ranks evaluated in (1) and the expertise network $G$ defined in sec. 2 to evaluate both $e(o)$ and $e_v(o)$. For the sake of clarity, we initially take into account only a topic, thus we label each arc of $G(V, L, Lab)$ with $r$ instead of $\{r_i, c_i\}$.

If we want to neglect personal user preferences, i.e. we focus on $e(o)$, it is reasonable that we use the $u(o)$ defined in eq. 8 as the prediction value, so we assume that $e(o) = u(o)$. As stated previously however, a better contribution is given by $e_v(o)$, thus we define $e_v(o)$ as follows:

$$e_v(o) = \frac{\sum_{i \in N_v^o} r_i \cdot r_i^o}{|N_v^o|} \tag{9}$$

where $N_v^o$ is the set of $v$'s neighbours that rated the product $o$. The formula can be extended by considering the (local) expertise rank each neighbour is given by $v$, in order to properly weight their contribution, thus we define:

$$\begin{aligned} e_v^{local}(o) &= \frac{\sum_{i \in N_v^o} r_{vi} \cdot r_i^o}{\sum_{i \in N_v} r_{vi}} \\ e_v^{global}(o) &= \frac{\sum_{i \in N_v^o} r_i \cdot r_i^o}{\sum_{i \in N_v} r_i} \\ e_v(o) &= \delta \cdot e_v^{local} + (1 - \delta) \cdot e_v^{global} \end{aligned} \tag{10}$$

where $r_{vi}$ is the (local) expertise rating $v$ assigned to his neighbour $i$, $r_i$ is the global expertise rank about $i$ evaluated in eq. (1) and $\delta$ allows to balance local and global contributions. Note that we avoid the use of product to calculate expected value for a resource since it always gives a resulting value lower than the highest factors; this is due to the term related with expertise, which is always less than 1 (in some of the example in this paper it ranges over $[3 \cdot 10^{-6}, 5 \cdot 10^{-3}]$ with an average value of $2 \cdot 10^{-5}$).

### 4.2 Products rating prediction

To apply utility functions defined previously in the Epinion dataset to predict product ratings by others experience and expertise, we divided the dataset into a testing dataset and a training dataset, according to insertion date, i.e. reviews and nodes belong to testing or training dataset according to the date they has been stored into Epinions. Currently our dataset contains about 433 000 reviews starting from Jan, 17 2001 up to 4/1/2009, the chosen training set contains all reviews up to Dec, 31 2007 i.e. about 381 000 entries that represents about 88% of total entries, the related ratings on reviews are about 11 800 000 that are 90% of total entries. Once created the dataset, we evaluate the global expertise using only the training dataset, then we begin to add new review according with the date they were actually written, evaluating the expected function (10). Finally the expected function is compared with the real value the user assigned to the resource, in order to assess the effectiveness of such approach.

This comparison is performed using the Mean absolute error (*MAE*), used in statistics [29] to measure how close predictions are to outcomes; *MAE* is defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{n} = \frac{\sum_{i=1}^{n} |\epsilon_i|}{n} \qquad (11)$$

where $\epsilon_i$ is the absolute error between the prediction $f_i$ and the true value $y_i$, and $n$ is the number of samples.

The simulation takes into account all Epinions users that wrote at least one review. Starting from jan 1, 2008 we look for all reviews and compare the expected value to the real one. Results are summarized in table 2, where the $f_i$ used as the expected function is as follows:

1. expected value is calculated using *all* existing reviews starting from jan 1, 2008, i.e. the prediction of (11) is the $u(o)$ defined in eq. (8); this is indicated as approach 1 in table 2
2. expected value is calculated using *only* users that belong to the neighbours of review's author. This is the approach 2 in table 2 and as $f_i$ we used the eq. (9)
3. expected value is calculated using *only* users that belong to neighbours of review's author as in previous case (eq. (9) is used), though in this case just neighbours with some expertise in the category $o$ belongs to are considered (instead of considering all neighbours). This approach is named as approach 3 in table 2.
4. finally, in the approach 4 we used the eq. (10)

The experiments shown good results in all cases because the *MAE* is always less than 1 while rating ranges over the set $\{1, 2, 3, 4, 5\}$. This means that if an expected rate is 3 for instance, the real rate is in the worst case 2 or 4, which can be considered acceptable since it is not a complete turnover of the expressed opinion.

However different strategies highlight that base weighted average (i.e. approach 1) always performs worst than the approches that also exploits local information, in-fact approches 2,3 and 4 select the user to ask for their reference among own neighbourood. The best results are about 0.5 that highlight that error is quite marginal. Unfortunately using only own neighbourhood limits the number of resources (products) that can be ranked, therefore in a real system both approaches should be used

Table 2: mean absolute error

| biasing | approach | category | mae |
|---|---|---|---|
| smooth | 1 | 30 | 0.934 |
| smooth | 2 | 30 | 0.630 |
| smooth | 3 | 30 | 0.591 |
| smooth | 4 | 30 ($\delta = 0.8$) | 0.592 |
| strong | 1 | 30 | 0.821 |
| strong | 2 | 30 | 0.571 |
| strong | 3 | 30 | 0.539 |
| strong | 4 | 30 ($\delta = 0.8$) | 0.595 |
| smooth | 1 | 17 | 0.833 |
| smooth | 2 | 17 | 0.583 |
| smooth | 3 | 17 | 0.548 |
| smooth | 4 | 17 ($\delta = 0.8$) | 0.595 |

## 5 Conclusion

In this work we introduced the *expertise* as a global property of a node and we performed an assessment on the Epinions dataset. *Epinions.com*. Expertise has been defined using a biased random walker model and its corresponding probabilistic interpretation and has been applied to a dataset extracted from Epinions website, where a mechanism of expertise evaluation based on products review has been introduced, together with a similarity function used to exploit topic similarity for a better global expertise rank assessment. Results confirmed that the expertise can be considered a network property that depends on network structure and direct (local) users experience.

Moreover, we applied the global expertise ranks in recommender systems to predict the rating an user would assign to a given item based on other users reviews' on the same item, mediating such reviews with the reviewers expertise. The prediction aimed at integrating both local and global contributions, and simulations shown the effectiveness of such an approach.

Some questions still remains to be addressed:

- we have to investigate on different similarity functions (e.g. when more, different ontologies are present)
- we worked supposing that all information about products ratings and expertise were available, but users may also somehow hide their personal preferences [28]
- in [32] the question of expertise within online communities is addressed, and network structure as well as

algorithms are tailored to the case of Java Forum; this suggests that a fundamental role is played by the specific (possibily complex) network being considered and by its properties [18, 2]. hence we have to test the proposed approach in other networks, since network structure may have a significant impact on both the effectiveness and the efficiency of the approach

# References


[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, 2005.

[2] Reka Albert and Albert-Laszlo Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.

[3] Marko Balabanovic and Yoav Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40:66–72, 1997.

[4] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. *Computational Linguistics and Intelligent Text Processing*, pages 117–171, 2002.

[5] Pavel Berkhin. A survey on pagerank computing. *Internet Mathematics*, 2:73–120, 2005.

[6] Hui Fang and ChengXiang Zhai. Probabilistic models for expert finding. In *ECIR*, pages 418–430, 2007.

[7] Peter Fishburn. *Utility Theory for Decision Making*. Robert E. Krieger Publishing Co., 1970.

[8] Norman Foo, Brian J. Garner, Anand Rao, and Eric Tsui. *Semantic distance in conceptual graphs*. Ellis Horwood, Upper Saddle River, NJ, USA, 1992.

[9] Jared Freeman, Webb Stacy, Jean Macmillan, and Georgiy Levchuk. Capturing and building expertise in virtual worlds. In *FAC '09: Proceedings of the 5th International Conference on Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, pages 148–154, Berlin, Heidelberg, 2009. Springer-Verlag.

[10] B.J. Garner, D. Lukose, and E. Tsui. Parsing natural language through pattern correlation and modification. In *Proc. of the 7th International Workshop on Expert Systems & Their Applications*, pages 1285–1299, Avignon, France, 1987.

[11] T. Grandison and M. Sloman. A survey of trust in internet application. *IEEE Communication Surveys and Tutorials*, 4(4):2–16, 2000.

[12] A. John and D. Seligmann. Collaborative tagging and expertise in the enterprise. *Proc WWW 2006*, 2006.

[13] Henry Kautz, Bart Selman, and Mehul Shah. Referralweb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40:63–65, 1997.

[14] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632, 1999.

[15] Francisco J. Martin. Top 10 lessons learned developing, deploying, and operating real-world recommender systems. In *Proceedings of 3rd ACM Conference on Recommender Systems*, 2009.

[16] D.W. McDonald and M. S. Ackerman. Expertise Recommender: A flexible recommendation system and architecture. In *Proc. Int. Conf. on CSCW*, 2000.

[17] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge University Press, New York, NY, USA, 1995.

[18] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[19] John O'Donovan and Barry Smyth. Trust in recommender systems. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 167–174, New York, NY, USA, 2005. ACM.

[20] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[21] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.

[22] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of ACM*, 18(11):613–620, November 1975.

[23] Hector Garcia-Molina Sepandar D. Kamvar, Mario T. Schlosser. The eigentrust algorithm for reputation management in P2P networks. In *proceedings of the Twelfth International World Wide Web Conference, 2003.*, 2003.

[24] Upendra Shardanand and Patti Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1, pages 210–217, 1995.

[25] Shopping.com Network. Epinions.com ©, http://www.epinion.com, 1999-2010.

[26] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. Phoaks: a system for sharing recommendations. *Communications of ACM*, 40(3):59–62, 1997.

[27] Loren Terveen and David W. McDonald. Social matching: A framework and research agenda. *ACM Trans. Comput.-Hum. Interact.*, 12(3):401–434, 2005.

[28] Frank E. Walter, Stefano Battiston, and Frank Schweitzer. A model of a trust-based recommendation system on a social network. *Journal of autonomous agents and multi-agent systems*, 16:57, 2008.

[29] Frank Edward Walter, Stefano Battiston, and Frank Schweitzer. Personalised and dynamic trust in social networks. In Lawrence D. Bergman, Alexander Tuzhilin, Robin D. Burke, Alexander Felfernig, and Lars Schmidt-Thieme, editors, *RecSys*, pages 197–204. ACM, 2009.

[30] Wolfram MathWorld. http://mathworld.wolfram.com/periodicmatrix.html - http://mathworld.wolfram.com/reduciblematrix.html, 1999-2010.

[31] Jing Zhang, Jie Tang, and Juanzi Li. *Expert Finding in a Social Network*. LNCS, 2008.

[32] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM.

# BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis

Yu Weng
National Language Resource Monitoring Research Centre Minority Language Branch
College of Information Engineering
Minzu University of China, 100081, China
E-mail : mr.wengyu@gmail.com

Changjun Hu, Xuechun Zhang and Liyong Zhao
School of Information and Engineering
University of Science and Technology Beijing, 100083, China
E-mail: zlyong1981@163.com

**Keywords:** blogger reputation evaluation, opinion analysis, distributed computing

**Received:** November 21, 2009

*As a booming virtual community platform, blogosphere has won more and more public attention and preference. For improving the social status analysis ability of blogosphere more effectively, a distributed blogger reputation evaluation model based on opinion analysis is presented (named BREM). The model not only evaluates the reputation level of blogger in the inner-network domain, but also cooperatively schedules the blogger reputation information among the inter-network domains. In the application process, BREM firstly tracks the variation trend of various factors (including the amount of reviews, comments and the published time), identifies the comments opinions of each topic, and evaluates the reputation level of blogger in the single blogosphere periodically. On the other hand, through cooperatively scheduling the local reputation information of bloggers among different blogosphere, the model extends the scope of reputation evaluation and manages the bloggers in the virtual social community more comprehensively. To validate the performance, the experiments on the data corpus about "Unhealthy Campus Culture" demonstrate that BREM has higher application validity and practicality of blogger reputation evaluation in distributed environment.*

*Povzetek: Razvit je model ocenjevanja ugleda blogov na osnovi mnenj.*

## 1 Introduction

In the real society, people usually are classified into different groups by retrieving the personal information (such as age, sexual, job and etc.). However, due to the limits of user authority and personal privacy, the personal information of users could not be obtained freely and truthfully in virtual community-blogosphere. As a kind of novel analysis method, the reputation evaluation has been successfully applied in finance, insurance and the other domains. Using the reputation evaluation into blogosphere will group the virtual community users more effectively and provide the data support for various complex applications.

Recent years, lots of efforts have been made to the research of reputation evaluation. The common idea is to use the number of page links as the estimation of its reputation [1, 2,]. S. Brin and L. Page [3] (1998) modeled the page links graph for the reputation computing, where vertices represent pages and edges represent the links between pages. Klessius Berlt and Nivio Ziviani [4] (2007) proposed a representation of web pages and improved the page links hypergraph evaluation model by reducing the impact of non-votes links. Combined user's individual activity analysis approach and collaborative activity analysis approach, Fusheng Jin and Zhendong Niu [5] (2008) proposed a user reputation model and applied it to the DLDE Learning 2.0 community. Jennifer Golbeck and James Hendler [6] (2004) presented a voting based algorithm for aggregating reputation ratings on the Semantic Web. Some business companies [7, 8] also proposed the online reputation systems to rate and find the more potential customers.

Different from the traditional online reputation calculation methods which mostly focus on the individual activities, the reputation evaluation of blogosphere should give more emphasis on the social relations analysis of bloggers. By mining the comment opinion attitudes of other bloggers (e.g. positive,

negative or neutral), the blogger reputation status in the whole virtual community would be reflected. According to the scenario above, we present a distributed blogger reputation evaluation model based on opinion analysis (named BREM). The model not only evaluates the local reputation of blogger in the single blogosphere, but also cooperatively schedules the blogger reputation information in the other blogospheres. On one side, BREM analyses the semantic orientation (SO) of blog comments and tracks the opinion relations between bloggers. Two calculation methods for long text and short text are adopted respectively. For the long comment text, BREM calculates the SO weight of each character and the distribution density of opinion characters in target text. Through constructing the text opinion case base for long text, the model reuses the evaluation result of historical case and shortens the execution time effectively. For the short comment text, the text opinion is calculated by summing the SO weight of each character. Then, BREM tracks the supportive degree of blog topics and evaluates the reputation of blogger. On the other hand, the model schedules the reputation information of blogger in the other blogosphere periodically and improves the analytical ability of blogger reputation in distributed environment.

This paper is organized as follows. Section 2 outlines the previous approaches of opinion analysis. In section 3, some problems and the general process are described. In section 4, each part of BREM is presented in details. In section 5, experimental results on the corpus of "Unhealthy Campus Culture" are given. Finally section 6 concludes the work with some possible extensions.

## 2 Related works

With the rapid development of Web 2.0 technology, text opinion analysis is attracting more and more attention. Hatzivassiloglou and McKeown [9] (1997) used textual conjunctions such as fair and legitimate or simplistic but well received to separate similarly and oppositely connoted. Pang [10] (2002) classified the documents by sentiment analysis and showed that machine learning approaches on sentiment classification do not perform as well as that on traditional topic-based categorization at document level. Hu and Cheng [11] (2005) illustrated an opinion summarization of bar graph style, categorized by product features. Soo-Min Kim and Eduard Hovy [12] (2006) describe a sentence-level opinion analysis system. The experiment based on MPQA (Wiebe et al. [13], 2005) and TREC (Soboroff and Harman [14], 2003) showed that automatic method for obtaining opinion-bearing words can be used effectively to identify opinion-bearing sentences. Lun-Wei Ku, Hsiu-Wei Ho and Hsin-Hsi Chen [15] (2006) selected TREC, NTCIR[16], and some web blogs as the opinion information sources and proposes an algorithm for opinion extraction at word, sentence and document level. Ruifeng Wong and et al. [17] (2008) Proposed an opinion analysis system based on linguistic knowledge which is acquired from small-scale annotated text and raw topic-relevant webpage. The system used a classifier based on support vector machine to classify the opinion features, identify opinionated sentences and determine their polarities. Veselin and Claire [18] (2008) presented a novel method for general-purpose opinion topic identification and evaluate the validity of this approach by the MPQA corpus. Table 1 shows the comparison of four methods of text opinion analysis.

These technologies above could be applied in the comments opinion analysis in single blogosphere successfully. However, since neglecting the blogger reputation influences of the other network domains, the applied scope and the precision of reputation evaluation would be affected sharply. Through cooperatively scheduling the blogger reputation information among the inter-network domains, BREM comprehensively considers the impacts of topic opinion in multi-blogosphere, strengthens the analysis ability of blogger reputation evaluation and improves the bloggers management level of the whole virtual social community

Table 1: Text Opinion Analysis Comparison.

| Author | Method Description | Testing Results |
|---|---|---|
| Hatzivassiloglou[9] | Identifying the constraints from conjunctions on the positive or negative SO of the conjoined adjectives (e.g. and, but, either-or, etc.). | 21 million words (Wall Street Journal) annotated with part-of-speech tags using the PARTS (Church, 1988). Accuracy: 82%. |
| Peter D. Turney[19] | The classification of a review is predicted by the average SO of the phrases in the review that contain adjectives or adverbs. | 410 reviews from Epinions, sampled from domains (including banks, , movies, travel and automobiles). Accuracy: 74% |
| Lun-Wei Ku[16] | A major topic detection method is proposed to capture main concepts of the relevant documents. Then retrieving all the sentences related to the major topic, determining the opinion polarity of each relevant sentence, and summarizing positive and negative sentences. | TREC corpus, NTCIR corpus and articles from web blogs. TREC corpus is in English, the other two are in Chinese. Accuracy 40% |
| Soo-Minkim [12] | An approach of exploiting the semantic structure of a sentence, anchored to an opinion bearing verb or adjective. This model uses | 2028 annotated sentences from FrameNet data set. (834 from frames related to opinion verb and 1194 from opinion |

| | | |
|---|---|---|
| | semantic role labeling as an intermediate step to label an opinion holder and topic using data from FrameNet. | adjectives) and 100 sentences selected from online news sources (New York Times and BBC) . Accuracy: 47.9% |

## 3 Problems description and general process

To evaluate the blogger reputation more reasonably, in the design process of BREM, the following three parts should be considered:

(1) Comment Opinion Analysis. The aim of comment opinion monitoring is to analyse the attitudes of reviewers to topics (e.g. positive, negative or neutral), and evaluate the blogger reputation more precisely. The calculation process, BREM considers the comprehensive influence of the length of text, the SO (Semantic Orientation) of characters, and the distribution status of opinion characters and identifies the SO of blog comments

(2) Blogger Reputation Evaluation. The reputation of blogger is the reflection of blogger social status in virtual community. Through monitoring the amount of comments, reviews and the semantic opinion of blog comments, BREM could effectively analyse and calculate the supportive degree of the each blog topic and evaluate the reputation of the blogger.

(3) Blogger Reputation Cooperative Scheduling. In the virtual social community-blogosphere, bloggers could publish or comment the topic logs in different blogosphere freely. So the blogger reputation evaluation would be affected by the multi-network domains. BREM simulates the dynamic spreading process, schedules the local blogger information of other network domains and strengthens the blogger reputation analysis ability in the distributed environment.

In Figure 1, the general process of BREM is given. As the part of data preprocessing, firstly BREM analyzes the compositions of blog and represents them by Resource Description Frame (RDF) [20]. Through monitoring the semantic opinion of blog comment, BREM tracks the supportive ratio of the other reviewers to a specific blog topic and evaluates the reputation of bloggers. For improving the practicality of BREM, the model schedules the local blogger reputation information among different blogosphere periodically and manages the bloggers more effectively.

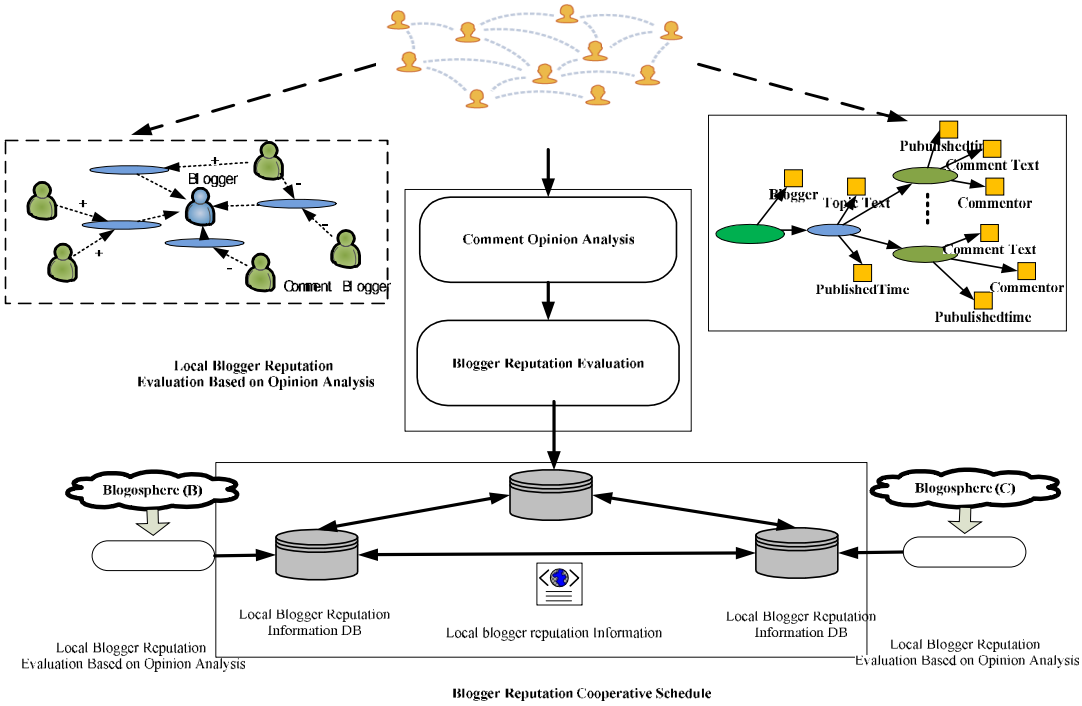



Figure 1: The general process of BREM.

# 4 Distributed blogger reputation evaluation based on opinion analysis

## 4.1 Blog knowledge representation

From the perspective of composition, blogosphere is made up of lots of blogs and the related page links [21]. Each blog includes a series of topics which are ordered by the published time. The author of a blog is named as "blogger" who owns the unique blog sphere. As shown in Figure2 (A) and Figure2 (B), BREM extracts some blog information (blogger, topic title, topic text, published time, comment text and the reviewers) and represents them as the format of RDF [22, 23]. In Figure2 (C) and Figure2 (D), for improving the performance of the blog comment opinion analysis, some typical blog comments are abstracted as the opinion cases.

With the excellent knowledge representation ability of RDF, The opinion case is described as the following three-triples:

$$OpinionCase = \langle Case\ Subject, Case\ Predicate, CaseObject \rangle \quad (1)$$

Here, *Subject* represents the case resources which are uniquely identified by a Uniform Resource Identifier (URI). *Object* denotes the specific literals. Predicate is the binary relation between *Subject* and *Object*. Seven kinds of predicate attributes are defined as *Semantic Opinion*, *Positive Distribution Threshold*, *Negative Distribution Threshold*, *Positive Character Frequency*, *Negative Character Frequency*, *Positive Characters Corpus* and *Negative Characters Corpus*.

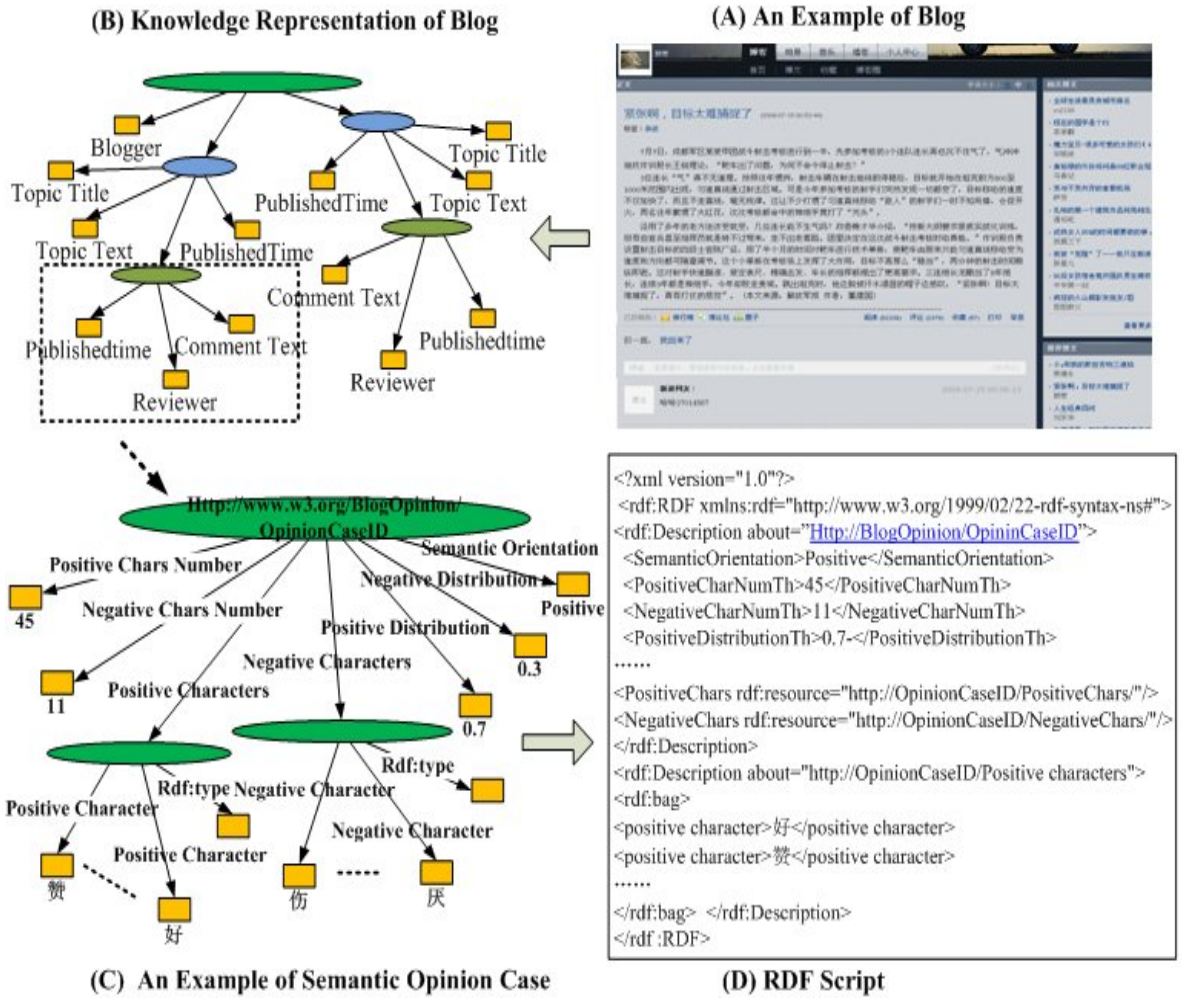



Figure 2: Blog Knowledge Representation Based on RDF.

## 4.2 Blog comment opinion analysis

When we read English text, people could identify the specific word by the blank character. However, in Chinese text, there is no any label between any two words. It greatly increases the difficulty of Chinese text mining. The traditional Chinese text opinion analysis methods usually split the words by some Chinese dictionaries firstly [24]. While, due to being limited by the Chinese segmentation technology, the precision of opinion analysis could not meet the actual application requirement.

Raymond W.M. Yuen and Terence Y.W. Chan [25] (2004) presented a general strategy for inferring SO for Chinese words from their association with some strongly-polarized morphemes. The experimental results proved that using polarized morphemes is more effective than strongly-polarized words. Based on this scenario, BREM improves the calculation model [15] (Liu-Wei Ku and Yu-Ting Liang 2006) and evaluates the text opinion by analyzing the semantic orientation of Chinese characters. In the Blogosphere, users could publish or comment the topics freely. Some comments maybe consist of hundreds of words. Nevertheless, some ones only have dozens of words. For fitting in with the open environment of blogosphere, BREM adopts the different opinion calculation methods for the long text and short text, respectively.

In table 2, $T$ is a paragraph of comment text, $C_i$ represents the i-*th* character of $T$, $N_{count}$ is the amount of words of $T$. $fp_{ci}$ and $fn_{ci}$ stands for the occurring frequency of $C_i$ in positive and negative corpus. $S_{ci}$ denotes the opinion degree of $C_i$. $OpDensity(S_{ci})$ is the distribution density of positive characters. $\| S_{ci} \|$ is the amount of positive characters in $T$. m and n denote the total number of unique characters in positive and negative words corpus. $Th_{LongText}$ is the boundary threshold of long text and short text.

Table 2: Opinion Analysis Algorithm for Blog Comment.

| | |
|---|---|
| Input: | Comment Text $T$, $C_i$, $N_{count}$, $Th_{LongText}$ , $m$, $n$ |
| Output: | Semantic Orientation of $T$ ->$S(T)$ |
| Step 1: | //Initialize Inputs<br>$S(T)=0$;<br>Calculate the SO of each character<br>//transverse all the characters of $T$<br>For each character $C_i$ |
| Step 2: | // where $P_{ci}$ and $N_{ci}$ denote the weights of $c_i$ as positive and negative characters.<br>$$P_{ci} = \frac{fp_{ci} / \sum_{j=1}^{n} fp_{cj}}{fp_{ci} / \sum_{j=1}^{n} fp_{cj} + fn_{ci} / \sum_{j=1}^{m} fn_{cj}} \quad (2)$$ |
| | $$N_{ci} = \frac{fn_{ci} / \sum_{j=1}^{m} fn_{cj}}{fp_{ci} / \sum_{j=1}^{n} fp_{cj} + fn_{ci} / \sum_{j=1}^{m} fn_{cj}} \quad (3)$$ |
| | //SO of character<br>$$S_{ci} = P_{ci} - N_{ci} \quad (4)$$ |
| | if ($\| S_{ci} \| \le Th_{NeutralChar}$)<br>$$S_{ci} = 0 \quad (5)$$ |
| Step 3: | //Evaluate the Comment Opinion of $T$<br>//judge the length of $T$<br>if ($N_{count} <= Th_{LongText}$)<br>// $T$ is short text.<br>then; $$S(T) = \sum_{i=1}^{N_{count}} S_{ci} \quad (6)$$<br>// $T$ is long text.<br>else<br>$S(T)=$<br>$$\sum_{j=1}^{\|S_c^+\|} S_{cj}^{+} \times OpDensity(S_c^{+}) - \sum_{j=1}^{\|S_c^-\|} S_{cj}^{-} \times OpDensity(S_c^{-}) \quad (7)$$ |
| Step 4: | Return $S(T)$; |

In Step2, BREM traverses all the characters of target text $T$ and calculates the SO value of each one. Considering the quantitative difference of positive and negative words corpora, BREM normalizes the occurring frequencies of $C_i$ and evaluates them respectively. In formula 4, through comparing $P_{ci}$ (the character occurring frequency in positive words corpus) and $N_{ci}$ (the character occurring frequency in negative words corpus), the semantic orientation of $C_i$ is determined. If the certain character appears more times in positive words, then it is a positive value; and vice versa. To shorten the calculation error, in formula 5, BREM sets a threshold for the neutral sentiment character in advanced and returns to zero the absolute value less than $Th_{NeutralChar}$.

In Step3, BREM adopts two calculation methods to solve the different length of blog comments respectively. If the length of $T$ is less than the threshold $Th_{LongText}$, the opinion of target text is determined by the SO sum total of all the characters. Otherwise, the length of $T$ is greater than the threshold. We traverse the opinion cases and reuse the historical evaluation result. If there is not any case matching with the target text, as shown in formula 7, the SO of $T$ is evaluated by comprehensively considering the mutual influence of the semantic orientation of characters and the opinion distribution density. In formula 8, through clustering the subjective characters of $T$, BREM analyzes the ratio of the sum of cluster radiuses to the whole amount of characters and calculates the opinion distribution density of subjective characters.

$$OpDensity(S_c^{+}) = \frac{\sum_{i=1}^{k} R_{Cluster}[Position(S_c^{+})]}{N_{count} / 2} \quad (8)$$

Where, *Position* ($S_c^+$) represents the position of $S_c^+$ in *T*. *k* denotes the amount of clusters. Some examples are given in Table 3 and 4.

Table 3: Short Comment Opinion Analysis Examples.

| Short Text 53 words | Short Text 31 words |
|---|---|
| 只有我们自己发扬助人为乐的精神，与人为善，我们才能得到别人的帮助和尊敬, 才能在互动的真诚中感到真正的快乐。<br>(We will not get help and respect from the others until we are willing to help and be kind to anyone else. Then, the real happy will come in good faith. )<br>Score: +19.89<br>Classification: Positive | 我被骗了！我被误了！他们是罪魁祸首，我开始喊了出来，我要宣泄。<br>(I was cheated! I was missed! They are the culprit, and I want to cry loudly and abreact.)<br>Score:-2.22<br>Classification: Negative |

Table 4: Long Comment Opinion Analysis Examples.

| Long Text 201 words | Long Text 145 words |
|---|---|
| 多少年来，人们一直把教师比作红烛，赞颂她默默发光、无怨无悔奉献的精神，教师也一直以蜡烛精神来鞭策自己。“教师”虽说只是一个职业的称谓，而在现实中，教师的职业行为似乎成了他们生活的全部。他们以牺牲自我来换得学生茁壮成长，他们情系学生，情倾讲坛。讲台催人老，粉笔染白头，但教师们却无怨无悔，失志不改，耕耘不辍。这种甘为人梯、无私奉献精神就如同红烛一般，燃烧自己来照亮别人，用自己的付出换来一批批学生的成长。<br>(For so many years, people always use the burning candle to analogy the devotion spirit of teachers. Teachers also use this spirit to encourage themselves. Although “teacher” is a kind of job, it becomes the only part of their lives ……)<br>Sum(S+) : 56.56<br>Sum(S-) : -18.52<br>OpDensity(S+):0.65<br>OpDensity (S-):0.31<br>Score: +30.86<br>Classification: Positive | 商家打着学雷锋的旗号，却利用雷锋的名字和图像乃至其精神进行炒作，甚至恶搞，行为极不严肃，这是对雷锋精神的漠视、曲解和颠覆，是对雷锋精神和形象的一种玷污，也是对时代进步的莫大讽刺。对于这种行为应该严肃对待，坚决抵制，否则我们辛苦树立的英雄形象将会被毁灭，真正的雷锋精神也将会消失在我们的手中。<br>(Some ones use the name the pictures and the spirit of “LeiFeng” to obtain the business interests. That disregards, distorts and subverts the spirit of “Lei Feng”, and satirizes the progress of our times sharply. We should resist these behaviours seriously. Otherwise, the spirit of “LeiFeng” would disappear for ever.)<br>Sum(S+) : 37.15<br>Sum(S-) : -35.50<br>OpDensity (S+):0.54<br>OpDensity (S-):0.70<br>Score: -4.72<br>Classification: Negative |

## 4.3 Blogger reputation evaluation

Given a blogosphere *CBlogosphere*, *A* is any blogger of *CBlogosphere*. In the blogosphere, each blogger could publish the topics in the personal space or comment some ones of other blogs. In formula 9, *Reputation(A,t)* and *Reputation(A,t+1)* represent the reputation of *A* at *t* and *t+1* respectively. $\triangle$ *Reputation(A,t,t+1)* is the increment reputation of *A* within t to t+1.

$$Reputation(A,t+1)= f(Reputation(A,t), \triangle Reputation(A,t,t+1)) \quad (9)$$

Formula 9 is further expanded. As shown in formula 8, through tracking the supportive ratio of blog topics, the reputation of blogger is evaluated dynamically.

$$Reputation(A,t)= \sum_{i=1}^{\|A_{Topic}\|} \left( \frac{\left\|Comments_i^+\right\| - \left\|Comments_i^-\right\|}{\left\|Comments_i\right\|} \right) \bullet \left\|View_i\right\| \quad (10)$$

Where, $\| A_{Topic} \|$ denotes the blog topics of *A*, $\|View_i\|$ and $\left| Comments_i \right|$ represent the reviewers and comments of the i-*th* topic. $\|Comments_i^+\|$ and $\|Comments_i^-\|$ are the number of positive and negative comments. The more the positive comments are, the more reliable the blogger is, and the reputation is higher.

On the contrary, with the increment of negative comments, the reputation of blogger is declined.

In formula 11, $P(A,\triangle t)$ is the increment of reputation between $t$ and $t+1$. BREM analyzes the reputation fluctuation of $A$ in deeply and projects the influence into the range of 0 to 1 by the exponent function.

$$\Delta Reputation(A,t,t+1) = e^{-|P(A,\Delta t)|} = e^{-\left|\frac{Reputation(A,t+1)-Reputation(A,t)}{Reputation(A,t)}\right|} \quad (11)$$

Through monitoring the number variation of positive comments, three cases should be discussed as follow:

(1) With the increment of positive comments (namely P（A,△ t）>0 ), the reliability of blogger is ascended and the reputation is increased.

$$Reputation(A,t+1) = Reputation(A,t) * [1 + Reputation(A,t,t+1)] \quad (12)$$

(2) If the positive comments of two times are equal(namely $P(A,\triangle t)=0$ ), the reputation of blogger keeps invariant.

$$Reputation(A,t+1) = Reputation(A,t) \quad (13)$$

(3) With the reduction of positive comments (namely $P(A,\triangle t)<0$), the reputation of blogger is decreased.

$$Reputation(A,t+1) = Reputation(A,t) * [1 - Reputation(A,t,t+1)] \quad (14)$$

## 4.4 Blogger reputation information cooperative scheduling

To balance the different reputation of the same blogger in multi-blogosphere, BREM further cooperatively schedules the local blogger reputation information and improves the blogger reputation evaluation ability in the global virtual social community.

Given any blogger α . *DomainA* and *DomainB* represent two blogosphere. $BRDB_{DomainA}$ and $BRDB_{DomainB}$ denote the local blogger reputation information database of *DomainA* and *DomainB*, respectively. $\triangle t$ is the time interval of cooperative scheduling. $\beta$ is any blogger of $BRDB_{DomainB}$, and *Th* is the threshold of local reputation variation. The cooperative schedule algorithm of local blogger reputation is as follow:

Table 5: Blogger Reputation Information Cooperative Scheduling Algorithm.

| | |
|---|---|
| Input: | Bloggerα and $\beta$, $BRDB_{DomainA}$ , $BRDB_{DomainB}$, $Th$ , $\triangle t$ |
| Output: | Target reputation information database $BRDB_{DomainB}$ |
| Step 1: | //Initialize Inputs<br>$SendListDomainA, ReceivedListDomainB \leftarrow \phi$ |
| Step 2: | //Local Blogger Reputation Distribution<br>//Traverse all bloggers of $BRDB_{DomainA}$<br>For any blogger $\alpha$ of $BRDB_{DomainA}$<br>//Local Blogger Reputation Evaluation.<br>if ( $Reputation(\alpha,\Delta t)>=Th$ )<br>//Prepare to be scheduled by other network domains<br>then $\alpha$-> $SendListDomainA$ ; |
| Step 3: | //Blogger Reputation Scheduling;<br>//Retrieving $\alpha$ from Domain A<br>$\alpha -> ReceivedListDomainB$ ;<br>//Traverse all blogger reputation information of $BRDB_{DomainB}$<br>For blogger $\beta$ of $BRDB_{DomainB}$<br>//Analysing whether two bloggers are same or not by comparing with the *Email Address*<br>if( $\alpha.email == \beta.email$ )<br>then<br>// If they are same one, update the reputation and take the bigger one.<br>If( $\beta.reputation < \alpha.reputation$ )<br>$\beta.reputation = \alpha.reputation$<br>Elseif ( $\beta.reputation >= \alpha.reputation$ )<br>//prepare to send $\beta$ to *DomainA* and modify the reputation of α<br>$\beta$-> $SendListDomainB$ ;<br>// if they are not the same one, insert new blogger reputation into $BRDB_{DomainB}$<br>else,<br>*Insert* $\alpha$ *into* $BRDB_{DomainB}$ ; |
| Step 4: | //Output<br>Return $BRDB_{DomainB}$; |

In step2, BREM analyses the blogger reputation fluctuation in single blogosphere, and distributes the ones which have the higher number variation to the other network domain. In step3, the model retrieves the local blogger reputation information and queries whether there is the same one by comparing with the email address which is used as the unique identity of blogger. We do not consider wether two or more email addresses belong to the same blogger. If there exists the same blogger in target network domain, BREM updates the reputation

and takes the bigger one. Otherwise, BREM inserts the new blogger reputation information into target database.

# 5 Experiment

## 5.1 Experiment corpus

To validate the performance of BREM, we download over 70,000 blogs (time span from February 4 to May 30, 2009) from the Sina (http://blog.sina.com) and Renren (http://www.renren.com), construct an experimental corpus about "Unhealthy Campus Culture" (named UCC) and test the validity of the comment opinion analysis algorithm and the blogger reputation evaluation. Table 6 presents the information of UCC (Average increment of topics, reviews, comments, long comments and short comments) at four testing time.

Table 6: Information on UCC corpus.

| | Feb. | Mar. | April | May |
|---|---|---|---|---|
| □Topics | 22.2 | 24.1 | 18.3 | 27.4 |
| □Reviews | 164.11 | 126.65 | 185.5 | 144.6 |
| □Comments | 74.30 | 74.88 | 90.4 | 85.4 |
| □ Comments(L) | 30.6 | 31.01 | 40.4 | 35.2 |
| □ Comments (S) | 44.2 | 26.6 | 40.5 | 53.3 |

As the basis of opinion analysis, we collected and revised two sets of opinion words as the testing corpus, including General Positive-Negative Dictionary (abbreviated as GPND) and Chinese Network Sentiment Dictionary (abbreviated as CNSD). Table 7 shows the statistics of the revised testing corpus.

Table 7: Testing Corpus of Opinion Words.

| Dictionary | Positive Corpus | Negative Corpus | Total |
|---|---|---|---|
| GPND | 5,421 | 3,514 | 8,935 |
| CNSD | 1,431 | 1,948 | 3,379 |
| Total | 6,852 | 5,462 | 12,314 |

## 5.2 Experimental results

**Testing 1: The Comparison Testing of Comment Opinion Analysis**

To compare the validity of opinion analysis, we took a comparison testing among OSNB [15], Morpheme [25] and BREM. We selected 40,000 blogs from BREM as the testing set and divided them into long text corpus and short text corpus, respectively. Through calculating the Precision (P), Recall(R) , F-measure (F) and Average Execution Time (T), the performance of three methods was compared.

From the results of the comparison testing, we noticed that, BREM could adapt the different features of long text and short text, and improve the validity and practicability of opinion analysis.

Table 8: Opinion Analysis Comparison Testing for Long Comments.

| | Long Comments Corpus | | |
|---|---|---|---|
| | OSNB | Morpheme | BREM |
| P | 59.87% | 73.85% | 79.49% |
| R | 78.48% | 74.11% | 82.24% |
| F | 67.92% | 73.98% | 80.84% |
| T | 1.5s | 1.1s | 0.5s |

For the long comments corpus, BREM reuses the evaluation results of historical case and comprehensively considers the mutual influence (the semantic orientation of Chinese characters, the distribution density of positive and negative characters). The precision of BREM (P 79.49%, R 82.24%, 0.5 second) is much than OSNB (P 59.87%, R 78.48%, 1.5 second) and Morpheme (P 73.85%, R 74.11%, 1.1second).

Table 9: Opinion Analysis Comparison Testing for Short Comments .

| | Short Comments Corpus | | |
|---|---|---|---|
| | OSNB | Morpheme | BREM |
| P | 54.22% | 70.23% | 72.55% |
| R | 65.79% | 75.03% | 74.28% |
| F | 54.55% | 72.55% | 73.40% |
| T | 1.1s | 0.22s | 0.15s |

For the short comments corpus, BREM adopts the similar method with Morpheme, avoids the limit of Chinese segmentation technology and had better performance than OSNB (P 54.22%, R 65.79%).

**Testing 2: The Validity Testing of Blogger Reputation Evaluation**

Six blogs were constructed to evaluate the validity of blogger reputation evaluation ability of BREM. As shown in Table 10, six kinds of topics about "Unhealthy Campus Culture" are selected from UCC: Unhealthy Psychology (UP, 311, topics), Bad Habits (BH, 165 topics), Warning Speeches (WS, 264 topics), Corruptible Learning (CL, 242 topics), Campus Violence (CV, 202 topics) and Campus Eroticism (CE, 153 topics).

We further input different kinds of alert topics into three blogs of the two blogosphere respectively (Blogosphere A-> Unhealthy Psychology, Bad Habits and Warning Speeches; Blogosphere B->Corruptible Learning, Campus Violence and Campus Eroticism) at three time spans (t1, t2 and t3). Through comparing the number variation of (including the whole amount of comments, comments (+), comments (-) and the reviews) and the trend of blogger reputation, the validity of blogger reputation evaluation will be validated.

Table 10: the Input Data Statistics for Blog Reputation Validity Testing.

| | Av Comments(+)/Page, Av Comments(-)/Page, Av Comments(all)/Page | | |
|---|---|---|---|
| | t1 | t2 | t3 |
| UP | 15(+), 8(-),26 | 35(+),12(-),57 | 64(+), 21(-),96 |
| BH | 14(+), 10(-),33 | 31(+),16(-),72 | 43(+), 22(-),85 |
| WS | 15(+), 12(-),31 | 24(+),19(-),46 | 36(+), 28(-),73 |
| CL | 8(+), 5(-),33 | 19(+),14(-),54 | 22(+), 16(-),68 |
| CV | 12(+), 11(-),33 | 31(+),17(-),58 | 35(+), 37(-),79 |
| CE | 6(+), 8(-),21 | 14(+),11(-),32 | 23(+), 26(-),57 |

| | Av Reviews/Page | | |
|---|---|---|---|
| Type | T1 | T2 | T3 |
| UP | 53 | 84 | 119 |
| BH | 42 | 93 | 133 |
| WS | 56 | 121 | 88 |
| CL | 72 | 163 | 135 |
| CV | 89 | 86 | 94 |
| CE | 74 | 78 | 82 |

The testing results show that, BREM has the good blogger reputation evaluation ability and practicality.

As shown in figure 3, with the amount fluctuation of topic comments and reviews, BREM analyzes the supportive ratio of the other blogger to the topics in deeply and tracks the variation trend within the time span t1 to t3.

Take "Blog A-> Unhealthy Psychology (UP)" and "Blog E->Campus Violence (CV)" for example. The positive comments of blog A increase in the whole time, so the reputation of blogger A ascends. On the contrary, at t2 the supportive ratio of Blog E begins to reduce, BREM captures this trend and lowers the reputation level.

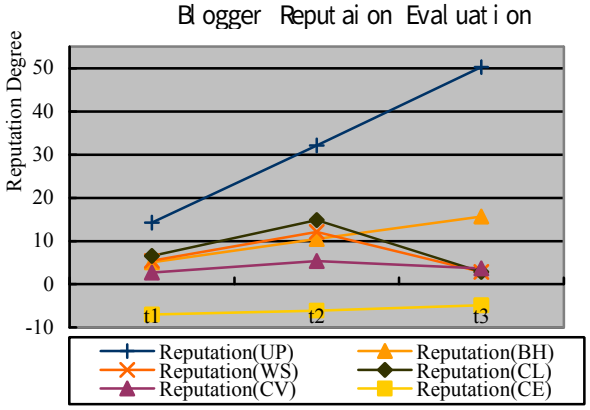



Figure 3.Blogger Reputation Evaluation Validity Testing

# 6 Conclusion & future work

In this paper, a distributed blogger reputation evaluation model based on opinion analysis (named BREM) is proposed. Different with traditional reputation computing methods based on the page links, BREM analyzes the SO of each blog comment, tracks the semantic opinion attitudes of the bloggers and evaluates the blogger reputation level dynamically. Oriented to the length of blog comment, BREM designs two kinds of semantic orientation identification methods by calculating the mutual impacts of opinion weight of Chinese characters and the distribution density of opinion characters comprehensively. To balance the different reputation of the same blogger in the different network domains, BREM cooperatively schedules the local blogger reputation information among the multi-blogosphere and strengthens the management and analysis ability of blogosphere effectively.

In the experiment, we constructed a corpus about "Unhealthy Campus Culture" to validate the comment opinion analysis and the blogger reputation evaluation. The statistics results showed that, with increment of testing corpus, the model had higher opinion analysis ability (Long Comment: Precision 79.49%, Recall 82.24%, Average Executive Time 0.5 second; Short Comment: Precision 72.55%, Recall 74.28%, Average Executive Time 0.15 second) and the validity of blogger reputation evaluation. The statistics results of corresponding compared experiments are showed in table 8 and table 9 which also illustrate the advantage of our method.

In the future work, for improving the calculation scalability of BREM, we will transplant and deploy the original system into the distributed environment or cloud computing platform. With the help of the Map/Reduce technology [26], a blogger reputation evaluation service will be built to strengthen the social status analysis ability of the virtual community - blogosphere

## Acknowledgement

We would like to thank Dr. Liyong Zhao for discussing some issues about this paper. The work reported in this paper was supported by the Key Science-Technology Plan of the National 'Eleventh Five-Year-Plan' of China under Grant No. 2006BAK11B03 and No. 2008AA01Z109, Natural Science Foundation of China under Grant No. 60373008.

# References

[1] T. Bray. Measuring the web. In Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems, Elsevier, pp. 993–1005, Amsterdam, Netherlands, 1996.

[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp.668–677, San Francisco, California, USA, January 1998.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of

the 7th International World Wide Web Conference, pp. 107–117, April 1998.
[4] Klessius Berlt, Edleno Silva de Moura, Andr´e Carvalho and etc. A Hypergraph Model for Computing Page Reputation on Web Collections, SBBD 2007.
[5] Fusheng Jin, Zhendong Niu, Quanxin Zhang, and etc. A User Reputation Model for DLDE Learning 2.0 Community, ICADL 2008, LNCS 5362, pp. 61–70, 2008.
[6] Jennifer Golbeck, James Hendler. Inferring Reputation on the Semantic Web, WWW 2004, May 17-22, 2004, New York, NY USA.
[7] A Resnick, P, Kuwabara, K, A Zeckhauser, R and etc. Reputation systems, 2000, ACM New York, NY, USA.
[8] Yang, M, Feng, Q, Dai, Y and Zhang, Z. A multi-dimensional reputation system combined with trust and incentive mechanisms in P2P file sharing systems. 27th International Conference on Distributed Computing Systems Workshops, 2007. pp. 29–29
[9] HATZIVASSILOGLOU, V., AND MCKEOWN. Predicting the Semantic Orientation of Adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, Madrid, Spain, 1997, pp.174–181.
[10] Pang, B., Lee, L., and Vaithyanathan, Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on EMNLP, 2002, pp. 79–86.
[11] Liu B., Hu M. and Cheng, J. Opinion Observer: Analyzing and Comparing Opinions on the Web. the 14th International World Wide Web Conference, 2005, pp.342–351.
[12] Kim Soo-Min and Hovy Eduard. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, 2006, pp. 1–8.
[13] Wiebe, J., T. Wilson and C. Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 2005.
[14] Soboroff, I. and Harman, D. Overview of the TREC 2003 novelty track. The Twelfth Text REtrieval Conference, National Institute of Standards and Technology, 2003, pp. 38–53.
[15] NTCIR Project, http://research.nii.ac.jp/ntcir/index-en.html.
[16] Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog Corpora. Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006, pp. 100–107.
[17] Ruifeng Xu, Kam-Fai Wong, et al. Learning Knowledge from Relevant Webpage for Opinion Analysis. Web Intelligence and Intelligent Agent Technology, 2008, pp. 307–313.
[18] Veselin Stoyanov and Claire Cardie. Topic Identification for Fine-Grained Opinion Analysis. Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 817–824.
[19] TURNEY, P.D. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the Association for Computational Linguistics 40th Anniversary.
[20] RDF Primer, http://www.w3.org/TR/rdf-primer.
[21] Blog, http://en.wikipedia.org/wiki/Blog
[22] Thomas (2008), Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance. In :Informatica 32, pp.27–38.
[23] S Muñoz, J Pérez, C Gutierrez. Web Semantics: Science, Services and Agents on the world wide web, 2009, Elsevier.
[24] Yohei Seki, David Kirk Evans, Lun-Wei Ku and etc. Overview of opinion analysis pilot task at NTCIR-6, Proceedings of the Workshop Meeting of the National Institute of Informatics (NII) Test Collection for Information Retrieval Systems (NTCIR), pp. 265–278, 2007
[25] R W M Yuen, T Y W Chan et al.Morpheme-based Derivation of Bipolar Semantic Orientation of Chinese Words[A].In:Proceedings of the 20th International Conference on Computational Linguistics,2004,pp.1008–1014.
[26] Hung-chih Yang, Ali Dasdan, Ruey-Lung Hsiao and D. Stott Parker. Map-reduce-merge: simplified relational data processing on large clusters. Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, pp.1029–1040, 2007.

# Trusted Reasoning Services for Semantic Web Agents

Kalliopi Kravari, Efstratios Kontopoulos and Nick Bassiliades
Dept. of Informatics, Aristotle University of Thessaloniki, Greece, GR-54124
E-mail: {kkravari, skontopo, nbassili}@csd.auth.gr

**Keywords:** semantic web, intelligent agents, multi-agent system, reasoning

**Received:** January 16, 2010

*The Semantic Web aims at enriching information with well-defined semantics, making it possible both for people and machines to understand Web content. Intelligent agents are the most prominent approach towards realizing this vision. Nevertheless, agents do not necessarily share a common rule or logic formalism, neither would it be realistic to attempt imposing specific logic formalisms in a rapidly changing world like the Web. Thus, based on the plethora of proposals and standards for logic- and rule-based reasoning for the Semantic Web, a key factor for the success of Semantic Web agents lies in the interoperability of reasoning tasks. This paper reports on the implementation of trusted, third party reasoning services wrapped as agents in a multi-agent system framework. This way, agents can exchange their arguments, without the need to conform to a common rule or logic paradigm – via an external reasoning service, the receiving agent can grasp the semantics of the received rule set. Finally, a use case scenario is presented that illustrates the viability of the proposed approach.*

*Povzetek: Semantični spletni agenti potrebujejo oceno zaupanja storitev za kvalitetno delovanje.*

## 1 Introduction

The *Semantic Web* (*SW*) is a rapidly evolving extension of the World Wide Web that derives from Sir Tim Berners-Lee's vision of a universal medium for data, information and knowledge exchange [1]. The SW aims at augmenting Web content with well-defined semantics (i.e. meaning), making it possible both for people and machines to comprehend the available information and better satisfy their requests. So far, the fundamental SW technologies (content representation, ontologies) have been established and researchers are currently focusing their efforts on logic and proofs.

*Intelligent agents* (*IA*s – software programs extended to perform tasks more efficiently and with less human intervention) are considered the most prominent means towards realizing the SW vision [2]. The gradual integration of *multi-agent systems* (*MAS*) with SW technologies will affect the use of the Web in the imminent future; its next generation will consist of groups of intercommunicating agents traversing it and performing complex actions on behalf of their users.

IAs, on the other hand, are considered to be greatly favored by the interoperability that SW technologies aim to achieve. Thus, IAs will often interact with other agents, belonging to service providers, e-shops, Web enterprises or even other users. However, it is unrealistic to expect that all intercommunicating agents will share a common rule or logic representation formalism; neither can W3C impose specific logic formalisms in a drastically dynamic environment like the Web. In order for agent interactions to be meaningful, nevertheless, agents should somehow share an understanding of each other's position justification arguments (i.e. logical conclusions based on corresponding rule sets and facts). This heterogeneity in representation and reasoning technologies comprises a critical drawback in agent interoperation.

A solution to this compatibility issue could emerge via equipping each agent with its own inference engine or reasoning mechanism, which would assist in "grasping" other agents' logics. Nevertheless, every rule engine possesses its own formalism and, consequently, agents would require a common interchange language. Since generating a translation schema from one (rule) language into the other (e.g. *RIF – Rule Interchange Format* [3]) is not always plausible, this approach does not resolve the agent intercommunication issue, but only moves the setback one step further, from argument interchange to rule translation/transformation.

An alternative, more pragmatic, approach is presented in this work, where reasoning services are wrapped in IAs. Although we have embedded these reasoners in a common framework for interoperating SW agents, called *EMERALD*[1], they can be added in any other multi-agent system. The motivation behind this approach is to avoid the drawbacks outlined above and propose utilizing third-party reasoning services, instead, that allow each agent to effectively exchange its arguments with any other agent, without the need for all involved agents to conform to the same kind of rule paradigm or logic. This way, agents remain lightweight and flexible, while the tasks of inferring knowledge from agent rule bases and verifying the results is conveyed to the reasoning services.

Flexibility is a key aim for our research, thus a variety of popular inference services that conform to various

[1] *http://lpis.csd.auth.gr/systems/emerald/emerald.html*

types of logics is offered and the list is constantly expanding. Furthermore, the notion of trust is vital, since agents need a mechanism for establishing trust towards the reasoning services, so that they can trust the generated inference results. Towards this direction, reputation mechanisms (centralized and decentralized) were proposed and integrated in the EMERALD framework.

The rest of the paper is structured as follows: Section 2 presents a brief overview of the framework, followed by a more thorough description of the reasoning services, in Section 3. Section 4 features the implemented trust mechanisms, while Section 5 reports on a brokering use case scenario that illustrates the use of the reasoning services and the reputation methodology. Finally, the paper is concluded with an outline of related work paradigms, as well as the final remarks and directions for future improvements.

## 2 Framework overview

The EMERALD framework is built on-top of JADE[2] and, as mentioned in the introduction, it involves trusted, third-party reasoning services, deployed as agents that infer knowledge from an agent's rule base and verify the results. The rest of the agents can communicate with these services via ACL message exchange.

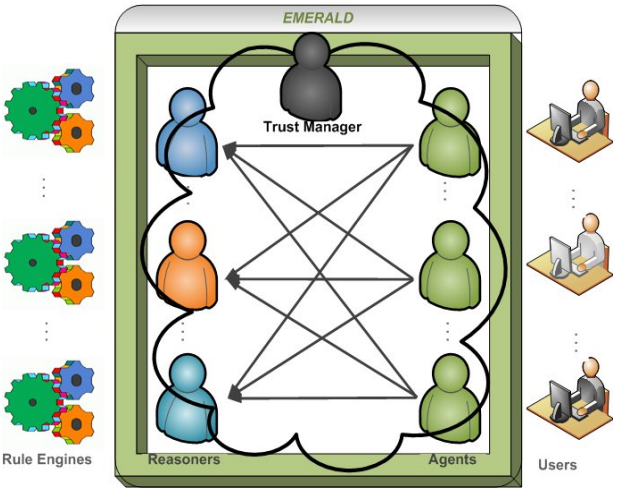



Figure 1: Generic Overview.

Figure 1 illustrates a generic overview of the framework: each human user controls a single all-around agent; agents can intercommunicate, but do not have to "grasp" each other's logic. This is why third-party, reasoning services are deployed. In our approach, reasoning services are "wrapped" by an agent interface, called the Reasoner (presented later), allowing other agents to contact them via *ACL* (*Agent Communication Language*) messages.

The element of trust is also vital, since an agent needs to trust the inference results returned from a Reasoner and is established via centralized and decentralized reputation mechanisms integrated in EMERALD. Figure 1 displays the aspect of the former (centralized) mechanism, where a specialized "*Trust Manager*" agent keeps the reputation scores for the reasoning services given from the rest of the IAs.

Overall, the goal is to apply as many standards as possible, in order to encourage the application and development of the framework. Towards this affair, a number of popular rule engines that comply with various types of (monotonic and non-monotonic) logics are featured in EMERALD (see section 3). Additionally, *RDF*/*S* (*Resource Description Framework/Schema*) and *OWL* (*Web Ontology Language*) serve as language formalisms, using in practice the Semantic Web as infrastructure for the framework.

## 3 Reasoning services

EMERALD currently implements a number of Reasoner agents that offer reasoning services in two main formalisms: deductive and defeasible reasoning.
Table **1** displays the main features of the reasoning engines described in the following sections.

Table 1: Reasoning engine features.

| | **Type of logic** | **Implementation** |
|---|---|---|
| **R-DEVICE** | deductive | RDF/CLIPS/RuleML |
| **Prova** | deductive | Prolog/Java |
| **DR-DEVICE** | defeasible | RDF/CLIPS/RuleML |
| **SPINdle** | defeasible | XML/Java |
| | **Order of Logic** | **Reasoning** |
| **R-DEVICE** | 2nd order | fwd chaining |
| **Prova** | 1st order | bwd chaining |
| **DR-DEVICE** | 2nd order | fwd chaining |
| **SPINdle** | 1st order | fwd chaining |

*Deductive reasoning* is based on classical logic arguments, where conclusions are proved to be valid, when the premises of the argument (i.e. rule conditions) are true. *Defeasible reasoning* [4], on the other hand, constitutes a non-monotonic rule-based approach for efficient reasoning with incomplete and inconsistent information. When compared to more mainstream non-monotonic reasoning approaches, the main advantages of defeasible reasoning are enhanced representational capabilities and low computational complexity [5]. The following subsection gives a brief insight into the fundamental elements of defeasible logics.

### 3.1 Defeasible logics

A *defeasible theory D* (i.e. a knowledge base or a program in defeasible logic) consists of three basic ingredients: a set of facts ($F$), a set of rules ($R$) and a superiority relationship ($>$). Therefore, $D$ can be represented by the triple ($F$, $R$, $>$).

In defeasible logic, there are three distinct types of rules: strict rules, defeasible rules and defeaters. *Strict rules* are denoted by $A \rightarrow p$ and are interpreted in the typical sense: whenever the premises are indisputable, so

[2] *JADE (Java Agent Development Environment): http://jade.tilab.com/*

is the conclusion. An example of a strict rule is: "*Apartments are houses*", which, written formally, would become: `r1: apartment(X) → house(X).`

*Defeasible rules* are rules that can be defeated by contrary evidence and are denoted by $A \Rightarrow p$. An example of such a rule is "*Any apartment is considered to be acceptable*", which becomes: `r2: apartment(X)` $\Rightarrow$ `acceptable(X).`

*Defeaters*, denoted by $A \in p$, are rules that do not actively support conclusions, but can only prevent some of them. In other words, they are used to defeat some defeasible rules by producing evidence to the contrary. An example of a defeater is: `r3: pets(X), gardenSize(X,Y), Y>0` $\in$ `acceptable(X)`, which reads as: "*If pets are allowed in the apartment, but the apartment has a garden, then it might be acceptable*". This defeater can defeat, for example, rule `r4: pets(X)` $\Rightarrow$ $\neg$`acceptable(X).`

Finally, the *superiority relationship* among the rule set $R$ is an acyclic relation > on $R$. For example, given the defeasible rules `r2` and `r4`, no conclusive decision can be made about whether the apartment is acceptable or not, because rules `r2` and `r4` contradict each other. But if a superiority relation > with `r4` > `r2` is introduced, then `r4` overrides `r2` and we can indeed conclude that the apartment is considered unacceptable. In this case rule `r4` is called *superior* to `r2` and `r2` *inferior* to `r4`.

Another important element of defeasible reasoning is the notion of *conflicting literals*. In applications, literals are often considered to be conflicting and at most one of a certain set should be derived. An example of such an application is price negotiation, where an offer should be made by the potential buyer. The offer can be determined by several rules, whose conditions may or may not be mutually exclusive. All rules have `offer(X)` in their head, since an offer is usually a positive literal. However, only one offer should be made. Therefore, only one of the rules should prevail, based on superiority relations among them. In this case, the conflict set is:

$$C(offer(x,y)) = \{\neg offer(x,y)\} \cup \{offer(x,z) \mid z \neq y\}$$

For example, the following two rules make an offer for a given apartment, based on the buyer's requirements. However, the second one is more specific and its conclusion overrides the conclusion of the first one.

```
r5: size(X,Y),Y≥45,garden(X,Z)
    ⇒ offer(X,250+2Z+5(Y-45))
r6: size(X,Y),Y≥45,garden(X,Z),central(X)
    ⇒ offer(X,300+2Z+5(Y-45))
r6 > r5
```

## 3.2 Deductive reasoners

EMERALD currently deploys two deductive reasoners, based on the logic programming paradigm: *R-Reasoner* and *Prova-Reasoner*, which deploy the R-DEVICE and Prova rule engines, respectively.

### 3.2.1 R-DEVICE

*R-DEVICE* [6] is a deductive object-oriented knowledge base system for querying and reasoning about RDF metadata. The system is based on an OO RDF data model, which is different from the established triple-based model, in the sense that resources are mapped to objects and properties are encapsulated inside resource objects, as traditional OO attributes. More specifically, R-DEVICE transforms RDF triples into CLIPS (COOL) objects and uses a deductive rule language for querying and reasoning about them, in a forward-chaining Datalog fashion. This transformation leads to fewer joins required for accessing the properties of a single resource, subsequently resulting in better inference/querying performance.

Furthermore, R-DEVICE features a deductive rule language (in OPS5/CLIPS-like format or in a RuleML-like syntax) for reasoning on top of RDF metadata. The language supports a second-order syntax, which is efficiently translated into sets of first-order logic rules using metadata, where variables can range over classes and properties, so that reasoning over the RDF schema can be performed. A sample rule in the CLIPS-like syntax is displayed below:

```
(deductiverule test-rule
  ?x <- (website (dc:title ?t) (dc:creator
  "John Smith"))
  =>
  (result (smith-creations ?t))
)
```

Rule `test-rule` above seeks for the titles of websites (class website) created by "`John Smith`". Note that namespaces, like DC, can also be used.

The semantics of the rule language of R-DEVICE are similar to Datalog [7] with a semi-naive evaluation proof procedure and an OO syntax in the spirit of F-Logic [8]. The proof procedure of R-DEVICE dictates that when the condition of the rule is satisfied, then the conclusion is derived and the corresponding object is materialized (asserted) in the knowledge base. R-DEVICE supports non-monotonic conclusions. So, when the condition of a rule is falsified (after being satisfied), then concluded object is retrieved (retracted). R-DEVICE also supports negation-as-failure.

### 3.2.2 Prova

*Prova* [9] is a rule engine for rule-based Java scripting, integrating Java with derivation rules (for reasoning over ontologies) and reaction rules (for specifying reactive behaviors of distributed agents). Prova supports rule interchange and rule-based decision logic, distributed inference services and combines ontologies and inference with dynamic object-oriented programming.

As a declarative language with derivation rules, Prova features a Prolog syntax that allows calls to Java methods, thus, merging a strong Java code base with Prolog features, such as backtracking. For example, the following Prova code fragment features a rule, whose body consists of a number of Java method calls:

```
hello(Name):-
       S = java.lang.String("Hello "),
       S.append(Name),
       java.lang.System.out.println(S).
```

On the other hand, Prova reaction rules are applied in specifying agent behavior, leaving more critical operations (e.g. agent messaging etc.) to the language's Java-based extensions. In this affair, various communication frameworks can be deployed, like JADE, JMS[3] or even Java events generated by Swing (G.U.I.) components. Reaction rules in Prova have a blocking `rcvMsg` predicate in the head and fire upon receipt of a corresponding event. The `rcvMsg` predicate has the following syntax: `rcvMsg(Protocol, To, Performative, [Predicate|Args] | Context)`. The following code fragment shows a simplified reaction rule for the FIPA *queryref* performative:

```
rcvMsg(Protocol,From,queryref,[Pred|Args]|
Context):-
   derive([Pred|Args]),
   sendMsg(Protocol,From,reply,[Pred|Args]
   |Context).
rcvMsg(Protocol,From,queryref,[Pred|Args],
Protocol):-
   sendMsg(Protocol,From,end_of_transmissi
   on,[Pred|Args]|Context).
```

The `sendMsg` predicate is embedded into the body of derivations or reaction rules and fails only if the parameters are incorrect or if the message could not be sent due to various other reasons, like network connection problems. Both code fragments presented above were adopted from [9].

Prova is derived from *Mandarax* [10], an older Java-based inference engine, and extends it by providing a proper language syntax, native syntax integration with Java, agent messaging and reaction rules.

## 3.3 Defeasible reasoners

Furthermore, EMERALD also supports two defeasible reasoners: *DR-Reasoner* and *SPINdle-Reasoner*, which deploy DR-DEVICE and SPINdle, respectively.

### 3.3.1 DR-DEVICE

*DR-DEVICE* [11] is a defeasible logic reasoner, based on R-DEVICE presented above. DR-DEVICE is capable of reasoning about RDF metadata over multiple Web sources using defeasible logic rules. More specifically, the system accepts as input the address of a defeasible logic rule base. The rule base contains only rules; the facts for the rule program are contained in RDF documents, whose addresses are declared in the rule base. After the inference, conclusions are exported as an RDF document. Furthermore, DR-DEVICE supports all defeasible logic features, like rule types, rule superiorities etc., applies two types of negation (strong, negation-as-failure) and conflicting (mutually exclusive) literals.

Similarly to R-DEVICE, rules can be expressed either in a native CLIPS-like language, or in a (further) extension of the OORuleML syntax, called *DR-RuleML*, that enhances the rule language with defeasible logic elements. For instance, rule $r_2$ from section 3.1 can be represented in the CLIPS-like syntax as:

```
(defeasiblerule r2
  (apartment (name ?X))
  =>
  (acceptable (name ?X)))
```

For completeness, we also include the representation of rule $r_4$ from section 3.1 in the CLIPS-based syntax, in order to demonstrate rule superiority and negation:

```
(defeasiblerule r4
  (declare (superior r2))
  (apartment (name ?X) (pets "no"))
  =>
  (not (acceptable (name ?X))))
```

The reasoner agent supporting DR-DEVICE is *DR-Reasoner* [12].

### 3.3.2 SPINdle

*SPINdle* [13] is an open-source, Java-based defeasible logic reasoner that supports reasoning on both standard and modal defeasible logic. It accepts defeasible logic theories, represented via a text-based pre-defined syntax or via a custom XML vocabulary, processes them and exports the results via XML. More specifically, SPINdle supports all the defeasible logic features (facts, strict rules, defeasible rules, defeaters and superiority relationships), modal defeasible logics [14] with modal operator conversions, negation and conflicting (mutually exclusive) literals.

A sample theory that follows the pre-defined syntax of SPINdle is displayed below (adopted from the SPINdle website[4]):

```
>> sh          #Nanook is a Siberian husky.
R1: sh -> d    #Huskies are dogs.
R2: sh => -b   #Huskies usually do not bark.
R3: d => b     #Dogs usually bark.
R2 > R3        #R2 is more specific than R3.
#Defeasibly, Nanook should not bark.
#That is, +d -b
```

Additionally, as a standalone system, SPINdle also features a visual theory editor for editing standard (i.e. non-modal) defeasible logic theories.

## 3.4 Reasoner functionality

The reasoning services, as already mentioned, are wrapped by an agent interface, the *Reasoner*, allowing other IAs to contact them via ACL messages. The Reasoner can launch an associated reasoning engine, in order to perform inference and provide results. In essence, the Reasoner is a service and not an autonomous agent; the agent interface is provided in order to integrate Reasoner

[3] *JMS (Java Message Service): http://java.sun.com/products/jms/*

[4] *http://spin.nicta.org.au/spindleOnline/index.html*

agents into EMERALD or even any other multi-agent system.

The procedure is straightforward (Figure 2): each Reasoner constantly stands by for new requests (ACL messages with a "*REQUEST*" communication act). As soon as it gets a valid request, it launches the associated reasoning engine that processes the input data (i.e. rule base) and returns the results. Finally, the Reasoner returns the above result through an "*INFORM*" ACL message.

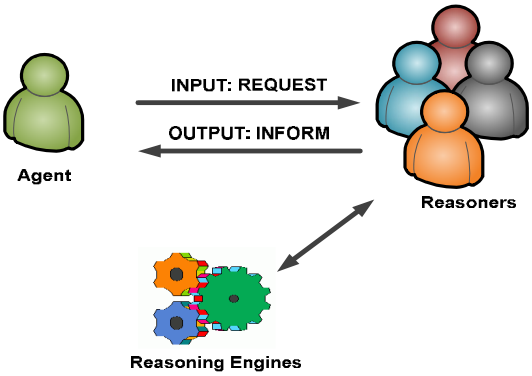



Figure 2: Reasoners' functionality.

A sample ACL message, based on Fipa2000[5] description, in the CLIPS-like syntax is displayed below:

```
(ACLMessage
    (communicative-act REQUEST)
    (sender AgentA@xx:1099/JADE)
    (receiver xx-Reasoner@xx:1099/JADE)
    ....
    (protocol protocolA)
    (language "English")
    (content C:\\rulebase.ruleml)
)
```

where AgentA sends to a Reasoner (`xx-Reasoner`) a RuleML file path (`C:\\rulebase.ruleml`).

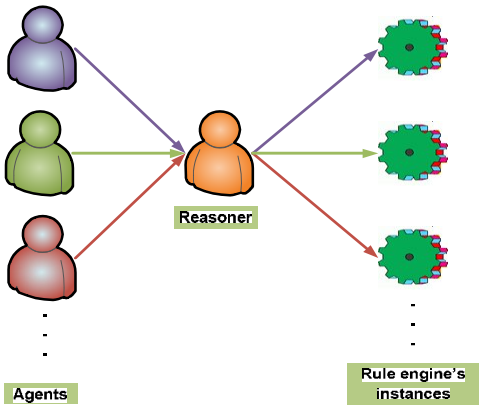



Figure 3: Serving multiple requests.

An important feature of the procedure is that whenever a Reasoner receives a new valid request, it launches a new instance of the associated reasoning engine. Therefore, multiple requests are served concurrently and independently (see Fig. 3). As a result, new requests are served almost immediately, avoiding burdening the framework's performance, because the only sequential operation of the reasoner is the transfer of requests and results between reasoning engines and the requesting agents, which are very low demanding in time.

Finally, note that Reasoners do not use a particular rule language. They simply transfer file paths (in the form of Java Strings) via ACL messages either from a requesting agent to a rule engine or from the rule engine to the requesting agent. Obviously, the content of these files has to be written in the appropriate rule language. For instance an agent who wants to use either the DR-DEVICE or the R-DEVICE rule engine has to provide valid RuleML files. Similarly, valid Prova or XML files are required by the Prova and SPINdle rule engine, respectively. Hence, it is up to the requesting agent's user to provide the appropriate files, by taking each time into consideration the rule engines' specifications.

Thus, new reasoners can be easily created and added to the platform by building a new agent that manages messages between the requesting agent and the rule engine. Furthermore, it has to launch instances of the rule engine according to the specific requirements of the engine.

## 4 Trust mechanisms

Tim Berners-Lee described trust as a fundamental component of his vision for the Semantic Web [1], [15], [16]. Thus, it is not surprising that trust is considered critical for effective interactions among agents in the Semantic Web, where agents have to interact under uncertain and risky situations. However, there is still no single, accepted definition of trust within the research community, although it is generally defined as the expectation of competence and willingness to perform a given task. Broadly speaking, trust has been defined in various ways in literature, depending on the domain of use. Among these definitions, there is one that can be used as a reference point for understanding trust, provided by Dasgupta [17]: "*Trust is a belief an agent has that the other party will do what it says it will (being honest and reliable) or reciprocate (being reciprocative for the common good of both), given an opportunity to defect to get higher payoffs.*"

There are various trust metrics, some involving past experience, some giving relevance to opinions held by an agent's neighbours and others using only a single agent's own previous experience. During the past decade, many different metrics have been proposed, but most have not been widely implemented. Five such metrics are described in [18], among them *Sporas* [19] seems to be the most used metric, although *CR* (*Certified Reputation*) [20] is one of the most recently proposed methodologies.

Our approach adopts two reputation mechanisms, a decentralized and a centralized one. Notice that in both approaches newcomers start with a neutral value. Otherwise, if their initial reputation is set too low, it may be rather difficult to prove trustworthiness through one's

[5] *Fipa2000 description for the ACL Message parameters: www.fipa.org*

actions. If, on the other hand, the reputation is set too high, there may be a need to limit the possibility for users to "start over" after misbehaving. Otherwise, the punishment from having behaved badly becomes void.

### 4.1 Decentralized reputation mechanism

The decentralized mechanism is a combination of Sporas and CR, where each agent keeps the references given from other agents and calculates the reputation value, according to the formula:

$$R_{i+1} = \frac{1}{\theta}\sum_{1}^{t} \Phi(R_i) R_{i+1}^{other}(W_{i+1} - E(W_{i+1})) \quad \textbf{(1)}$$

$$\Phi(R) = 1 - \frac{1}{1 + e^{\frac{-(R-D)}{\sigma}}} \quad \text{and} \quad E(W_{i+1}) = \frac{R_t}{D}$$

where: $t$ is the number of ratings the user has received thus far, $\theta$ is a constant integer greater than 1, $W_i$ represents the rating given by user $i$, $R^{other}$ is the reputation value of the user giving the rating, $D$ is the range of reputation values (maximum rating minus minimum rating) and $\sigma$ is the acceleration factor of the damping function $\Phi$ (the smaller the value of s, the steeper the dumping factor $\Phi$). Note that the value of $\theta$ determines how fast the reputation value of the user changes after each rating. The larger the value of $\theta$, the longer the memory of the system is.

The user's rating value $W_i$ is based on four coefficients:

- *Correctness* ($Corr_i$): refers to the correctness of the returned results.
- *Completeness* ($Comp_i$): refers to the completeness of the returned results.
- *Response time* ($Resp_i$): refers to the Reasoner's response time.
- *Flexibility* ($Flex_i$): refers to the Reasoner's flexibility in input parameters.

The four coefficients are evaluated, based on the user's (subjective) assessment for each standard and their ratings vary from 1 to 10. The final rating value ($W_i$) is the weighted sum of the coefficients (equation (2) below), where $a_{i1}$, $a_{i2}$, $a_{i3}$ and $a_{i4}$ are the respective weights and $nCorr_i$, $nComp_i$, $nResp_i$ and $nFlex_i$ are the normalized values for correctness, completeness, response time and flexibility, accordingly:

$$w_i = a_{i1}nCorr_i + a_{i2}nComp_i + a_{i3}nResp_i + a_{i4}nFlex_i \quad \textbf{(2)}$$

New users start with a reputation equal to 0 and can advance up to the maximum of 3000. The reputation ratings vary from 0.1 for "terrible" to 1 for "perfect". Thus, as soon as the interaction ends, the Reasoner asks for a rating. The other agent responds with a new message containing both its rating and its personal reputation and the Reasoner applies equation (1) above to update its reputation.

### 4.2 Centralized reputation mechanism

In the centralized approach, a third-party agent keeps the references given from agents interacting with Reasoners or any other agent in the MAS environment. Each reference is in the form of:

$$Ref_i=(a, b, cr, cm, flx, rs)$$

where: $a$ is the *truster agent*, $b$ is the *trustee agent* and *cr* (*Correctness*), *cm* (*Completeness*), *flx* (*Flexibility*) and *rs* (*Response time*) are the evaluation criteria.

Ratings ($r$) vary from -1 (*terrible*) to 1 (*perfect*), while newcomers start with a reputation equal to 0 (*neutral*). The final reputation value ($R_b$) is based on the weighted sum of the relevant references stored in the third-party agent and is calculated according to the formula:

$$\sum R_b = w_1*cr + w_2*cm + w_3*flx + w_4*rs$$

where: $w_1+w_2+w_3+w_4=1$. Two options are supported for $R_b$, a default where the weights are equivalent, namely $w_{k\ s[1,4]}=0.25$ each and a user-defined, where the weights vary from 0 to 1 depending on user priorities.

### 4.3 Comparison

The simple evaluation formula of the centralized approach, compared to the decentralized one, leads to time gain as it needs less calculation time. Moreover, it provides more guaranteed and reliable results ($R_b$), as it is centralized, overcoming the difficulty to locate references in a distributed mechanism.

In addition, in the decentralized approach an agent can interact with only one agent per time and, thus, requires more interactions, in order to discover the most reliable agent, leading to further time loss.

Agents can use either of the above mechanisms or even both complementarily. Namely, they can use the centralized mechanism, in order to find the most trusted service provider and/or they can use the decentralized approach for the rest of the agents.

## 5 Use case: a brokering scenario

Defeasible reasoning (see section 3) is useful in various applications, like brokering [21], bargaining and agent negotiations [22]. These domains are also extensively influenced by agent-based technology [23]. Towards this direction, a defeasible reasoning-based brokering scenario is adopted from [24]. In order to demonstrate the functionality of the presented technologies, part of the above scenario is extended with deductive reasoning. Four independent parties are involved, represented by intercommunicating intelligent agents.

- The *customer* (called Carlo) is a potential renter that wishes to rent an apartment based on his requirements (e.g. location, floor) and preferences.
- The *broker* possesses a number of available apartments stored in a database. His role is to match Carlo's requirements with the features of the available apartments and eventually propose suitable flats to the potential renter.
- Two *Reasoners* (independent third-party services), *DR-Reasoner* and *R-Reasoner*, with a high reputation rating that can conduct inference on defeasible and

deductive logic rule bases, accordingly, and produce the results as an RDF file.

## 5.1 Scenario overview

The scenario is carried out in eight distinct steps, as shown in Fig. 4 Carlo's agent retrieves the corresponding apartment schema (Appendix A), published in the broker's website, formulates his requirements accordingly and submits them to the broker, in order to get back all the available apartments with the proper specifications (Fig. 4 – step 1). These requirements are expressed in defeasible logic, in the DR-DEVICE RuleML-like syntax (Fig 5 and Fig 6). For the interested reader, Appendix B features a full description of the customer's requirements in d-POSL (see Appendix E), a POSL[25]-like dialect for representing defeasible logic rule sets in a more compact way.

The broker, on the other hand, has a list of all available apartments, along with their specifications (stored as an RDF database – see **Figure** 7 for an excerpt), but does not reveal it to Carlo, because it's one of his most valuable assets. However, since the broker cannot process Carlo's requirements using defeasible logic, he requests a trusted third-party reasoning service. The DR-Reasoner, as mentioned, is an agent-based service that uses DR-DEVICE, in order to infer conclusions from a defeasible logic program and a set of facts in an RDF document. Hence, the broker sends the customer's requirements, along with the URI of the RDF document containing the list of available apartments, and stands by for the list of proper apartments (step 2).

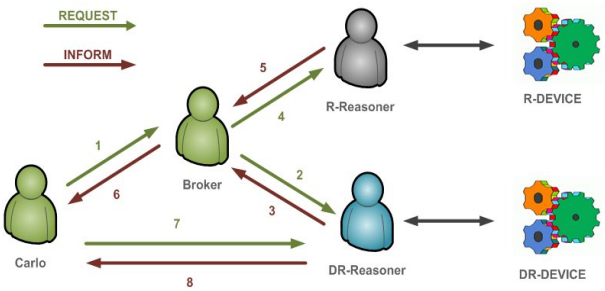



Figure 4: The distinct steps featured in the scenario.

Then, DR-Reasoner launches DR-DEVICE, which processes the above data and returns an RDF document, containing the apartments that fulfil all requirements (Fig. 8). When the result is ready, the Reasoner sends it back to the broker's agent (step 3). The latter should forward the results to Carlo's agent; however, the broker possesses a private "agenda", i.e. a rulebase that infers broker's proposals, according to his/her own strategy, customized to Carlo's case, i.e. selected from the list of apartments compatible to Carlo's requirements. A sample of these rules is shown in Appendix C; one rule proposes the biggest apartment in the city centre, while the other one suggests the apartment with the largest garden in the suburbs. These rules are formulated using deductive logic, so the broker sends them, along with the results of the previous inference step, to the R-Reasoner that launches R-DEVICE (step 4). Finally, the broker gets the appropriate list with proposed apartments that fulfil his "special" rules (step 5).

```
<RuleML rdf_import="...carlo_ex.rdf" rdf_export="export-carlo.rdf" >
.....
<Implies ruletype = "defeasiblerule">
 <oid><Ind uri = "&carlo_rb;r1">r1</Ind></oid>
 <head>
  <Atom>
   <op><Rel>acceptable</Rel></op>
   <slot><Ind>apartment</Ind><Var>x</Var></slot>
  </Atom>
 </head>
 <body>
  <Atom><op><Rel uri = "carlo:apartment"/></op>
   <slot><Ind>carlo:name</Ind><Var>x</Var></slot>
  </Atom>
 </body>
</Implies>
.....
</rulebase>
```

Figure 5: Rule base fragment – rule $r_1$.

```
<RuleML rdf_import="...carlo_ex.rdf" rdf_export="export-carlo.rdf" >
.....
<Implies ruletype = "defeasiblerule">
 <oid><Ind uri = "&carlo_rb;r2">r2</Ind></oid>
 <head>
  <Neg>
   <Atom>
    <op><Rel>acceptable</Rel></op>
    <slot><Ind>apartment</Ind><Var>x</Var></slot>
   </Atom>
  </Neg>
 </head>
 <body>
  <Atom>
   <op><Rel uri = "carlo:apartment"/></op>
   <slot><Ind>carlo:name</Ind><Var>x</Var></slot>
   <slot><Ind>carlo:bedrooms</Ind>
    <Constraint>
     <and_constraint><Var>y</Var>
      <Function_call name = "&lt;">
       <Var>y</Var><Ind>2</Ind>
      </Function_call>
     </and_constraint>
    </Constraint>
   </slot>
  </Atom>
 </body>
 <superior><Ind uri = "&carlo_rb;r1"/></superior>
</Implies>
.....
</rulebase>
```

Figure 6: Rule base fragment – rule $r_2$.

Eventually, Carlo receives the appropriate list (step 6) and has to decide which apartment he prefers. However, his agent does not want to send Carlo's preferences to the broker, because he is afraid that the broker might take advantage of that and will not present him with his most preferred choices. Thus, Carlo's agent sends the list of acceptable apartments (an RDF document) and his preferences (once again as a defeasible logic rule base) to the Reasoner (step 7). The latter calls DR-DEVICE and

gets the single most appropriate apartment. It replies to Carlo and proposes the best transaction (step 8). The procedure ends and Carlo can safely make the best choice based on his requirements and personal preferences. See Appendix D for a d-POSL version of Carlo's specific preferences. Notice that Carlo takes into consideration not only his preferences and requirements, but also broker's proposals, as long as they are compatible with his own requirements.

```
<rdf:RDF... xmlns:carlo="&carlo;">
  <carlo:apartment rdf:about="&carlo_ex;a1">
    <carlo:bedrooms rdf:datatype="&xsd;integer">1</carlo:bedrooms>
    <carlo:central>yes</carlo:central>
.....
  </carlo:apartment>
.....
</rdf:RDF>
```

Figure 7: RDF document excerpt for available apartments.

As for the reputation rating, after each interaction with the Reasoners, both the Broker and the Customer are requested for their ratings. For instance, after the successful end of step 3, the Broker not only proceeds to step 4, but also sends its rating to the Reasoner or/and the third-party agent. As a result, the latter updates the reputation value.

```
<!DOCTYPE rdf:RDF [
<!ENTITY dr-device "http://.../dr-device/export/export-carlo.rdf#"> ]>
<rdf:RDF... xmlns:dr-device='&dr-device;'>
.....
<dr-device:acceptable rdf:about="&dr-device;acceptable5">
<dr-device:apartment>a5</dr-device:apartment>
<dr-device:truthStatus>defeasibly-proven</dr-device:truthStatus>
</dr-device:acceptable >
.....
</rdf:RDF>
```

Figure 8: Results of defeasible reasoning exported as an RDF document.

## 5.2 Brokering protocol

Although FIPA provides standardized protocols, we found that none is suitable for our brokering scenario, since 1-1 automated brokering cannot be supported. As a result, a brokering protocol was implemented that encodes the allowed sequences of actions for the automation of the brokering process among the agents. The protocol is depicted in Fig. 9 and is based on specific performatives that conform to the FIPA ACL specification.

$S_0$ to $S_6$ represent the states of a brokering trade and $E$ is the final state. Predicates *Send* and *Receive* represent the interactions that cause state transitions. For instance, the sequence of transitions for the customer is: $S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4 \rightarrow S_5 \rightarrow E$, which means that the agent initially sends a *REQUEST* message ($S_1 \rightarrow S_2$) to the broker, then waits and finally gets an *INFORM* message with the response ($S_2 \rightarrow S_3$). After that, the customer decides to send a new request message to the DR-Reasoner ($S_3 \rightarrow S_4$), receives an *INFORM* message from him ($S_4 \rightarrow S_5$) and successfully terminates the process ($S_5 \rightarrow E$).

On the other hand, the transition sequence for the broker is: $S_0 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4 \rightarrow S_5 \rightarrow S_6 \rightarrow E$. Initially, the agent is waiting for new requests; as soon as one is received ($S_0 \rightarrow S_1$), he sends an enriched *REQUEST* message to the DR-Reasoner ($S_1 \rightarrow S_2$) and waits for results. Finally, he gets the *INFORM* message from the DR-Reasoner ($S_2 \rightarrow S_3$) and sends a new enriched *REQUEST* message to the R-Reasoner ($S_3 \rightarrow S_4$). Eventually, the broker receives the appropriate *INFORM* message from the R-Reasoner ($S_4 \rightarrow S_5$) and forwards it to the customer ($S_5 \rightarrow S_6$), terminating the trade ($S_6 \rightarrow E$).

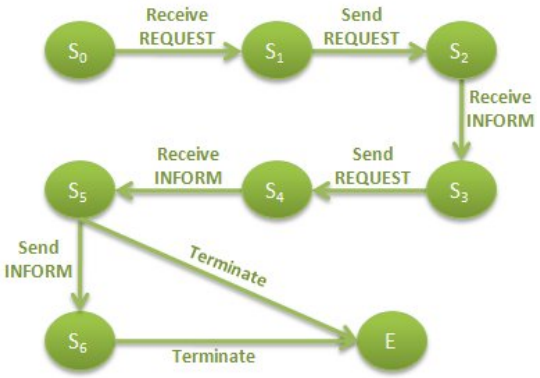



Figure 9: Agent brokering communication protocol.

In case that an agent receives a wrong performative, it sends back a NOT-UNDERSTOOD message and the interaction is repeated.

# 6 Related work

A similar architecture for intelligent agents is presented in [26], where various reasoning engines are employed as plug-in components, while agents intercommunicate via FIPA-based communication protocols. The framework is build on top of the OPAL agent platform [27] and, similarly to EMERALD, features distinct types of reasoning services that are implemented as reasoner agents. The featured reasoning engines are *3APL* [28], *JPRS* (*Java Procedural Reasoning System*) and *ROK* (*Rule-driven Object-oriented Knowledge-based System*) [29]. 3APL agents incorporate BDI logic elements and first-order logic features, providing constructs for implementing agent beliefs, declarative goals, basic capabilities and reasoning rules, through which an agent's goals can be updated or revised. JPRS agents perform goal-driven procedural reasoning and each JPRS agent is composed of a world model (agent beliefs), a plan library (plans that the agent can use to achieve its goals), a plan executor (reasoning module) and a set of goals. Finally, ROC agents are composed of a working memory, a rule-base (consisting of first-order, forward-chaining production rules) and a conflict set. Thus, following a similar approach to EMERALD, the framework integrates the three reasoning engines into OPAL in the form of OPAL micro-agents.

The primary difference between the two frameworks lies in the variety of reasoning services offered by EMERALD. While the three reasoners featured in [26] are all based on declarative rule languages, EMERALD proposes a variety of reasoning services, including deductive, defeasible and modal defeasible reasoning, thus, comprising a more integrated solution. Furthermore, the framework does not feature a trust and reputation mechanism. Finally, and most importantly, the approach of [26] is not based on Semantic Web standards, like EMERALD, for rule and data interchange.

The *Rule Responder* [30] project builds a service-oriented methodology and a rule-based middleware for interchanging rules in virtual organizations, as well as negotiating about their meaning. Rule Responder demonstrates the interoperation of various distributed platform-specific rule execution environments, based on Reaction RuleML as a platform-independent rule interchange format. We have a similar view of reasoning service for intelligent agents and usage of RuleML. Also, both approaches allow utilizing a variety of rule engines. However, contrary to Rule Responder, our framework (EMERALD) is based on FIPA specifications, achieving a fully FIPA-compliant model and proposes two reputation mechanisms to deal with trust issues. Finally, and most importantly, our framework does not rely on a single rule interchange language, but allows each agent to follow its own rule formalism, but still be able to exchange its rule base with other agents, which will use trusted third-party reasoning services to infer knowledge based on the received ruleset.

*DR-BROKERING*, a system for brokering and matchmaking, is presented in [31]. The system applies RDF in representing offerings and a deductive logical language for expressing requirements and preferences. Three agent types are featured (Buyer, Seller and Broker). Similarly, our approach identifies roles such as Broker and Buyer. On the other hand, we provide a number of independent reasoning services, offering both deductive and defeasible logic. Moreover, our approach takes into account trust issues, providing two reputation approaches in order to guarantee the interactions' safety.

In [32] a negotiation protocol and a framework that applies it are described. Similarly to our approach, the proposed framework also uses JADE. Additionally, a taxonomy of declarative rules for capturing a wide variety of negotiation mechanisms in a well-structured way is derived. The approach offers the same advantages with EMERALD, namely, the involved mechanisms are being represented in a more modular and explicit way. This makes agent design and implementation easier, reducing the risks of unintentional incorrect behaviour. On the other hand, EMERALD comprises a more generic framework, allowing the adoption of various scenarios that are not only restricted in negotiations. Moreover, reasoning services are provided, along with two reputation models for agents.

# 7 Conclusions

The paper argued that agent technology will play a vital role in the realization of the Semantic Web vision and presented a variety of reasoning services, wrapped in an agent interface, embedded in a common framework for interoperating SW IAs, called *EMERALD*, a JADE multi-agent framework designed specifically for the Semantic Web. This methodology allows each agent to effectively exchange its argument base with any other agent, without the need for all involved agents to conform to the same kind of rule paradigm or logic. Instead, via EMERALD, IAs can utilize third-party reasoning services, that will infer knowledge from agent rule bases and verify the results.

The framework offers a variety of popular inference services that conform to various types of logics. Additionally, since agents need a mechanism for establishing trust towards the reasoning services, reputation mechanisms (centralized and decentralized) were integrated in the framework and were also described in this work. Finally, the paper presents a use case brokering trade scenario that illustrates the usability of the technologies described in the paper.

As for future directions, it would be interesting to verify our model's capability to adapt to a variety of different scenarios other than brokering. An appealing field could be contract negotiation; the incorporation of negotiation elements into the agents' behavior would demand alterations in the protocol. The latter would now have to include the agents' negotiation strategy as well. Another goal is to integrate an even broader variety of distinct reasoning engines, thus, forming a flexible, generic environment for interoperating agents in the SW. Finally, our intention is to test our reasoning services (reasoners) in data intensive applications.

# References

[1] Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. *Scientific American*, 284(5):34-43

[2] Hendler J (2001) Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2):30-37

[3] Boley H, Kifer M. *RIF Basic Logic Dialect*. Latest version available at http://www.w3.org/TR/rif-bld/.

[4] Nute D. (1987) Defeasible Reasoning. *20th International Conference on Systems Science*, IEEE Press, pp. 470-477.

[5] Maher MJ (2001) Propositional defeasible logic has linear complexity. *Theory and Practice of Logic Programming* 1(6):691–711.

[6] Bassiliades N, Vlahavas I (2006) R-DEVICE: An Object-Oriented Knowledge Base System for RDF Metadata. *International Journal on Semantic Web and Information Systems*, 2(2):24-90.

[7] Abiteboul S, Hull R, Vianu V (1995) *Foundations of Databases*. Addison-Wesley, p. 305.

[8] Kifer M, Lausen G, Wu J (1995) Logical foundations of object-oriented and frame-based languages. *J. ACM* 42(4):741-843.

[9] Kozlenkov A, Penaloza R, Nigam V, Royer L, Dawelbait G, Schroeder M (2006) Prova: Rule-based Java Scripting for Distributed Web Applications: A Case Study in Bioinformatics. In Sebastian Schaffert (Ed.) *Workshop on Re-*

*activity on the Web at the International Conference on Extending Database Technology (EDBT 2006)*, Springer.

[10] Dietrich J, Kozlenkov A, Schroeder M, Wagner G (2003) Rule-based agents for the semantic web. *Electronic Commerce Research and Applications*, 2(4):323–338.

[11] Bassiliades N, Antoniou G, Vlahavas I (2006) A Defeasible Logic Reasoner for the Semantic Web. *International Journal on Semantic Web and Information Systems*, 2(1):1-41.

[12] Kravari K, Kontopoulos E, Bassiliades N (2009) Towards a Knowledge-based Framework for Agents Interacting in the Semantic Web. *2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'09)*, Italy, Vol. 2, pp. 482-485.

[13] Lam H, Governatori G (2009) The Making of SPINdle. *RuleML-2009 International Symposium on Rule Interchange and Applications*, Springer, pp. 315-322.

[14] Governatori, G, Rotolo, A (2008). BIO logical agents: Norms, beliefs, intentions in defeasible logic. *Journal of Autonomous Agents and Multi Agent Systems* 17:36–69.

[15] Berners-Lee T (1999) Weaving the Web, Harper San Francisco, ISBN: 0062515861.

[16] Berners-Lee T, Hall W, Hendler J, O'Hara K, Shadbolt N, Weitzner D (2006) A Framework for Web Science. *Foundations and Trends in Web Science*, Vol 1, No 1.

[17] Dasgupta P (2000) Trust as a commodity. Gambetta D. (Ed.). *Trust: Making and Breaking Cooperative Relations*, Blackwell, pp. 49-72.

[18] Macarthur K (2008) Tutorial: Trust and Reputation in Multi-Agent Systems. *International Conference on Autonomous Agents and Multiagent Systems* (*AAMAS 2008*), Portugal.

[19] Zacharia G, Moukas A, Maes P (2000) Collaborative reputation mechanisms for electronic marketplaces. *Decision Support Systems*, 29:371-388.

[20] Huynh T, Jennings N, Shadbolt N (2006) Certified Reputation: how an agent can trust a stranger. In *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, Hokkaido, Japan.

[21] Benjamins R, Wielinga B, Wielemaker J, Fensel D (1999) An Intelligent Agent for Brokering Problem-Solving Knowledge. *International Work-Conference on Artificial Neural Networks IWANN* (2), pp. 693-705.

[22] Governatori G, Dumas M, Hofstede A ter, Oaks P (2001) A Formal Approach to Protocols and Strategies for (Legal) Negotiation. *International Conference on Artificial Intelligence and Law (ICAIL 2001)*, pp. 168-177.

[23] Skylogiannis T, Antoniou G, Bassiliades N, Governatori G, Bikakis A (2007) DR-NEGOTIATE - A System for Automated Agent Negotiation with Defeasible Logic-based Strategies. Data & Knowledge Engineering (*DKE)*, 63(2):362-380.

[24] Antoniou G, Harmelen F van (2004) *A Semantic Web Primer*. MIT Press.

[25] Boley H.: POSL: An Integrated Positional-Slotted Language for Semantic Web Knowledge. http://www.ruleml.org/submission/ruleml-shortation.html

[26] Wang M, Purvis M, Nowostawski M. (2005) An Internal Agent Architecture Incorporating Standard Reasoning Components and Standards-based Agent Communication. In: *IEEE/WIC/ACM international Conference on intelligent Agent Technology (IAT'05)*, IEEE Computer Society, Washington, DC, pp. 58-64.

[27] Purvis M, Cranefield S, Nowostawski M, Carter D (2002) Opal: A Multi-Level Infrastructure for Agent-Oriented Software Development. In: *Information Science Discussion Paper Series*, number 2002/01, ISSN 1172-602. University of Otago, Dunedin, New Zealand.

[28] Dastani M, van Riemsdijk M B, Meyer J-J C. (2005) Programming multi-agent systems in 3APL. In: R. H. Bordini, M. Dastani, J. Dix, and A. El Fallah Seghrouchni (Eds.) *Multi-Agent Programming: Languages, Platforms and Applications*, Springer, Berlin.

[29] Nowostawski, M. (2001) Kea Enterprise Agents Documentation.

[30] Paschke A, Boley H, Kozlenkov A, Craig B (2007) Rule responder: RuleML-based Agents for Distributed Collaboration on the Pragmatic Web. *2nd International Conference on Pragmatic Web*. ACM, pp. 17-28, vol. 280, Tilburg, The Netherlands.

[31] Antoniou G, Skylogiannis T, Bikakis A, Bassiliades N (2005) DR-BROKERING – A Defeasible Logic-Based System for Semantic Brokering. *IEEE International Conference on E-Technology, E-Commerce and E-Service*, IEEE, pp. 414-417.

[32] Bartolini C, Preist C, Jennings N (2002) A Generic Software Framework for Automated Negotiation. *1st International Joint Conference on the Autonomous Agents and Multi-Agent Systems (AAMAS)*, Italy.

# Appendix A – Apartment Schema

The RDF Schema file for the broker's apartments and proposals (Section 5):

```
<!DOCTYPE rdf:RDF [ <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <!ENTITY carlo "http://lpis.csd.auth.gr/systems/dr-device/carlo/carlo.rdf#">
   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#">
   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#"> ]>
<rdf:RDF xmlns:rdf="&rdf;" xmlns:carlo="&carlo;"
  xmlns:rdfs="&rdfs;" xmlns:xsd="&xsd;">
<rdfs:Class rdf:about="&carlo;apartment" rdfs:label="apartment">
   <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdf:Property rdf:about="&carlo;bedrooms" rdfs:label="bedrooms">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;central" rdfs:label="central">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;floor" rdfs:label="floor">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;gardenSize" rdfs:label="gardenSize">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;lift" rdfs:label="lift">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;name" rdfs:label="name">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;pets" rdfs:label="pets">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;price" rdfs:label="price">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Property>
<rdf:Property rdf:about="&carlo;size" rdfs:label="size">
   <rdfs:domain rdf:resource="&carlo;apartment"/>
   <rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Property>
<rdfs:Class rdf:about="propose">
   <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdf:Property rdf:about="apartment">
   <rdfs:domain rdf:resource="propose"/>
   <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
</rdf:RDF>
```

# Appendix B – Carlo's Requirements

Carlo's requirements (Section 5) in d-POSL:

```
r1: acceptable(apartment->?x) :-
    apartment(name->?x).
r2: ~acceptable(apartment->?x) :-
    apartment(name->?x,bedrooms->?y), ?y > 2.
r3: ~acceptable(apartment->?x) :-
    apartment(name->?x,size->?y), ?y < 45.
r4: ~acceptable(apartment->?x) :-
    apartment(name->?x,pets->"no").
r5: ~acceptable(apartment->?x) :-
    apartment(name->?x,lift->"no",floor->?y), ?y > 2.
r6: ~acceptable(apartment->?x) :-
    apartment(name->?x,price->?y), ?y > 400.
r9: ~acceptable(apartment->?x) :-
    offer(apartment->?x,amount->?y),
    apartment(name->?x,price->?z), ?y < ?z.
r7: offer(apartment->?x,amount->?a) :-
    apartment(name->?x,size->?y,gardenSize->?z,central->"yes"),
    ?a is 300+2*?z+5*(?y-45).
r8: offer(apartment->?x,amount->?a) :-
    apartment(name->?x,size->?y,gardenSize->?z,central->"no"),
    ?a is 250+2*?z+5*(?y-45).
info-copy: apartment-info(apartment->?x,price->?p,
           size->?s,gardenSize->?gs) :-
               acceptable(apartment->?x),
               apartment(name->?x,price->?p,
               size->?s,gardenSize->?gs).
r2 > r1.
r3 > r1.
r4 > r1.
r5 > r1.
r6 > r1.
r9 > r1.
```

Rules $r_1$-$r_6$ express Carlo's requirements regarding the apartment specifications. Rules $r_7$ and $r_8$ indicate the offer Carlo is willing to make for an apartment that fits his needs, while rule $r_9$ ensures that the amount offered by the customer will not be higher than the apartment's actual rental price. Finally, rule `info-copy` stores all the characteristics of appropriate apartments that are of interest to Carlo, so that he can later refer to them.

# Appendix C – Broker's "Hidden Agenda"

Broker's "hidden agenda" (Section 5) in d-POSL:

```
propose(apartment->?x) :-
        acceptable(apartment->?x),
        apartment(name->?x,central->"yes",size->?s),
        \+ (acceptable(apartment->?y), ?x \= ?y,
        apartment(name->?y,central->"yes",size->?s1), ?s < ?s1).
propose(apartment->?x) :-
        acceptable(apartment->?x),
        apartment(name->?x,central->"no",gardenSize->?gs),
        \+ (acceptable(apartment->?y), ?x \= ?y,
        apartment(name->?y,central->"no",gardenSize->?gs1),
        ?gs < ?gs1).
```

The broker does not propose to Carlo all appropriate apartments, but only a subset of them, according to his "hidden agenda". The two rules depicted above are an example: the broker proposes to the customer the largest of all appropriate centrally located apartments or a non-centrally located one with the biggest garden size. Of course, the broker's hidden agenda could potentially consist of more (and possibly more adept) rules.

# Appendix D – Carlo's Preferences

Carlo's apartment preferences (Section 5) in d-POSL:

```
find_cheapest: cheapest(apartment->?x) :-
               acceptable(apartment->?x),
               apartment-info(apartment->?x,price->?z),
               \+ (acceptable(apartment->?y),
               apartment-info(apartment->?y,price->?w),
               ?x \= ?y, ?w < ?z).
find_largest: largest(apartment->?x) :-
              acceptable(apartment->?x),
              apartment-info(apartment->?x,size->?z),
              \+ (acceptable(apartment->?y),
              apartment-info(apartment->?y,size->?w),
              ?x \= ?y, ?w < ?z).
find_largestGarden: largestGarden(apartment->?x) :-
                    acceptable(apartment->?x),
                    apartment-info(apartment->?x,gardenSize->?z),
                    \+ (acceptable(apartment->?y),
                    apartment-info(apartment->?y,gardenSize->?w),
                    ?x \= ?y, ?w < ?z).
r10: rent(apartment->?x) :-
     propose(apartment->?x), cheapest(apartment->?x).
r11: rent(apartment->?x) :-
     propose(apartment->?x), cheapest(apartment->?x),
     largestGarden(apartment->?x).
r12: rent(apartment->?x) :-
     propose(apartment->?x), cheapest(apartment->?x),
     largestGarden(apartment->?x),largest(apartment->?x).
r11 > r10.
r12 > r10.
r12 > r11.
:- rent(apartment->?x), rent(apartment->?y), ?x \= ?y.
```

Carlo will choose among the apartments proposed by the broker and the ones that are compatible with his own preferences.

# Appendix E – d-POSL

*POSL* (positional-slotted language) [26] is an ASCII language that integrates Prolog's positional and F-logic's

slotted syntaxes for representing knowledge (facts and rules) in the Semantic Web. POSL is primarily designed for human consumption, since it is faster to write and easier to read than any XML-based syntax. We devised an extension to POSL, called *d-POSL*, which handles the specifics of defeasible logics and is a secondary contribution included in this work. Variables are denoted with a preceding "?". A deeper insight into core POSL, its unification scheme, the underlying webizing process (i.e. the introduction of URIs as names in a system to scale it to the Web – orthogonal to the positional/slotted distinction), and its typing conventions along with examples is found in [26].

Furthermore, d-POSL maintains all the critical components of POSL, extending the language with elements that are essential in defeasible logics:

- Rule Type: Binary infix functors are introduced (":-", ":=",":~") to denote the rule type ("strict", "defeasible", "defeater", respectively).
- Rule Label: The rule label is a vital feature in defeasible logic, since it satisfies the need to express superiorities among rules. Consequently, d-POSL employs a mechanism for expressing rule labels and superiority relationships.
- Conflicting Literals: Conflicting literals are represented as headless rules, i.e. constraints that have the following format:
  `:= predicate(?x), predicate(?y), ?x\=?y.`
  See, for example, Appendix D above.

# A Software System for Viewing and Querying Automatically Generated Topic Maps in the E-learning Domain

Liana Stanescu, Gabriel Mihai, Dumitru Burdescu, Marius Brezovan, and Cosmin Stoica Spahiu
University of Craiova, Faculty of Automation, Computers and Electronics, Romania
E-mail: {Stanescu_Liana, Burdescu_Dumitru, Brezovan.Marius, Stoica.Cosmin}@software.ucv.ro

**Keywords:** e-learning, topic maps, relational database, tolog, topic maps query

**Received:** February 8, 2010

*Topic Maps represent a recent technology for structuring and retrieval of information, based on principles used in traditional indexes and thesauri, drawing inspiration from semantic networks. In the e-learning domain Topic Maps have a greater importance as a content management technology, but also because they fill in the gap between information and knowledge. Because many e-learning systems use a relational database for storing the learning content, this paper presents an original algorithm for automated building of a Topic Map starting from a relational database. The process is illustrated on a database used in TESYS e-learning system. The paper also presents a tool with two main functions: the Topic Map graphical view that allows learner navigation for studying the topics, and associations between them and Topic Map querying using tolog that facilitates the establishing of search criteria for learning resources filtering.*

*Povzetek: Predstavljena je platforma za poučevanje s prilagajanjem vsakemu učencu posebej.*

## 1 Introduction

In the last few years, web-based e-learning has gained an increased popularity in many domains: technical, economic, medical, etc. It is well established that the e-learning platforms should offer learners interactive and flexible interfaces for access to learning resources and should also adapt easily to one's individual needs [1][2].

At the University of Craiova an e-learning system called TESYS has been created, which is used in distance learning for certain domains (economic), but also in the hybrid learning in medical and engineering domains to complete face-to-face lectures [8].

Over the few years of use there has been a tendency of passing from "course-centric" learning systems to "subject-centric" learning systems. "Course-centric" learning systems are the traditional ones which assume sequential run-over learning resources along with the lecture time schedules. In this way, learners acquire knowledge step by step in the order established by teacher. In this case, less motivated students often lose enthusiasm in the middle of the course, having difficulties in knowledge understanding [16].

This is why an e-learning system should offer students the possibility of designing their learning method in order to stay motivated. This goal can be achieved with the "subject-centric" learning systems based on Topic Maps (TM). Thus, the learners can choose their subjects and Topic Maps not only permit subjects to be visualized but also to relate with each other [16].

The paper presents an original algorithm for automated representation of a relational database with a Topic Map. This aspect is useful in the e-learning domain because many e-learning systems are based on a relational database. For example, the Moodle database has around 200 tables. The information about courses and their organization into categories is stored in the following tables: course, course_categories, course_display, course_meta, course_request. The information about activities and their arrangement within courses is stored in the next tables: modules, course_allowed_modules, course_modules and course_sections. The database structure is defined, edited and upgraded using the XMLDB system [3]. Also, the Blackboard Learning System uses a relational database for storing necessary data [14].

The proposed algorithm will be illustrated on a database used in TESYS e-learning system [8].

Human intervention can be necessary in the e-learning domain during the TM generation process. As a result, the paper presents an improvement of the automated algorithm. It assumes a configuration file that allows user to specify the part of the database that will be reflected in the content of the Topic Map. This way the size of the Topic Map is substantially reduced because it will contain only the selected items. Also the semantic content of the generated TM is improved. The generation process becomes semi-automatic because the human intervention is needed for obtaining this file.

The paper also presents a Topic Map graphical view that allows learner navigation for studying the topics and associations that shape in fact the relationships between items in the database. Associations provide the context

information necessary to better understand a topic. Associations simulate the way humans think and hence are essential for knowledge modelling. This tool also allows the learning resources filtering by establishing the search criteria in Topic Map based on tolog. To achieve this goal, the paper proposes an original graphical interface that allows the user to build interactive queries on topic map even without knowing the tolog syntax. These queries are automatically generated and sent to the tolog engine for processing.

# 2 Related work

Topic Maps represent a recent technology for the structuring and retrieval of information, based on principles used in traditional indexes and thesauri, inspired from semantic networks. Topic Maps work with topics, the relationships between topics and links to resources about those topics. Topic Maps are independent of the resources they describe and are used in many different situations. As a result, the Topic Maps can be used in information access on the Web, in reference book publishing, or in the integration of corporate information repositories [7, 9, 10]. Most applications of Topic Maps fall into four broad categories: enterprise information integration, knowledge management, e-learning, and Web publishing. In the e-learning domain, Topic Maps play an important role as a content management technology, but also due to the fact that they fill in the gap between information and knowledge [19].

There are three possibilities for the creation of TM: manually, automatically or a combination of them. In Topic Map building there are two phases: the design phase and the authoring one. Classes are created in the design phase and instances in the authoring phase [9].

Manual Topic Map population may require lots of resources: time, money and human interventions. As a result, it is considered as a weak point in process of Topic Maps self-population. The available resources that can act as a source of input to auto-population are: ontology, relational or object-oriented database, metadata about resources, index glossary, thesaurus, data dictionary, document structures and link structures or unstructured documents [9]. In related literature the authors explain only the general principles of mapping these structures to the Topic Maps model [7, 9, 19], without any specific algorithm.

A transformation of ontology into a topic map is straightforward because ontologies use the same concepts. There are several proposals as regards how to map RDF and DAML+OIL ontologies to XTM topic maps on the model level [9]. For example in [21] the authors present a topic map-driven web portal of conference papers. The paper also discusses the tools for automatically creating topic maps, with particular emphasis on how the synergies between topic maps and RDF can be exploited in the process of auto-generating topic maps from structured and semi-structured source data.

In the case of a relational database consisting of tables, columns, rows, keys and foreign keys, the mapping can be done using the following principles [9]:

Table -> topic class;

Row -> topic instance of corresponding class;

Column -> name or occurrence;

Key -> topic id;

Foreign key -> association

As we have already mentioned, one of the most important domain of successfully using the topic map concept is the e-learning. In this context we can mention papers that present interesting and modern modalities of using Topic Maps in e-learning. For example, TM4L is an e-learning environment providing editing and browsing support for developing and using Topic Maps-based digital course libraries. The TM4L functionality is enhanced by an interactive graphical user interface that combines a hierarchical layout with an animated view, coupled with context sensitive features [4, 5]. In Norway, school students are encouraged to create Topic Maps to record what they have learned, and the National School Curriculum itself now has its definitive expression in the form of a Topic Map [19].

Another author proposed Topic Map ontology focusing on both students and teachers as active producers of learning resources. Topic maps customize the interface, and the interface should also provide possibilities for online students to share learning resources like “on campus” students do [6].

In [15] we proposed original ways of using Topic Maps in medical e-learning. The TM is used for visualizing a thesaurus containing medical terms. The paper presents also the way in which the TM can be used for semantic querying a multimedia database with medical information and images.

There are some available TM authoring tools, but they are used by experts in knowledge representation, not by end-users (Ontopia Knowledge Suite [12], Mondeca Intelligent Topic Manager [13]). Also, there are few specialized education-oriented TM tools that can be used to facilitate the creation, maintenance, search, and visualization of Topic Maps-based learning resources.

Along with the Topic Map graphical visualization searching the information that satisfies a number of criteria is also important. A solution for querying a Topic Map is represented by the tolog language.

Tolog is a language for querying Topic Maps, inspired by Datalog (a subset of Prolog) and SQL. With tolog users can ask for all topics of a particular type, the names of all topics of a particular type, all topics used as association role types, all associations with more than two roles, and so on [17].

Tolog is a logic-based query language, which means that the basic operation consists of asking tolog in which cases a certain assertion holds true, and tolog will return all the sets of values that make the assertion true.

As a result, the paper presents a software system with the following contributions:

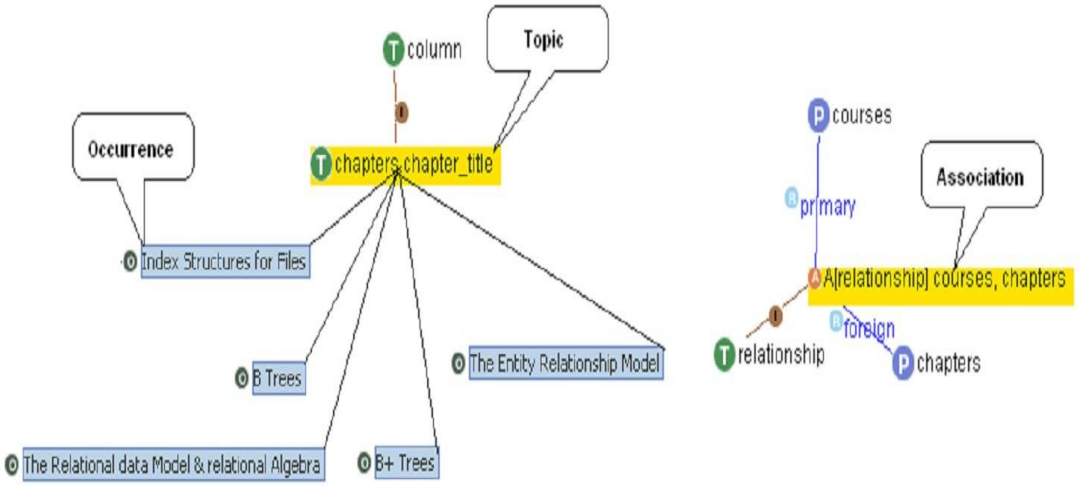



Figure 1: The basic concepts in a topic map.

1. An original algorithm of mapping any relational database to a topic map. The algorithm steps are exemplified on TESYS database used in the e-learning domain.
2. An improvement of this automated algorithm based on a configuration file that allows user to specify the database parts that will be reflected in the TM content.
3. A graphical interface for TM querying using tolog language. Users do not need to know tolog syntax and this aspect is important in the e-learning domain.

## 3 Topic Maps basic concepts

Topic Maps represent an ISO standard that provides concepts for describing knowledge and linking it it to information resources. Topic Maps allow organizing and representing complex structures with the next basic concepts [7, 9, 10]: topics, occurrences of a topic, associations that connect related topics, topic classes, occurrence classes and association classes that help to distinguish different kind of topics, occurrences and associations, respectively (figure 1).

A topic can be defined as a syntactic construct that represents a subject inside a computer system. Through the agency of it, the real-world concept becomes a machine interpretable object. A topic can be an instance of zero, one or more classes and the classes itself are also topics. A topic has three characteristics:

- Names (base names)
- Occurrences
- Playing roles in associations

Occurrences link information resources to topics. It is important to notice that topics and information resources are placed on different layers and the users may navigate at the abstract layer that is the topic layer rather than within data. In fact, occurrences bind real resources to topics that are abstract concepts. An occurrence is either a resource reference or a resource data. A resource reference links the relevant resource to the topic using XLink/XPointer URI. A resource data occurrence assigns a value to the topic [7, 9, 10].

The relationships between concepts are possible in Topic Maps through topic associations. The associations provide the context to better understand topics. They simulate the human thinking and so are important for knowledge modelling.

Topic Maps can improve navigation and information retrieval in large and complex information pools by adding semantics to these resources.

## 4 The system architecture

The system architecture is presented in figure 2. TMGenerator is the module that generates the content of the Topic Map using the information from the database

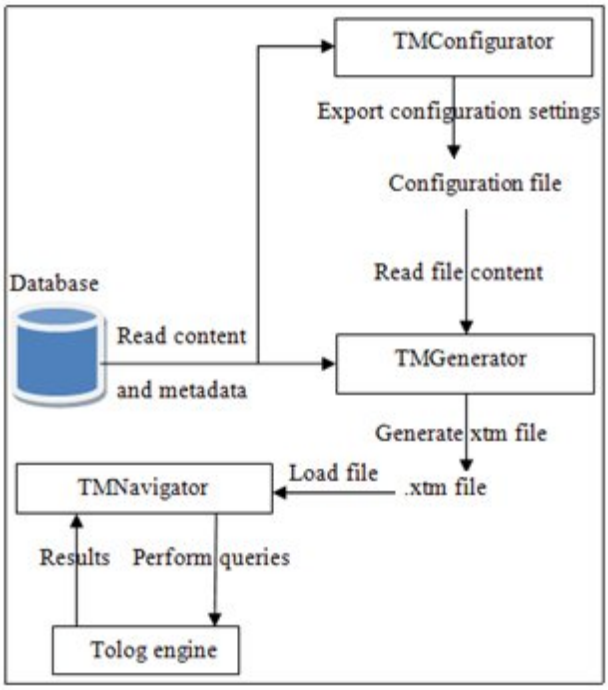



Figure 2: The system architecture.

and from a configuration file (when it is available). The content of a generated Topic Map is saved by the TMGenerator as an .xtm file.

TMConfigurator is the module used for making configurations. It uses the structure of the database and allows the user to select the tables and the columns, but also to specify the topics, the topics type, the associations and the association' types that will be generated. The settings are exported in an .xml file that will be used by the TMGenerator when generating the Topic Map. The user has the possibility to represent in the TM only the necessary information from the database, thus reducing the size of the Topic Map.

TMNavigator is the name of a graphical application used for exploring the content of a Topic Map and for running tolog queries. These queries are analysed by the tolog engine and the results are sent back.

The functions of these modules will be explained in detail in the next sections.

# 5 TESYS database structure

Figure 3 illustrates a part of the relational database used by the e-learning system called TESYS [8]. This database will be used later to better explain the Topic Map automated building and also it's graphical view.

The table named Courses stores data about electronic courses, each course being equivalent to a unit of curriculum or an academic subject in traditional learning. Usually, a course contains many chapters, and each chapter contains a number of topics that represent in fact learning objects. Each topic represents one unit of knowledge, being the smallest component. The learning object can be a piece of text, a video clip, a picture or a voiced text.

In this database structure the relationships between topics studied at the same course or different courses are important. If a topic uses knowledge presented in other topics, these topics must be linked. As a result, on Topics table a m:m recursive relationship is defined. This special relationship is implemented with Topic_connection table.

# 6 The algorithm for building the Topic Map starting from a relational database

The Topic Map generation process is a general one and it is realized by the module TMGenerator. This module uses the settings that were specified for accessing the database and for retrieving the content and the metadata (table's name, column's name, etc.). TMGenerator was designed to automatically generate the content of the Topic Map taking into account the xtm syntax [11]. When the generation process is completed the content is exported as an .xtm file that can be imported by the module TMNavigator for a graphical representation of the Topic Map. Initially, the TM generation process builds a topic type array: "database", "table", "row", "column", "relationship", "part", "whole", "primary", "foreign" that will be used as basic types for topics and associations.

From database meta-data the algorithm uses the following elements:

- The array with table names
- The relationships between tables

The general algorithm implemented in TMGenerator for generating a Topic Map stores internally (in memory) the content using the following data structures:

- *topicMap* – a structure that stores the content of a

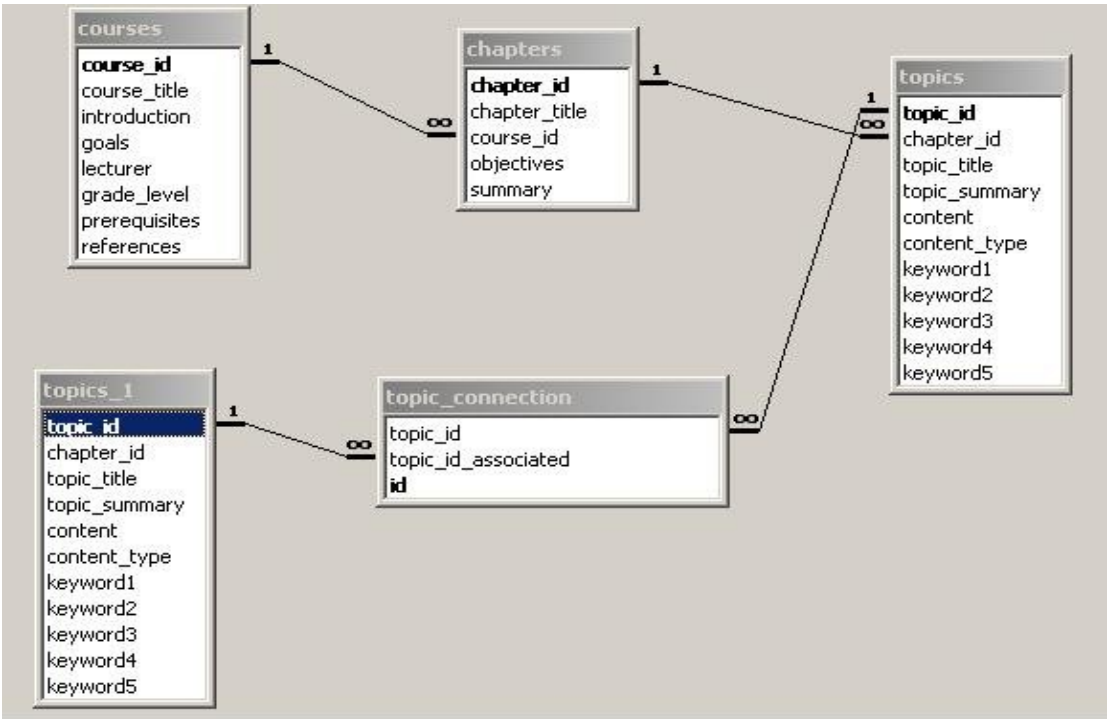



Figure 3: TESYS database structure.

Topic Map using 3 lists: a list of topics, a list of associations and a list of occurrences

- *topic* – a structure containing the information (id, instanceOf , baseName) about a topic item
- *association* – a structure containing the information (id, instanceOf, members) about an association item
- *occurrence* – a structure containing the information (instanceOf, resourceData) about an occurrence item

The algorithm uses the following procedures:

- GenerateTopic – used to generate a *topicMap* structure representing the content of a Topic Map
- GenerateTopic – used to generate a *topic* structure
- GenerateOccurrence – used to generate an *occurrence* structure
- GenerateAssociationForTables – used to generate an *association* structure representing the relationship between two tables
- GenerateAssociation – used to generate an *association* structure

The content of these procedures is presented bellow.

Procedure GenerateTopicMap()
a) initialize topicMap;
b) for * each topic type in topicTypes
topic <− GenerateTopic(id,instanceOf, baseName);
topicMap.Topics.Add(topic);
end;
c) databaseTopic <− GenerateTopic(dbName, “database”, dbName);
d) topicMap.Topics.Add(databaseTopic);
e) for * each table in database
tableTopic <− GenerateTopic(tblName, ”table”, tblName);
topicMap.Topics.Add(tableTopic);
end;
f) topics <− generate topics for table columns and rows
g) topicMap.Topics.Add(topics);
h) occurrences <− generate occurrences
i) topicMap.Occurrences.Add(occurrences);
j) for * relationship between tables
association <− GenerateAssociationForTables( params);
topicMap.Association.Add(association);
end;
k) association <− GenerateAssociation(params);
l) topicMap.Association.Add(association);
m) for *each table and his rows
association <− GenerateAssociation(params);
topicMap.Association.Add(association);
end;
n) for * each record involved in a 1:m relationship
association <− GenerateAssociation(params);
topicMap.Association.Add(association);
end;
return topicMap;

Procedure GenerateTopic(topicId,instanceOf,baseName)
a) initialize topic;
b) topic.Id <− topicId;
c) topic.InstanceOf <− instanceOf;
d) topic.BaseName <− baseName;
e) return topic;

Procedure GenerateOccurrence(instanceOf,resourceData)
a) initialize occurrence;
b) occurrence.InstanceOf <− instanceOf;
c) occurrence.ResourceData <− resourceData;
d) return occurrence;

Procedure GenerateAssociationForTables(associationId, instanceOf,primarySpecificRole,primaryTopicReference, foreignSpecificRole,foreignTopicReference)
a) initialize tblAssociation;
b) tblAssociation.Id <− associationId;
c) tblAssociation.PrimaryMember.SpecificRole <− primarySpecificRole;
d) tblAssociation.PrimaryMember.TopicReference <− primaryTopicReference;
e) tblAssociation.ForeignMember. SpecificRole <− foreignSpecificRole;
f) tblAssociation.ForeignMember.TopicReference <− foreignTopicReference;
g) return tblAssociation;

Procedure GenerateAssociation(associationId,instanceOf, members)
1) initialiaze association;
2) association.Id <− associationId;
3) association.InstanceOf <− instanceOf;
4) for * each member in members
5) association.Members.Add(member);
6) end;
7) return association;

The Topic Map generation process starting from the ELearning database is detailed bellow and has the following steps:

1. Topic generation for database and tables

The database is represented by the algorithm as a topic that is an instance of the topic database. This topic has an id that contains the database name.

Example: For ELearning database is used the next procedure:
GenerateTopic(“ELearning”,”database”,”ELearning”)

In the same way, the algorithm creates a topic for each table in the database. The topic id is given by table’s name, because this is unique in the database. For the database in figure 3 the topics for tables are generated: courses, chapters, topics, topics1, topic_connection.

Example: For table Courses is used the procedure: GenerateTopic(“courses”,”table”,”courses”)

2. Topic generation for table columns and records

The algorithm will generate a topic for each column in a table. The topic id is given by the next syntax: Table Name.ColumnName. This is our choice because each topic must have a unique id.

Example: The procedure used for column course_title in table Courses is: GenerateTopic("courses.course_title","column", "course_title").

For each table record the algorithm generates a topic that has the following format for the id: Table Name.Row.PrimaryKeyValue. The record content is considered as topic occurrence.

Example: For the record with primary key value 3 in table Chapter is used the procedure:
GenerateTopic("chapters.Row.3","row", "chapters.Row.3")

One of the columns in table Chapters is chapter_title. For this column content an occurrence is created using the procedure:
GenerateOccurrence("chapters.chapter_title","B Trees").

3. Associations generation process

3.1 Associations corresponding to relationships between tables

For each relationship in the database it is generated an association of type "relationship". The id of each association is based on the tables' names, the primary key and the foreign key. This development mode takes into consideration to offer information about the database structure for facilitating the learning process.

Example: For the relationship 1: m between tables Courses and Chapters an association with the next id is generated: "courses.course_id-chapters.course_id". This association is an instance of the topic relationship. In this association, the table courses contains the primary key and plays the role "primary" and the table chapters containing the foreign key plays the role "foreign". This association is generated by the procedure:
GenerateAssociationForTables("courses.course_id-chapters. course_id", "relationship","primary","courses", " foreign", " chapters").

3.2 Association between database and tables

The association between database and its tables is of the "part-whole" type. The topic representing the database plays the role "whole" and every topic representing a table plays the role "part".

Example: The association between the topic representing the database ELearning and topics representing the tables (courses, chapters, topics, topics1, topic_connection) is generated using the procedure:
GenerateAssociation ("Database: ELearning.Tables", "part-whole", members)

In this syntax, "members" is a list with member items, each member containing two fields: the role played in the association and the id of the corresponding topic.

3.3 Associations between table and records

The fact that a table contains records is represented by an association of "part-whole" type between table and its records.

Example: For table courses an association is generated using the next procedure:
GenerateAssociation( "Table:courses.Rows" ,"part-whole", members).

In this association the topic representing the table Courses plays the role "whole" and every topic representing a record plays the role "part".

3.4 Associations between records involved in a 1:m relationship

This association is of the "related-to" type. In order to be generated, for every value of the primary key, the records that contain the same value in the foreign key column must be founded. As a result, this association is established between the topics already generated for these records.

Example: Tables Courses and Chapters are defined by a 1:m relationship. The course entitled "Databases" contains 3 chapters stored in the table Chapters. This fact is represented by an association of the "related-to" type between the topic representing the corresponding record in the table courses and the topics representing connected records from the table chapters. Every topic plays the role "related".

The generation process is a generic one and it generates topics for each row, for each column value and a big number of associations. Other details of this process can be found in [20].

# 7 Improving the topic map building process

The process described in section 6 is a general one. It can be applied to mapping any relational database to a Topic Map. After a number of experiments we have decided to improve this process in order to have a better representation of the semantic content and to export only the necessary information from the database, thus reducing the size of the Topic Map. In the e-learning domain it is very important for the user to understand without a big effort the Topic Map content, otherwise the learner can be disoriented, especially when the Topic Map size is large.

By human intervention it is possible to generate a Topic Map that represents better the knowledge from the database. We have decided to complete the general algorithm by using configuration settings. For this purpose we have created an application that has a graphical interface called TMConfigurator. It retrieves the structure of a database and allows the user to select the tables and the columns, but also to specify the topics, the topics type, the associations and the association' types that will be generated. The settings are exported in an .xml file that will be used by the TMGenerator when generating the Topic Map. For the structure of the TESYS database used in this paper, an example of content from the configuration file is presented in figure 4.

Because the human intervention is needed for obtaining this file the generation process becomes semi-automatic.

# 8 Topic Map Graphical View

Topic map content is explored using the graphical interface with multiple views of the TMNavigator

```
<? xml version="1.0" encoding="UTF-8"?>
<configuration>
 <topics>
  <topic table_column="Courses.Course"
    topic_type="course"/>
  <topic table_column="Subchapters.Subchapter"
    topic_type ="subchapter">
    <occurrence table_column="LearningObjects.Name"
        type="resource"/>
  </topic>
 </topics>
 <associations>
  <association type="has_chapters">
    <member primary_table_column="Courses.Course"
        role_type="whole"/>
    <member foreign_table_column="Chapters.Chapter"
        role_type="part"/>
  </association>
  <association type="has_subchapters">
    <member primary_table_column="Chapters.Chapter"
        role_type="whole"/>
    <member foreign_table_column="Subchters.Subchapter"
        role_type="part"/>
  </association>
 </associations>
</configuration>
```

Figure 4: The configuration file.

module. TMNavigator has a menu with two options (figure 5): File and Search. Using the File option the user can load a generated .xtm file (representing the content of a Topic Map) and then explore the content. The Search option can be used by the user to run tolog queries. This way the user can find the information he is looking for.

The viewing interface for Topic Maps is organized in two windows: the left window displays a list of all topics, topic types, associations, association types, occurrence types and member types. As a topic type we have: column, database, row, table, etc.

In Topic Map there are multiple association types: "part-whole" (defined between the topics that represent database and tables or between topics representing a table and its records), "relationship" (defined between the topics representing tables implied in a 1:m relationship), "related-to" (defined between the topics representing the table records bound with a 1:m relationship).

The learner can select an item from the list displayed in the left window and he will see in the right window the item's graphical representation. Topic map viewing tool intends to offer learner much information about the selected item. Unlike TM4L, our viewing tool displays for each topic involved in an association its occurrence content also. Example: for the topic that represents a record in Courses table, the learner will see information like: lecturer, grade_level, introduction, etc. Topic content can be visualized separately by selecting it in the left window.

Another original element in this graphical window is that the learner can see directly the record content involved in 1:m relationship implemented in Topic Map by "related-to" association. Beside these associations viewing that offer better understanding to the topic, the learner can go directly to study the associated topic.

Figure 5 presents details of a "related-to" type association. Figure 6 contains a "part-whole" type association.

The users can use the Topic Map as a navigation tool with which they can navigate through Topic Map depending on their subject of interest and this is a big advantage. They don't have to be familiar with the logic of the database, they will learn about the semantic context, in which a collection and its single items are embedded and they may find useful items that they would not have expected to find.

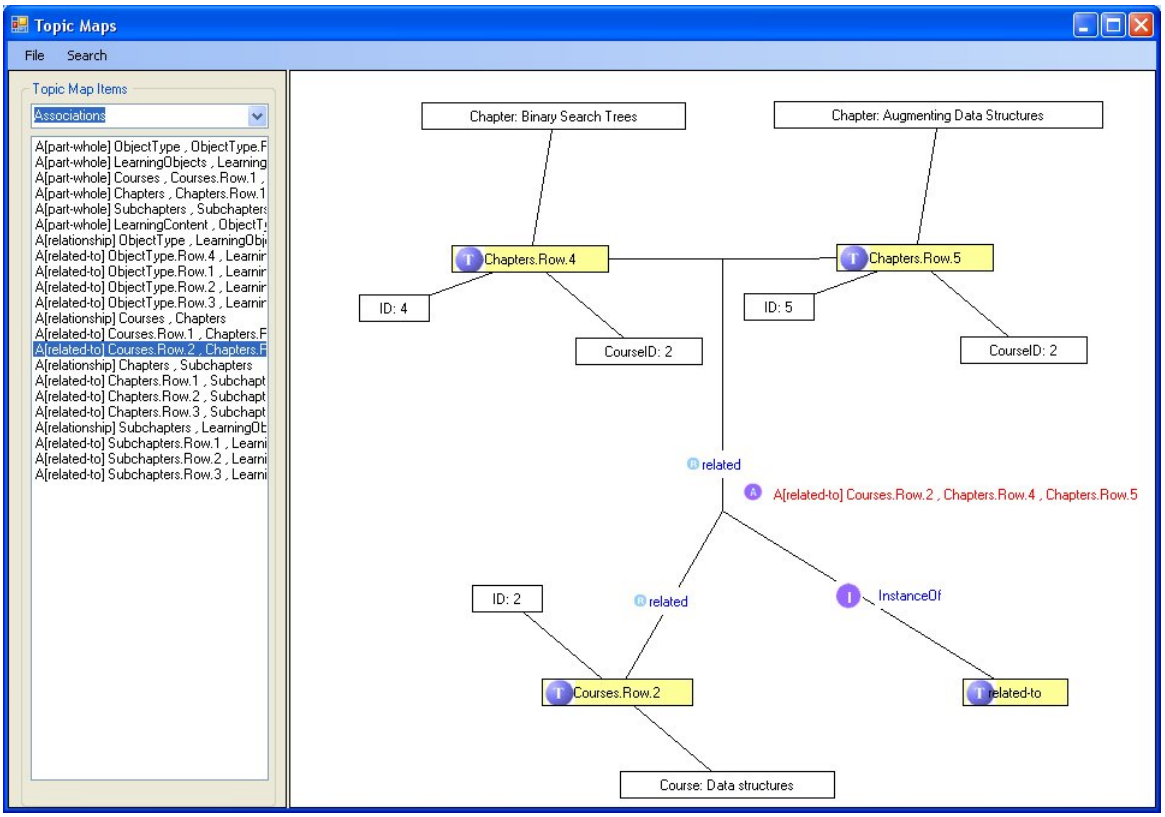



Figure 5: An association of type "related-to".

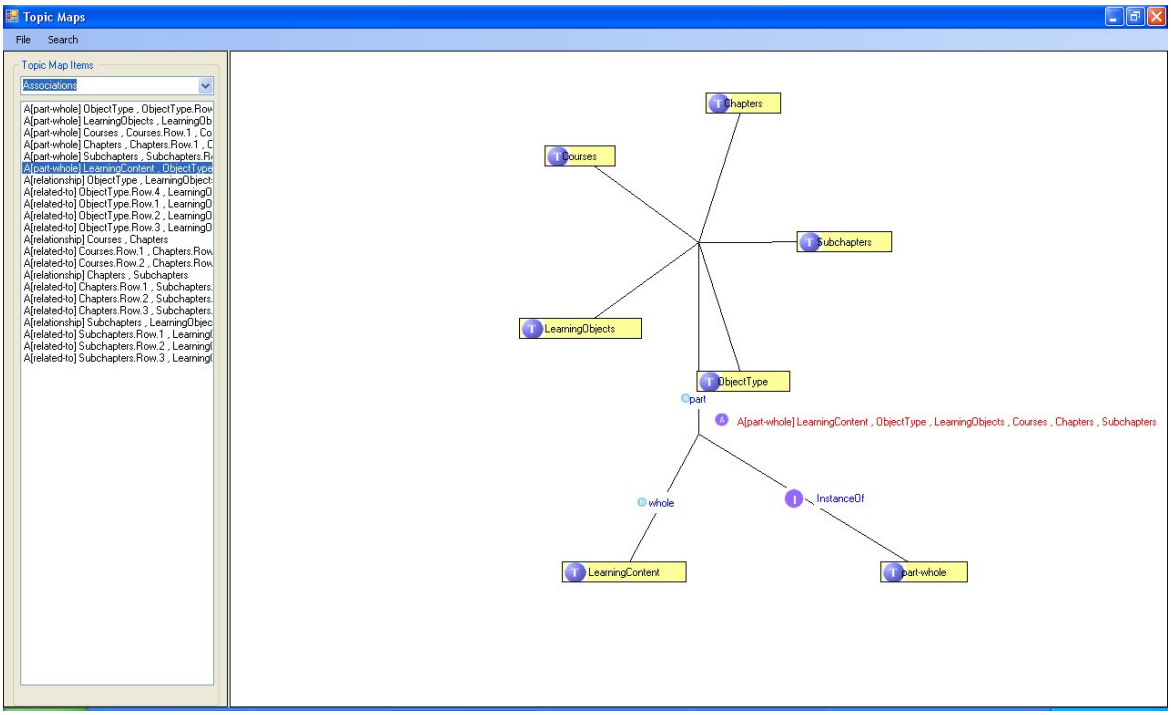



Figure 6: An association of type “part-whole”.

## 8 Topic Map querying with tolog query language

Along with the Topic Map graphical visualization searching the information that satisfies a number of criteria is also important. A solution for querying a Topic Map is represented by the tolog language. In our experiments we have used the tolog engine provided by Ontopia [17].

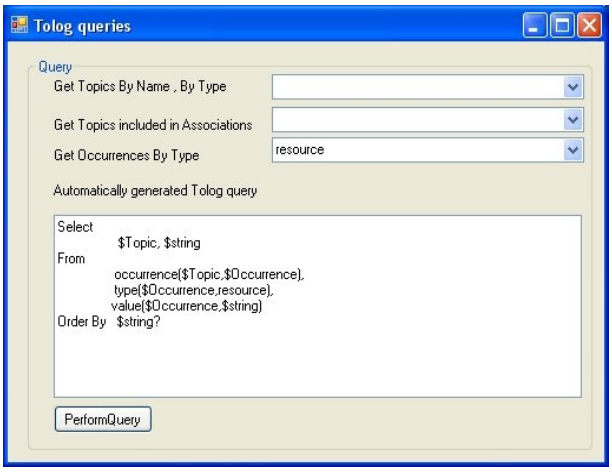



Figure 7: Graphical interface for tolog queries.

For running queries in this way, two operations are necessary:

- loading in engine the generated Topic Map
- building an interface for sending tolog queries to the engine

Tolog’s syntax is a complex one, normally used by experimented users and improper for the e-learning domain.

This is the reason why we propose an original graphical interface for running tolog queries. These queries are automatically generated as a result of user’s choice: he/she can select topics, topic types, associations or occurrence types (figure 7).

The user has the possibility to select a topic type from the available list of topics or the option “All topics” that will return a list with all topics defined in the Topic Map. For example, for the topic type “course”, the application generates the tolog query from the table 1 point a. After the execution, the result is a list with all topics that represent courses.

For example, for the association “has_chapters”, the application generates the tolog query from the table 1 point b. The result is the list of all topics involved in this type of association.

In some cases, a topic can have a number of occurrences. In the TESYS database, a subchapter has learning objects (image file, text file, etc.). All these are defined using the type “resource”. With the graphical interface, the learner can select an occurrence type, and the application generates a tolog query as in the table 1 point c.

Since the results from the tolog engine contain the topic’s name this information is used as a link. When a user clicks on a topic name the TMNavigator searches that name in the list of all topics (kept in memory after loading the xtm file). After the topic is found his details are automatically shown in the right window. If the results contain resource’s name (for example Projection.ppt) the TMNavigator generates a link that

can be used by the user to open/download the content of that file.

Table 1: Automatic generated queries and results.

| | Query | Obtained results (examples) | |
|---|---|---|---|
| a) | instance-of ($Topic, course)? | **Course** | |
| | | Databases | |
| | | Data structures | |
| b) | has_chapters ($Course: whole, $Chapter: part)? | **Course** | **Chapter** |
| | | Data structures | Binary Search Trees |
| | | Data structures | B-Trees |
| | | Databases | The Entity Relationship Model |
| | | Databases | Relational Algebra |
| c) | Select $Topic, $string From Occurrence ($Topic, $Occurrence) Type ($Occurrence, resource) Value ($Occurrence, $string) Order By $string? | **Topic** | **Title** |
| | | What is a binary search tree? | BinarySearchTree.ppt |
| | | Searching a B-tree | SearchBTree.ppt |
| | | Projection operator | Projection.ppt |

# 9 Users' feedback

A number of 60 students participated to the following experiment: they were asked to study the discipline Database using TESYS system, an on-line e-learning platform that uses a tree structure for displaying the learning content (the learner chooses the course, then a chapter, and finally a lesson). The existing relationships between learning objects are implemented as hyperlinks. The student can also use some search criteria. At the same time, they had to study the discipline Multimedia Applications Development using the Topic Map created with this software tool. The students' opinion over these two learning modalities was recorded with a number of questions presented in table 2.

Table 2: Questionnaire.

| Question | The number of "Yes answers |
|---|---|
| Is it easy to find the necessary subject in TM? | 45 |
| Is it easy to find associated subjects in TM? | 51 |
| Does TM enable the study in a free way? | 35 |
| Does TM keep you motivated? | 48 |
| Is TM easy to use? | 41 |
| Is the subjects order in the course difficult to understand using TM? | 48 |
| Is it easy to understand the whole structure of knowledge using TM? | 55 |
| Is the Topic Map querying process useful and fast? | 53 |
| Are there too many subjects displayed in TM? | 34 |
| Do you favour TM as a learning modality? Why? | 45 |
| Do you think that these 2 learning modalities complete one another? | 58 |

The students' answers emphasized the fact that the use of Topic Maps in the e-learning field presents positive aspects: they are easy to use and the student can easily pick a subject and see the relationships between subjects. Also, they found very useful the Topic Map querying process using tolog.

The results showed that many students prefer the new learning modality based on Topic Maps due to the following reasons:

- the graphical interfaces are more attractive and modern
- TM offers more dynamism in following the knowledge semantic relationships
- there are no limits in building queries (the user can specify many conditions and combine them in a free way) as in the TESYS system

It is very interesting to note that a large number of students (96%) consider that these two learning modalities can be used together, because they can complete one another.

# 10 Conclusions

The e-learning systems must contain powerful and intuitive tools for viewing the learning resources, for browsing the lessons or topics and relationships between them, and also for searching the relevant information. An important feature of an e-learning system is the presentation way of the semantic relationships between topics using an appropriate navigational structure.

This aim can be achieved using a modern concept - Topic Map. Topic Maps are an emerging Semantic Web technology that can be used as a means of organizing and retrieving information in e-learning repositories in a more efficient and meaningful way.

As a result, the paper presents the following important aspects:

1. The algorithm for Topic Map automated building starting from a relational database, which is not possible with existing Topic Maps software. This aspect is useful because there are many e-learning systems that store the educational content in a database.

2. An improvement of this algorithm with the help of a configuration file that is automatically generated as a result of user choices. This way, the Topic Map represents in a more appropriate manner the semantic content in the e-learning domain.

3. A graphical view that allows Topic Map navigation is useful in studying topics that represent in fact learning objects and associations between them.

4. A graphical window for building tolog queries on the Topic Map. This window allows learners to filter the information based on their interest in an interactive way.

The students have found this new modality of knowledge visualization and filtering useful, modern and interesting.

## References


[1] Rosenberg, M. (2001). E-Learning: Strategies for Delivering Knowledge in the Digital Age. In: New York, McGraw-Hill

[2] Wentling, T., Waight, C., Gallaher, J., La Fleur, J., Wang, C., Kanfer, A. (2000). E-Learning: A Review of Literature, http://learning.ncsa.uiuc.edu/-papers/elearnlit.pdf

[3] Moodle (2009). http://docs.moodle.org/en/-Development:Database_schema_introduction

[4] Dicheva, D., Dichev, C., Dandan, W. (2005). Visualizing topic maps for e-learning. *Advanced Learning Technologies (ICALT 2005)*, IEEE Computer Society, 950 – 951

[5] Dandan, W., Dicheva, D., Dichev, C., Akouala, J. (2007). Retrieving information in topic maps: the case of TM4L. *ACM Southeast Regional Conference*, pp.88-93

[6] Kolås, L. (2006.) Topic Maps in E-learning: An Ontology Ensuring an Active Student Role as Producer. *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Ed/ITLib Digital Library, Association for the Advancement of Computing in Education (AACE), pp. 2107-2113

[7] Garshol, L.M. (2004). Metadata? Thesauri? Taxonomies? Topic Maps! *Journal of Information Science*, vol. 30, no. 4, pp.378-391, CILIP (Chartered Institute of Library and Information Professionals), ISSN 0165-5515

[8] Stanescu, L., Mihaescu, M.C, Burdescu, D.D, Georgescu, E., Florea, L. (2008). An Improved Platform for Medical E-Learning. *Lecture Notes in Computer Science 4823*, Springer, pp.392-403

[9] Rath, H. (2003) *The Topic Maps Handbook*. Empolis GmbH, Gutersloh, Germany

[10] TopicMaps. Org, http://www.topicmaps.org/

[11] XML Topic Maps (XTM) 1.0. http://topicmaps.org/xtm/1.0/index.html

[12] Ontopia (2009). http://www.ontopia.net/

[13] Mondec,(2009). http://www.mondeca.com/

[14] Blackboard (2009). http://www.blackboard.com/-Teaching-Learning/ Learn-Platform.aspx

[15] Stanescu, L., Burdescu, D., Mihai, G., Ion, A., Stoica,C. (2008). Topic Map for Medical E-Learning, *Studies in Computational Intelligence*, Springer-Verlag, Vol. 162/2008, pp.305-310

[16] Matsuura, S., Naito, M. (2008). Creating a Topic Maps based e-Learning System on Introductory Physics, *Leipziger Beitrage zur Informatik*: XII, pp.247-260

[17] Tolog language tutorial, Ontopia, (2007) http://www.ontopia.net/omnigator/docs/query/tutorial.html

[18] Omnigator, (2010) http://topicobserver.com:8080/-omnigator/models/index.jsp.

[19] Pepper S. (2009) Topic Maps. *Encyclopedia of Library and Information Sciences*, Third Edition DOI: 10.1081/E-ELIS3-120044331

[20] Mihai, G., Stanescu, L., Burdescu, D, Brezovan, M., Stoica Spahiu, C. (2009) A Topic Map for "subject-centric" Learning. *Studies in Computational Intelligence* 237/2009, Springer, pp.141-150

[21] Pepper, S., Garshol, L. M. Lessons on Applying Topic Maps. http://www.ontopia.net/topicmaps/-materials/xmlconf.html#General-Principles

# Accommodating Learning Styles in an Adaptive Educational System

Elvira Popescu, Costin Badica and Lucian Moraret
University of Craiova, A.I.Cuza 13, 200585 Craiova, Romania
E-mail: popescu_elvira@software.ucv.ro, badica_costin@software.ucv.ro

**Keywords:** intelligent e-learning application, adaptive educational system, learning style, learner modeling

**Received:** May 11, 2010

*Integrating learning styles in adaptive educational systems is a relatively recent trend in technology enhanced learning. The rationale is that adapting courses to the learning preferences of the students has a positive effect on the learning process, leading to an increased efficiency, effectiveness and/or learner satisfaction. The purpose of this paper is twofold: i) to provide an extensive review of existing learning style-based adaptive educational systems (LSAES); ii) to propose an innovative system (called WELSA), which alleviates some of the encountered limitations. Specifically, WELSA is based on: i) a comprehensive set of learning style preferences; ii) an implicit and dynamic learner modeling method; iii) a dynamic adaptation approach. The system's architecture is presented, together with the main components responsible for its functionalities: authoring tool, data analysis tool and adaptation component. Encouraging experimental data are also reported.*

*Povzetek: V prispevku je podan pregled sistemov za učenje, ki se prilagajajo učencu, in nov sistem WELSA.*

## 1 Introduction

An important class of intelligent applications in e-learning are the adaptive ones, namely those that aim at individualizing the learning experience to the real needs of each student. The rationale behind them is that accommodating the individual differences of the learners (in terms of knowledge level, goals, learning style, cognitive abilities, etc.) is beneficial for the student, leading to an increased learning performance and/or learner satisfaction. A common feature of these systems is that they build a model of learner characteristics and use that model throughout the interaction with the learner [3]. An adaptive system must be capable of managing learning paths adapted to each user, monitoring user activities, interpreting them using specific models, inferring user needs and preferences and exploiting user and domain knowledge to dynamically facilitate the learning process [4].

The idea dates back to 1995-1996, when the first intelligent and adaptive Web-based educational systems (AIWBES) were developed [3]. Since then, both the intelligent techniques employed evolved and the range of learner characteristics that the systems adapt to expanded. A relatively recent characteristic that has started to be taken into account is the learning style of the student, i.e., the individual manner in which a person approaches a learning task, the learning strategies activated in order to fulfill that task. More formally, learning styles represent a combination of cognitive, affective and other psychological characteristics that serve as relatively stable indicators of the way a learner perceives, interacts with and responds to the learning environment [16].

For example, some learners prefer graphical representations and remember best what they see, others prefer audio materials and remember best what they hear, while others prefer text and remember best what they read. There are students who like to be presented first with the definitions followed by examples, while others prefer abstract concepts to be first illustrated by a concrete, practical example. Similarly, some students learn easier when confronted with hands-on experiences, while others prefer traditional lectures and need time to think things through. Some students prefer to work in groups, others learn better alone. These are just a few examples of the many different preferences related to perception modality, processing and organizing information, reasoning, social aspects, etc., all of which can be included in the learning style concept [24].

This paper deals with an intelligent learning environment that adapts to the learning style of the students, as its name suggests: WELSA - ***W***eb-based ***E***ducational system with ***L***earning ***S***tyle ***A***daptation. We start, in section 2, with an extensive review of related works, overviewing the adaptation techniques, as well as the modeling methods employed. Next, we introduce our innovative system, WELSA, based on: i) a comprehensive set of learning style preferences; ii) an implicit and dynamic learner modeling method; iii) a dynamic adaptation approach. The system architecture is presented in section 3, as well as an example of the platform at work. The following 3 sections present in more detail the main components responsible for the system's functionality: authoring tool (section 4), modeling component (section 5) and adaptation component (section 6). Finally, some conclusions are drawn in section 7.

## 2 Related works

In what follows, we will provide a summary of the state-of-the-art LSAES, classified from the point of view of the ***adaptation methods*** offered by these systems. Some of them combine adaptation provisioning based on several criteria: learning styles, knowledge level, goals, etc.; however, in what follows, we are only interested in the adaptation techniques used for learning style personalization. One of the most widely used techniques is the so-called ***fragment sorting*** [2], i.e., presenting the educational resources in an order considered most suitable for each student. So, basically, all the students are presented with the same learning resources, just ordered differently. This approach is used in several works, such as:

- [5] → The adaptation criteria in the CS383 system are represented by 3 constructs of the Felder-Silverman model (FSLSM) [9]: Sensing/Intuitive, Visual/Verbal, Sequential/Global. For each category of resources (i.e., hypertext, audio files, graphic files, digital movies, instructor slideshows, lesson objectives, note-taking guides, quizzes, etc.), the teacher has to mention its suitability (support) for each learning style (by rating it on a scale from 0 to 100). When a student logs into the course, a CGI executable loads the student profile (i.e., his/her learning style as resulted from answering a dedicated questionnaire); it then computes a unique ranking of each category of resources, by combining the information in the student's profile with the resource ratings. Next, the CGI dynamically creates an HTML page containing an ordered list of the educational resources, from the most to the least effective from the student's learning style point of view.
- [19] → The adaptation criteria in the INSPIRE system include the 4 learning styles in Honey and Mumford model [13]: Activist, Pragmatist, Reflector and Theorist. All learners are presented with the same knowledge modules, but their order and appearance (either embedded in the page or presented as links) differ for each learning style. Thus for Activists (who are motivated by experimentation and challenging tasks), the module "Activity" appears at the top of the page, followed by links to examples, theory and exercises. In case of Pragmatists (who are motivated by trying out theories and techniques), the module "Exercise" appears at the top of the page, followed by links to examples, theory and activities. Similarly, in case of Reflectors the order of modules is: examples, theory, exercises, and activities, while in case of Theorists the order is: theory, examples, exercises and activities. The system offers also the students the possibility to choose their preferred order of studying.
- [12] → The adaptation criteria are represented by three FSLSM dimensions (Active/Reflective, Sensing/Intuitive, Sequential/Global). The authors propose an add-on for Moodle Learning Management System [18], which supplies the required adaptation. More specifically, it provides an individualized sequence and number of learning objects of each type (i.e., examples, exercises, self assessment tests, content objects).

Another adaptation technique is to ***customize the system's interface*** according to students' preferences. This technique is used for example in [6]. The adaptation criterion is represented by the Felder-Silverman learning style model. The interface is adaptively customized: it contains 3 pairs of widget placeholders (text/image, audio/video, Q&A board/Bulletin Board), each pair consisting of a primary and a secondary information area. The space allocated on the screen for each widget varies according to the student's FSLSM learning style: e.g., for a Visual learner the image data widget is located in the primary information area, which is larger than the text data widget; the two widgets are swapped in case of a Verbal learner. Similarly, the Q&A Board and Bulletin Board are swapped in case of the Active versus Reflective learners.

A similar approach is used by [1]. However, besides layout customization, they also alter the sequencing and structure of the learning content, as well as the navigation options. The adaptation criterion is represented by the FSLSM Sequential / Global preference. The pages for Global students contain diagrams, table of contents, overview of information, summary, while pages for Sequential learners only include small pieces of information, and Forward and Back buttons.

A more complex adaptation approach is employed by [30]. They use both adaptive presentation technique and adaptive navigation support to individualize the information and the learning path to the field dependence (FD)/field independence (FI) characteristic of the students [32]. Specifically, the AES-CS system uses ***conditional text*** and ***page variants*** to present the information in a different style: from specific to general in case of FI learners (who have an analytic preference) and from general to specific in case of FD learners (who have a global preference). AES-CS offers also two control options: program control for FD learners, by means of which the system guides the learner through the learning material; learner control for FI learners, by means of which the learners can choose their own learning paths, through a menu. Since FD learners benefit more from instructions and feedback, an additional frame at the bottom of the page is used to provide them with explicit directions and guidance. This frame is missing in case of FI learners, who prefer few instructions and feedback. Similarly, in case of self-assessment tests, the feedback provided for FI learners is less extensive than in case of FD learners. Finally, FD learners are offered two navigational tools in order to help them structure the learning material and create the big picture: a concept map (a visual representation of the domain concepts and the relations between them) and a graphic path indicator (presenting the current, the previous and the next topic). Furthermore,

AES-CS allows students to modify the adaptation options provided by the system, making their own choices between program / learner control, minimal / maximal feedback, etc.

Another approach is the ***adaptive selection of learning objects***, among the set of equivalent ones (from the point of view of the domain concept that they explain). The learning object (LO) that best suits the learning style of the current student is included in the learning path. Two papers that use this method are:

- [27] → The adaptation criteria include the four FSLSM dimensions. Each LO is manually annotated by the teacher using IMS Metadata Standard [14]. Each of the possible "Learning Resource Type" metadata values (i.e., "Exercise", "Simulation", "Questionnaire", "Diagram", "Figure", "Graph", "Index", "Slide", "Table", "Narrative Text", "Exam", "Experiment", "ProblemStatement", "SelfAssesment") are classified with the help of pedagogic experts according to the Felder and Silverman's teaching styles. First, the system finds the set of necessary domain concepts to be taught to the current student, based on the domain ontology and student's knowledge level. Next, for each domain concept, the set of LOs that explain it are found; the system selects one of these LOs taking into account the value of the attribute "Learning Resource Type" and trying to minimize the distance between the learning style and teaching style (interpreted as Euclidian distance).

- [17] → Again, the adaptation criterion is represented by the Felder-Silverman model. Each learning object is annotated by the teacher with a set of weights corresponding to its suitability for each of the 4 FSLSM dimensions. First, the system automatically generates a personalized learning path by means of a planner which takes into account the student's knowledge level and her FSLSM score. At each step, the system can output a new Learning Object Sequence, in case the student model has changed. For each knowledge item on the learning path, the system selects the associated LO which is the most suited for the learning style of the student, based on the assigned weights (i.e., having the smallest Euclidian distance from the student's learning style).

A more ***generic adaptation approach*** is proposed by Stash [28]. She uses an XML Learning Style Adaptation Language, called LAG-XSL, based on the LAG language (i.e., generalized adaptation model for generic adaptive hypermedia authoring [8]). LAG-XSL is a high level language, including adaptation actions such as: selection of different representations of concepts (media, level of difficulty, type of activity) and sorting of concepts. By means of these actions, authors can define their own adaptation strategies for their own learning styles. However, there is a limitation in the types of strategies that can be defined and consequently in the set of learning preferences that can be used. Paper [28] includes examples of 3 such instructional strategies, for Verbalizer versus Imager style, Global versus Analytic style and Activist versus Reflector style.

As far as the ***method for identifying the learning style*** of the student is concerned, the existing LSAES can be classified in two categories:

1. those that use an explicit modeling method (i.e., rely on the measuring instruments associated to the learning style models for diagnosing purposes)

2. those that use an implicit modeling method (i.e., based on the analysis of students' observable behavior).

The main advantages of the second category of systems are:

1. they don't require any additional work form the part of the students (for filling in the questionnaires)

2. they overcome the psychometric flaws of the traditional measuring instruments (which sometimes lack internal consistency, test-retest reliability or construct and predictive validity)

3. the student model can be continuously updated - it doesn't have to be static, created at the beginning of the course and stored once and for all.

Examples of works that fall in the first category are: [1], [5], [17], [19], [30], [31]. Examples from the second category include: [7], [10], [11], [12], [20], [27], [28], [29], [33].

In this paper we report a system (WELSA), which uses an implicit modeling method, combined with adaptive sorting and adaptive annotations techniques. Furthermore, WELSA is based not on a single learning style model (as all the systems included above), but on a complex of features extracted from several such learning style models. Finally, WELSA was thoroughly tested and experimental data is available regarding the accuracy of the modeling method as well as the efficiency and effectiveness of the adaptation on the learning process.

# 3 WELSA Overview

WELSA's functionalities are primarily addressed at the students, who can learn by browsing through the course and performing the instructional activities suggested (play simulations, solve exercises, etc.). They can also communicate and collaborate with their peers by means of the forum and chat. Students' actions are logged and analyzed by the system, in order to create accurate learner models. Based on the identified learning preferences and the built-in adaptation rules, the system offers students individualized courses. WELSA provides also functionalities for the teachers, who can create courses by means of the dedicated authoring tool; they can also set certain parameters of the modeling process, so that it fits the particularities of their course.

Figure 1 shows how WELSA appears for a learner who is studying a course on Artificial Intelligence (more specifically the chapter on "Constraint Satisfaction Problems", based on the classical textbook of Poole, Mackworth and Goebel [21]).

A few notes should be made regarding the course pages: the first resource (LO) on the page is entirely visible (expanded form), while for the rest of LOs only the title is shown (collapsed form). Of course, the student may choose to expand or collapse any resource, as well as lock them in an expanded state by clicking the corresponding icons. Also, there are specific icons associated to each LO, depending on its instructional role and its media type, in order to help the learner browse more effectively through the resources. Finally, navigation can be done by means of the Next and Previous buttons, the course outline or the left panel with the chapter list.

### 3.1 Architecture

The overall architecture of WELSA is illustrated in Fig. 2. WELSA is composed of three main modules:

- an authoring tool for the teachers, allowing them to create courses conforming to the internal WELSA format (XML-based representation)
- a data analysis tool, which is responsible for interpreting the behavior of the students and consequently building and updating the learner model, as well as providing various aggregated information about the learners
- a course player (basic learning management system) for the students, enhanced with two special capabilities: i) learner tracking functionality (monitoring the student interaction with the system); ii) adaptation functionality (incorporating adaptation logic and offering individualized course pages).

The three modules will be presented in more details in the next three sections.

As far as the implementation is concerned, Java-based and XML technologies are employed for all WELSA components. Apache Tomcat 6.0 is used as HTTP web server and servlet container and MySQL 5.0 is used as DBMS.

## 4 WELSA authoring tool

The course structure that we propose in WELSA is a hierarchical one: each course consists of several chapters, and each chapter can contain several sections and subsections. The lowest level subsection contains the actual educational resources. Each such elementary learning object corresponds to a physical file and has a metadata file associated to it [22]. These metadata are independent of any learning style; they describe the LO from the point of view of media type, format, instructional role, abstractness level, prerequisite, hierarchical and similarity relations with other LOs. Apart from being widely used for organizing the teaching materials, this approach also insures a high reusability degree of the educational resources. Furthermore, due to the fine granularity level of the LOs, a fine granularity of adaptation actions can also be envisaged. Finally, since each LO has a comprehensive metadata file associated to it, we know all the information about the learning resource that is accessed by the learner at a particular moment, so we can perform a detailed learner tracking.

In order to support the teacher in creating courses conforming to WELSA internal format, we have designed a course editor tool, which allows authors to easily assemble and annotate learning resources, automatically generating the appropriate file structure. It should be noted that WELSA course editor does not deal with the creation of actual content (text, images, simulations, etc.) - a variety of existing dedicated tools can be used for this purpose (text editors, graphics editors, HTML editors, etc.). Instead, WELSA course editor provides a tool for adding metadata to existing learning resources and defining the course structure (specifying the order of resources, assembling learning objects in pages, sections and subsections). The teacher can define this chapter structure in a simple and intuitive way, by using the course editor, as shown in Fig. 3. The corresponding XML files are subsequently generated by the application and stored on the server [23].

## 5 WELSA analysis tool (modeling component)

The adoption of a suitable taxonomy of learning styles plays an important role in the overall quality of the system. The result of the adaptation process can only be as accurate and comprehensive as the underlying student model. As mentioned in section 2, WELSA is based not on a single learning style model, like the rest of the similar systems, but on a complex of features extracted from several such learning style models (called ULSM - Unified Learning Style Model). This model integrates characteristics related to: perception modality, way of processing and organizing information as well as motivational and social aspects (e.g., *Visual / Verbal, Abstract / Concrete, Serial / Holistic, Active experimentation / Reflective observation, Individual work / Team work, Intrinsic motivation / Extrinsic motivation*). A detailed description of the ULSM characteristics, together with the model's rationale and advantages, is included in [25].

For the identification of these ULSM preferences, WELSA uses an implicit modeling mechanism, by analyzing the interaction of the students with the educational system, in the form of behavioral patterns. Once the learner actions are recorded by the course player, they have to be processed by the Analysis tool, in order to yield the learning preferences of the students. The modeling mechanism

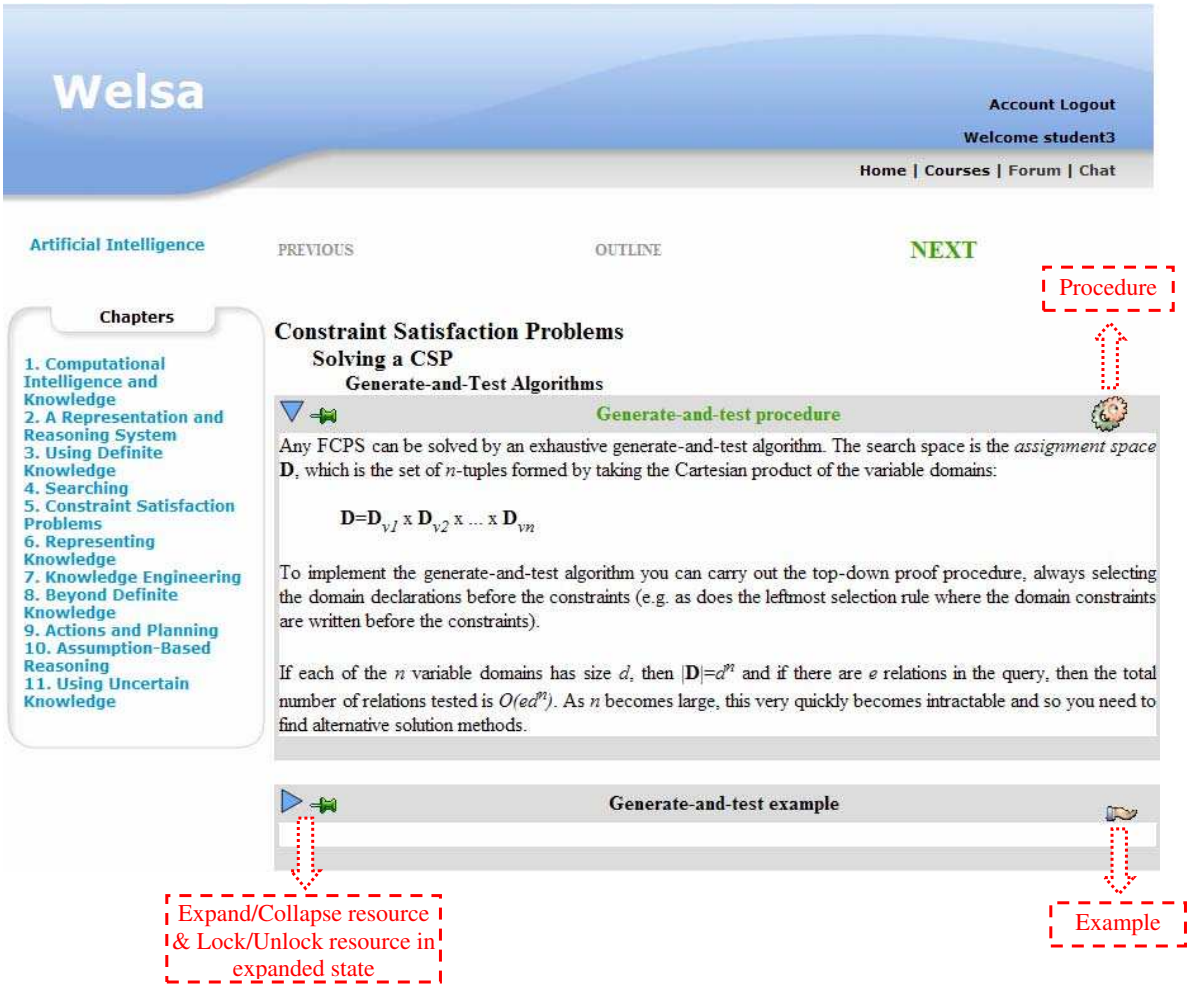



Figure 1: A snapshot of WELSA (student view)

is depicted in Fig. 4.

In order to compute the pattern values, a pre-processing phase of the raw data (i.e., the student actions and the associated timestamps) is necessary. The first step is to compute the duration of each action for each student, eliminating the erroneous values (for example, accessing the outline for more than 3 minutes means that the student actually did something else during this time). Next, the access time for each LO is computed, again filtering the spurious values (for example, an LO access time of less than 3 seconds was considered as random or a step on the way to another LO and therefore not taken into account). The data are then aggregated to obtain the pattern values for each student (e.g., total time spent on the course, total number of actions performed while logged in, time spent on each type of LO, number of hits on each category of LOs, the order of accessing the LOs, the number of navigation actions of a specific type, the number of messages in chat / forum, etc.). The reliability levels of these patterns are calculated as well (i.e., the larger the number of available relevant actions, the more reliable the resulted pattern). Next, the Analysis tool computes the ULSM preferences values, using modeling rules based on the pattern values, their reliability levels and their weights, as detailed in [24]. It should be noted that these rules also take into account the specificities of each course: the pattern thresholds as well as the importance of each pattern may vary with the structure and subject of the course. Therefore, the teachers should have the possibility to adjust the predefined values to correspond to the particularities of her/his course or even to eliminate some of the patterns, which are not relevant for that course. This is why the Analysis tool has a configuration option, which allows the teacher to modify the weight and threshold values, as seen in Fig. 5.

Beside the function of diagnosing the student learning preferences and correspondingly updating the learner model, the Analysis tool also offers various aggregated data that can be used for comparisons and statistical purposes. These tasks are accomplished by a researcher who interacts with the Analysis tool in the experimental version of WELSA. All the intermediate data (duration of learner actions, pattern values, pattern thresholds, reliability and confidence values) can be visualized by the researcher. Furthermore, at researcher's request, the analysis tool com-

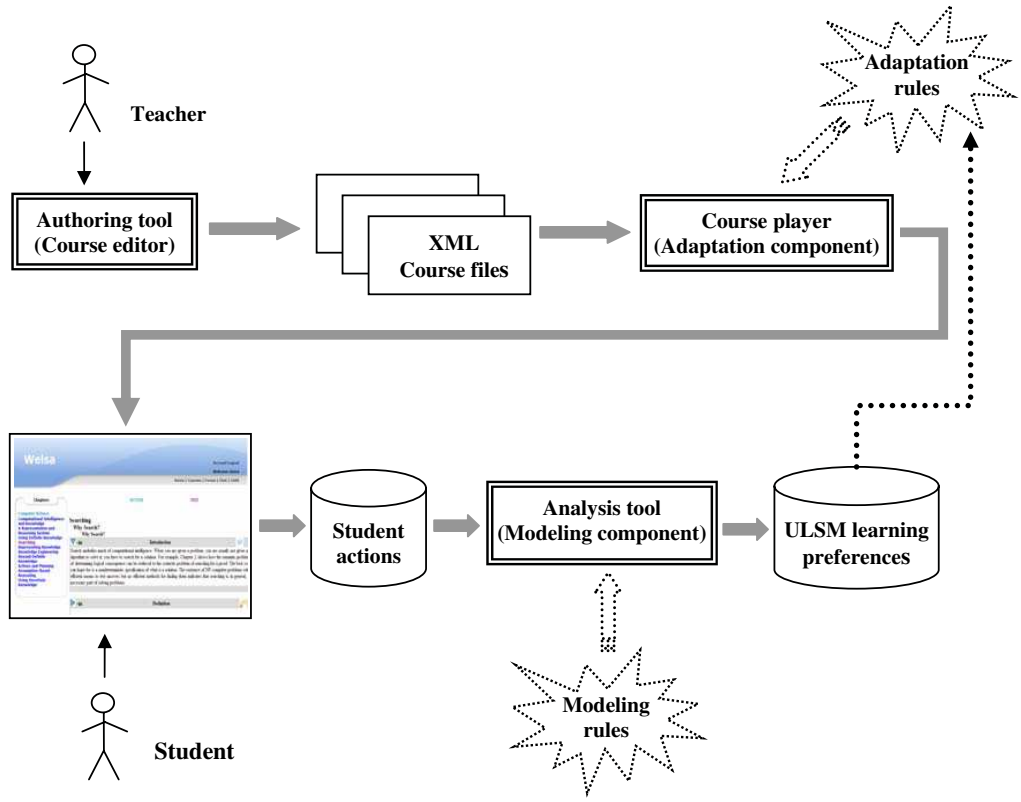



Figure 2: Overall WELSA architecture

putes and displays aggregated information, such as the total number of students with each ULSM preference, the total and average number of student actions, the average reliability and confidence values, etc. These data can be used for further analysis (e.g., by processing them in a dedicated statistical package). The roles and interactions of the actors with the Analysis tool are illustrated in Fig. 6.

In order to test the modeling method implemented in the Analysis tool, an experiment involving 71 undergraduate students was realized. The learners studied an AI course module on "Search strategies and solving problems by search" and all of their interactions with WELSA were recorded by the course player. Next, the Analysis tool computed the values of the behavioral patterns and applied the modeling rules, inferring the ULSM learning preferences of each student. In order to evaluate the validity of our modeling method, the results obtained by the Analysis tool (implicit modeling method) were compared with the reference results obtained using the ULSM questionnaire (explicit modeling method). Good precision results were obtained, with an average accuracy of 75.70%, as reported in [24].

# 6 WELSA course player (adaptation component)

WELSA course player is responsible with the generation of individualized web pages for each student; furthermore, it incorporates some basic LMS (learning management system) functions, such as: administrative support (registration and authentication) and communication and collaboration tools (discussion forum, chat).

Another function of the course player is to track student actions (down to click level) and record them in a database for further processing by the Analysis tool. This is done with the help of JavaScript code added to the HTML page, coupled with Ajax technology. Thus the application can communicate with the web server asynchronously in the background, without interfering with the display and behavior of the existing page.

In what follows we will give some details regarding the most important functionality of the course player, namely the adaptation mechanism, which allows the dynamic generation of individualized courses for each student.

Once the students' learning preferences are identified by the Analysis tool, the next step is to associate adaptation actions that are best suited for each preference. The development of these adaptation rules was a delicate task, since it involved interpretation of the literature in order to identify the prescriptive instructional guidelines. Indeed, apart

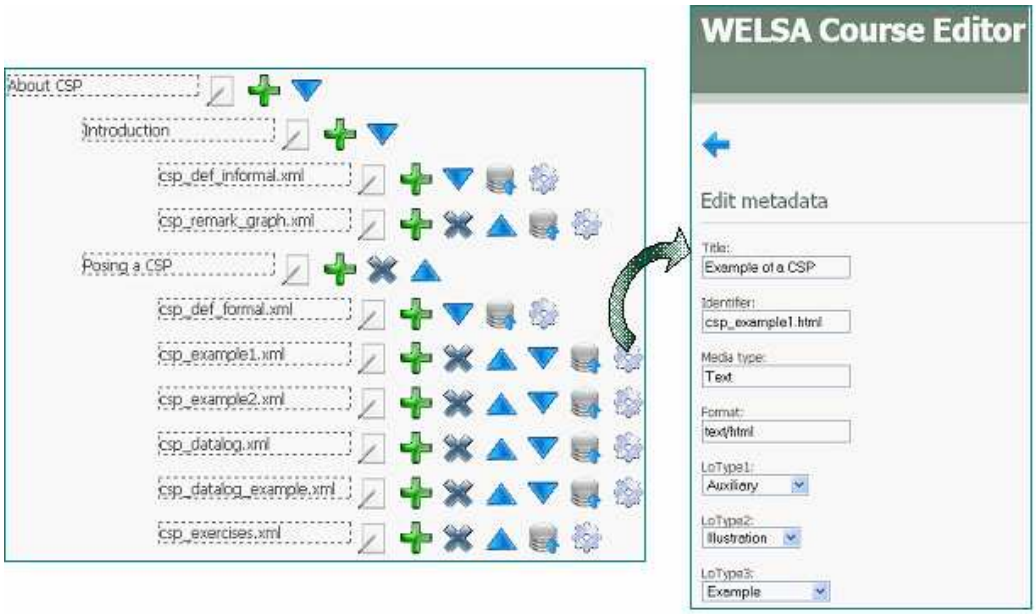



Figure 3: Snapshot of WELSA authoring tool: editing course structure (left-hand side) & editing metadata (right-hand side)

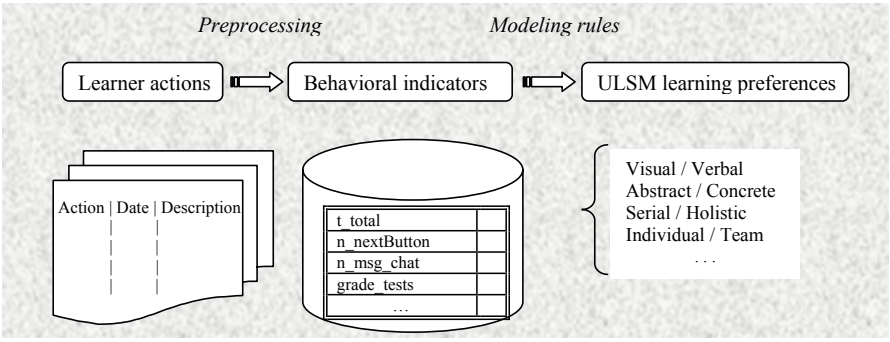
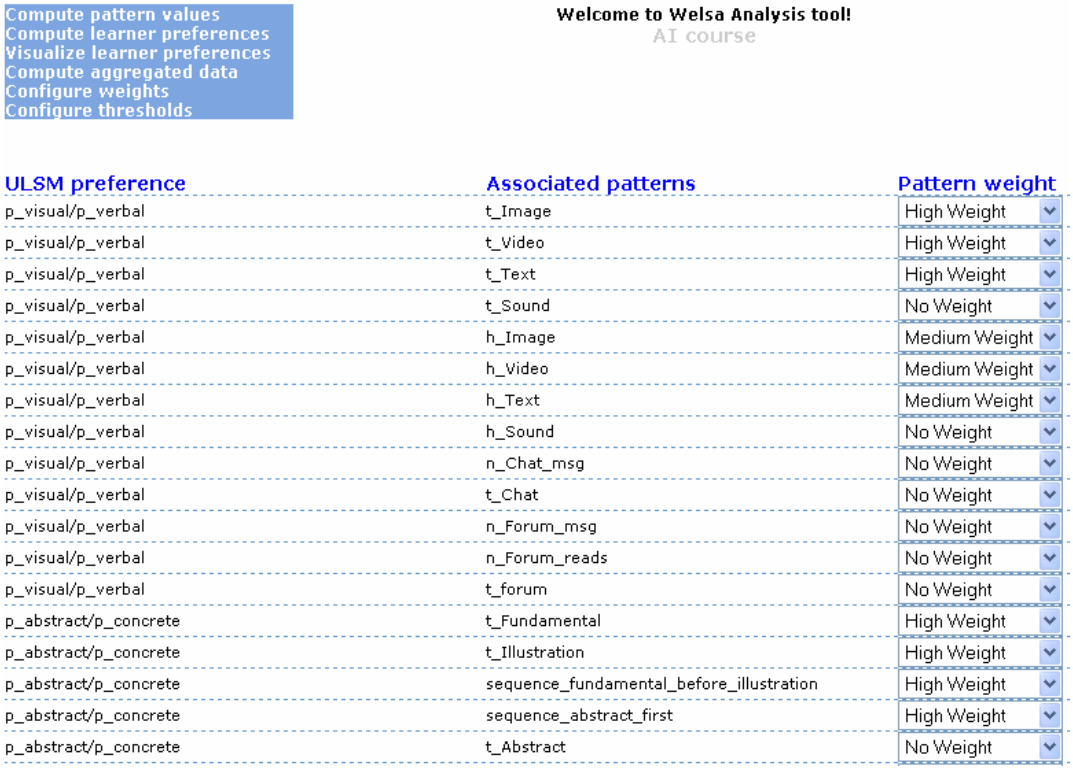



Figure 4: WELSA learner modeling mechanism

from defining the characteristics of the learners belonging to each learning style, for most of the models there are proposed teaching practices that effectively address the educational needs of students with the identified styles. However, as noted in [15], "learning styles models are usually rather descriptive in nature, in the sense that they offer guidelines as to what methods to use to best attain a given goal; they are not usually prescriptive in the sense of spelling out in great detail exactly what must be done and allowing no variation". Starting from these teaching methods (which only include a traditional learning view), enhancing them with e-learning specific aspects (technology-related preferences) and inspiring from other works that dealt with learning style based adaptation (as mentioned in section 2), we extracted the adaptation rules for our LSAES.

More specifically, we decided to use adaptive sorting and adaptive annotation techniques. The LOs are placed in the page in the order which is most appropriate to each learner; additionally, a "traffic light metaphor" was used to differentiate between recommended learning objects (LOs) (with a highlighted green title), standard LOs (with a black title) and not recommended LOs (with a dimmed light grey title) [26]. It should be mentioned however that the learning path suggested by the system is not compulsory: it is simply a recommendation that the student may choose to follow or not. We consider that offering control to students, instead of strictly guiding them, is a more flexible and rewarding pedagogical approach.

The adaptation mechanism is illustrated in Fig. 7, with a fragment of a Web page from an AI course generated for a student with a preference towards *Concrete, practical examples* rather than *Abstract concepts and generalizations*. The page is dynamically composed by selecting the appropriate LOs (mainly of type Example), each with its own status (highlighted in case of LOs of type Example and standard in case of LOs of type Definition) and ordered




| ULSM preference | Associated patterns | Pattern weight |
|---|---|---|
| p_visual/p_verbal | t_Image | High Weight |
| p_visual/p_verbal | t_Video | High Weight |
| p_visual/p_verbal | t_Text | High Weight |
| p_visual/p_verbal | t_Sound | No Weight |
| p_visual/p_verbal | h_Image | Medium Weight |
| p_visual/p_verbal | h_Video | Medium Weight |
| p_visual/p_verbal | h_Text | Medium Weight |
| p_visual/p_verbal | h_Sound | No Weight |
| p_visual/p_verbal | n_Chat_msg | No Weight |
| p_visual/p_verbal | t_Chat | No Weight |
| p_visual/p_verbal | n_Forum_msg | No Weight |
| p_visual/p_verbal | n_Forum_reads | No Weight |
| p_visual/p_verbal | t_forum | No Weight |
| p_abstract/p_concrete | t_Fundamental | High Weight |
| p_abstract/p_concrete | t_Illustration | High Weight |
| p_abstract/p_concrete | sequence_fundamental_before_illustration | High Weight |
| p_abstract/p_concrete | sequence_abstract_first | High Weight |
| p_abstract/p_concrete | t_Abstract | No Weight |

Figure 5: A snapshot from WELSA Analysis tool, illustrating the configuration options

correspondingly (first the notion of "Constraint satisfaction problem" is illustrated by means of two examples and only then a definition is provided).

Formally, the corresponding adaptation rules are included in Fig. 8. Note that *LoType* refers to the instructional role of the LO, as described in the metadata. More details regarding the LO indexing can be found in [22].

Adaptation rule for learners with "Abstract" preference
**IF**
$p\_abstract \in Pref(L)$
**THEN**
Sort $LoType$ {*Fundamental, Illustration*}
Highlight $LoType$ {*Fundamental*}

Adaptation rule for learners with "Concrete" preference
**IF**
$p\_concrete \in Pref(L)$
**THEN**
Sort $LoType$ {*Illustration, Fundamental*}
Highlight $LoType$ {*Illustration*}

Figure 8: Adaptation rules for *Abstract/Concrete* preference

The adaptation component consists of a Java servlet which automatically generates the individualized web page, each time an HTTP request is received by the server, as illustrated in Fig. 9. WELSA doesn't store the course web pages but instead generates them on the fly, following the structure indicated in the XML course and chapter files.

The adaptation servlet queries the learner model database, in order to find the ULSM preferences of the current student. Based on these preferences, the servlet applies the corresponding adaptation rules and generates the new HTML page. These adaptation rules involve the use of LO metadata, which as already stated in section 4, are independent of any learning style. However, they convey enough information to allow for the adaptation decision making (i.e., they include essential information related to the media type, the level of abstractness, the instructional role, etc.). Next the web page is composed from the selected and ordered LOs, each with its own status (highlighted, dimmed or standard).

This dynamic adaptation mechanism reduces the workload of authors, who only need to annotate their LOs with standard metadata and do not need to be pedagogical experts (neither for associating LOs with learning styles, nor for devising adaptation strategies). The only condition for LOs is to be as independent from each other as possible, without cross-references and transition phrases, to insure that the adaptation component can safely apply reordering techniques. Obviously, there are cases in which changing the order of the learning content is not desirable; in this case the resources should be presented in the predefined order only, independently of the student's preferences (the teacher has the possibility to specify these cases by means

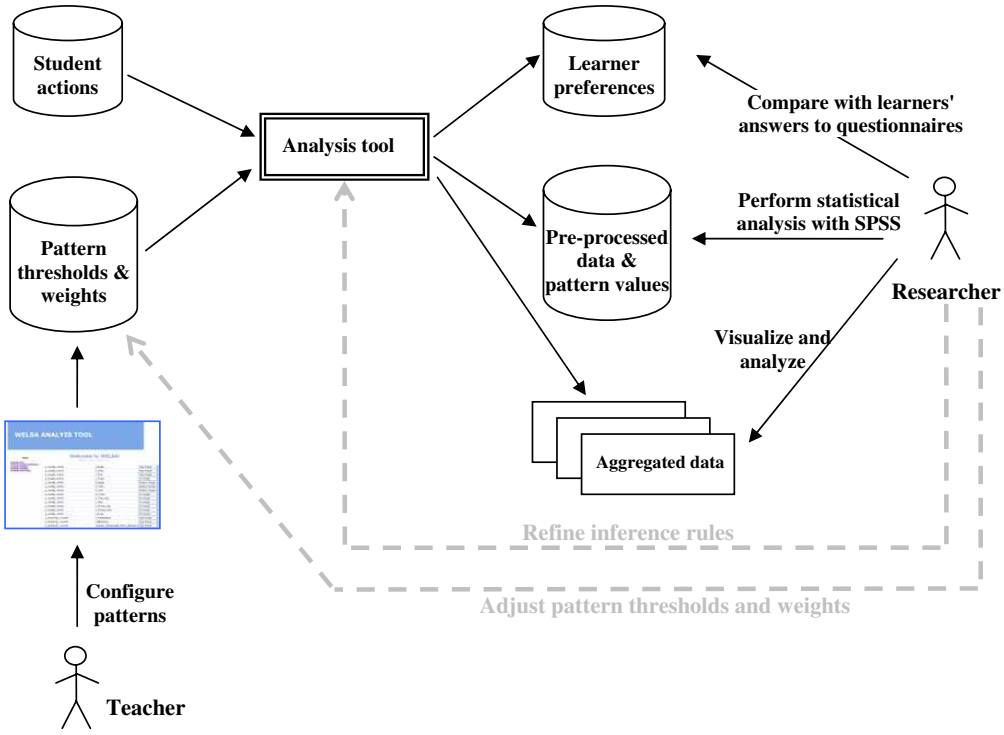



Figure 6: Users' interaction with the Analysis tool

of the prerequisites mechanism included in the metadata).

The validity and effectiveness of our adaptation approach were empirically confirmed by means of an experiment involving 64 undergraduate students in the field of Computer Science. The students were split in two groups: one which was provided with a matched version of the course (further referred to as "matched group") and one which was provided with a mismatched version of the course (further referred to as "mismatched group"), with respect to the students' learning preferences.

The objective evaluation consisted in performing a statistical analysis on the behavioral patterns exhibited by the students, comparing the values obtained for the matched and mismatched groups in order to find significant differences. The results showed that the matched adaptation approach increased the efficiency of the learning process, with a lower amount of time needed for studying and a lower number of randomly accessed educational resources (lower level of disorientation). The effectiveness of the matched adaptation and its suitability for addressing students' real needs are also reflected in the statistically significant higher time spent on recommended versus not recommended resources, as well the higher number of accesses of those recommended learning objects. Finally, the recommended navigation actions were followed to a larger extent than the not recommended ones.

As far as students' subjective evaluation of the system is concerned (as assessed by means of an opinion questionnaire), the students in the matched group reported significantly higher levels of enjoyment, overall satisfaction and motivation, compared to their mismatched peers. The overall results of the experimental study are very promising, proving the positive effect that our adaptation to learning styles has on the learning process. However, in order to allow for generalization, the system should be tested on a wider scale, with users of variable age, field of study, background knowledge and technical experience, which is one of our future research directions. Further details regarding the evaluation process can be found in [26].

# 7 Conclusion

The WELSA system described in this paper is an intelligent e-learning platform, aimed at adapting the course to the learning preferences of each student. We opened this paper with an extensive review of related LSAES. Starting from the existing systems, we introduced an innovative approach, based on an integrative set of learning preferences (ULSM). The technical and pedagogical principles behind WELSA were presented, focusing on the three main modules of the system. The learner modeling and adaptation methods were briefly introduced, together with their realization in WELSA.

As future work, improvements could be envisaged for each of the three main components. The authoring tool could be extended with an import/export facility, allowing for conversion between various course formats and standards (e.g., SCORM, IMS LD, etc.). The modeling component could also be extended to take into account the per-

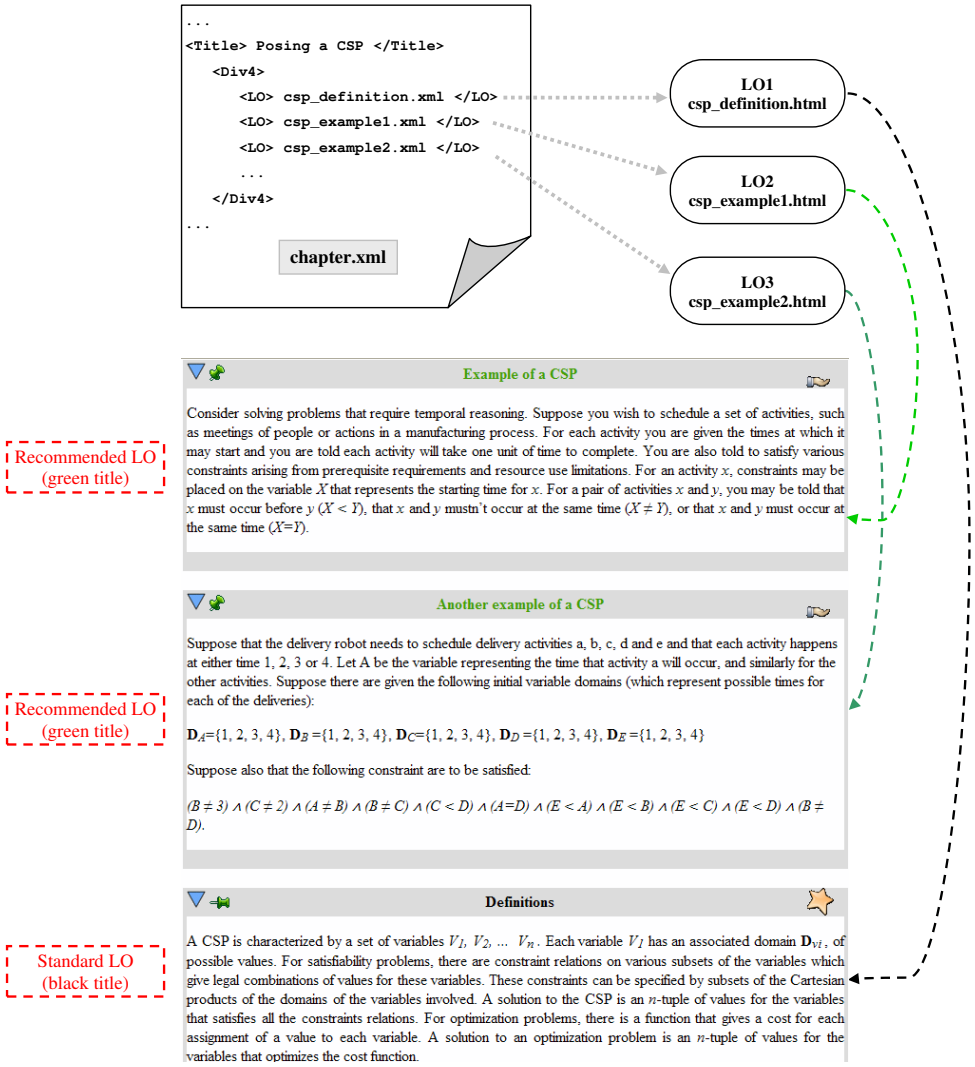



Figure 7: Composing a page from elementary LOs for a student with *Concrete* preference

turbations introduced by adaptation on students' actions; students' behavior in the adaptive version could be used as a valuable feedback on the effect of adaptation. Finally, the course player could incorporate a wider variety of adaptation actions, including also collaboration level adaptation techniques which are currently out of the scope of the system. In this respect, a wider range of communication and collaboration tools should be included in the system, including social software applications (e.g., blog, wiki, social bookmarking tool, etc.). Extending WELSA into a social and adaptive learning environment would be a challenging research direction.

## 8 Acknowledgment

This work was partially supported by the strategic grant POSDRU/89/1.5/S/61968, Project ID 61968 (2009), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007 – 2013.

## References

[1] Bajraktarevic, N., Hall, W., Fullick, P. (2003) Incorporating learning styles in hypermedia environment: Empirical evaluation. *Proc. Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems*, pp. 41–52.

[2] Brusilovsky, P. (2001) Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11, pp. 87-110.

[3] Brusilovsky, P., Peylo, C. (2003) Adaptive and Intelligent Web-based Educational Systems. *International Journal of Artificial Intelligence in Education*, 13 (2-4), pp. 159–172.

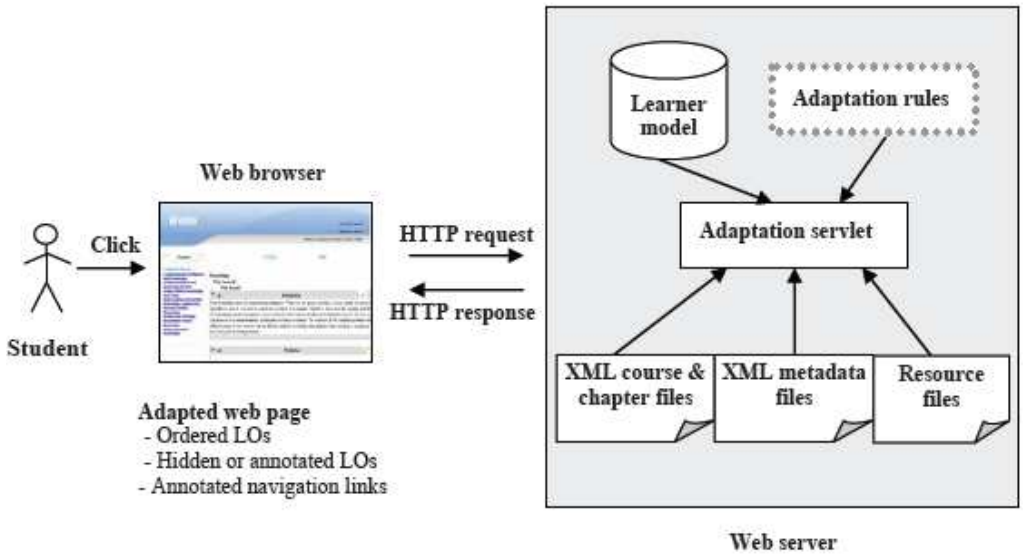



Figure 9: Adaptation component schematic architecture

[4] Boticario, J.G., Santos, O.C., van Rosmalen P. (2005) Issues in Developing Standard-based Adaptive Learning Management Systems. *EADTU 2005 Working Conference: Towards Lisbon 2010: Collaboration for Innovative Content in Lifelong Open and Flexible Learning*.

[5] Carver, C. A., Howard, R. A., Lane, W. D. (1999) Enhancing student learning through hypermedia courseware and incorporation of student learning styles. *IEEE Transactions on Education*, 42, pp. 33–38.

[6] Cha, H. J., Kim, Y. S., Lee, J. H., Yoon, T. B. (2006) An Adaptive Learning System with Learning Style Diagnosis based on Interface Behaviors. *Workshop Proceedings of Intl. Conf. E-learning and Games (Edutainment 2006)*.

[7] Cha, H. J., Kim, Y. S., Park, S. H., Yoon, T. B., Jung, Y. M., Lee J. H. (2006) Learning styles diagnosis based on user interface behaviors for the customization of learning interfaces in an intelligent tutoring system. *Procs. ITS 06*. Lecture Notes in Computer Science, Vol. 4053, Springer, pp. 513–524.

[8] Cristea, A., Calvi, L. (2003) The Three Layers of Adaptation Granularity. *Proc. UM 2003*, pp. 4–14.

[9] Felder, R. M., Silverman, L. K. (1988) Learning and Teaching Styles in Engineering Education. *Engineering Education*, 78(7). Preceded by a preface in 2002: `http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/LS-1988.pdf`.

[10] Garcia, P., Amandi, A., Schiaffino, S., Campo, M. (2007) Evaluating Bayesian NetworksŠ Precision for Detecting StudentsŠ Learning Styles. *Computers & Education*, 49(3), pp. 794–808.

[11] Gilbert, J.E., Han, C.Y. (1999) Adapting instruction in search of Śa significant differenceŠ. *Journal of Network and Computer Applications*, 22(3), pp. 149–160.

[12] Graf, S. (2007) *Adaptivity in Learning Management Systems Focussing on Learning Styles*. PhD Thesis, Vienna University of Technology, Austria.

[13] Honey, P., Mumford, A. (2000) *The Learning Styles Helper's Guide*. Maidenhead: Peter Honey Publications Ltd.

[14] IMS Metadata Standard (2010) `http://www.imsglobal.org/metadata/`.

[15] Karagiannidis, C., Sampson, D. (2004) Adaptation Rules Relating Learning Styles Research and Learning Objects Metadata. *Proc. Workshop on Individual Differences in Adaptive Hypermedia in AH2004*, pp. 60-69.

[16] Keefe, J.W. (1979) Learning style: an overview. *NASSP's Student Learning Styles: Diagnosing and Prescribing Programs*, pp. 1–17.

[17] Limongelli, C., Sciarrone, F., Temperini, M., Vaste, G. (2009) Adaptive Learning with the LS-Plan System: A Field Evaluation. *IEEE Transactions on Learning Technologies* 2(3), pp. 203–215.

[18] Moodle (2010) `http://moodle.org`.

[19] Papanikolaou, K.A., Grigoriadou, M., Kornilakis, H., Magoulas, G.D. (2003) Personalizing the interaction

in a Web-based educational hypermedia system: the case of INSPIRE. *User-Modeling and User-Adapted Interaction*, 13, pp. 213–267.

[20] Paredes, P., Rodriguez, P. (2004) A Mixed Approach to Modelling Learning Styles in Adaptive Educational Hypermedia. *Advanced Technology for Learning*, 1(4), pp. 210–215.

[21] Poole, D., Mackworth, A., Goebel, R. (1998) *Computational Intelligence: A Logical Approach*. Oxford University Press.

[22] Popescu, E., Badica, C., Trigano, P. (2008) Description and organization of instructional resources in an adaptive educational system focused on learning styles, *Procs. IDC 2007*, Studies in Computational Intelligence, Vol. 78, Springer, pp. 177–186.

[23] Popescu, E., Trigano, P., Badica, C., Butoi, B., Duica, M. (2008) A Course Authoring Tool for WELSA Adaptive Educational System. *Proc. ICCC 2008*, pp. 531–534.

[24] Popescu, E. (2009) Diagnosing Students' Learning Style in an Educational Hypermedia System. *Cognitive and Emotional Processes in Web-based Education: Integrating Human Factors and Personalization*, Advances in Web-Based Learning Book Series, IGI Global, pp. 187–208.

[25] Popescu, E. (2010) A Unified Learning Style Model for Technology-Enhanced Learning: What, Why and How?. *International Journal of Distance Education Technologies*, 8(3), IGI Global, pp. 65–81.

[26] Popescu, E. (2010) Adaptation Provisioning with respect to Learning Styles in a Web-Based Educational System: An Experimental Study. *Journal of Computer Assisted Learning*, 26(4), Wiley, pp. 243–257.

[27] Sangineto, E., Capuano, N., Gaeta, M., Micarelli, A. (2008) Adaptive course generation through learning styles representation. *Journal of Universal Access in the Information Society*, 7(1), pp. 1–23.

[28] Stash, N. (2007) *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*. PhD Thesis, Eindhoven University of Technology, Netherlands.

[29] Stathacopoulou, R., Grigoriadou, M., Samarakou, M., Mitropoulos, D. (2007) Monitoring studentsŠ actions and using teachersŠ expertise in implementing and evaluating the neural network-based fuzzy diagnostic model. *Expert Systems with Applications*, 32, pp. 955–975.

[30] Triantafillou, E., Pomportsis, A., Demetriadis, S. (2003) The design and the formative evaluation of an adaptive educational system based on cognitive styles. *Computers & Education*, 41, pp. 87–103.

[31] Wang, T., Wang, K., Huang, Y. (2008) Using a Style-based Ant Colony System for Adaptive Learning. *Expert Systems with Applications*, 34(4), pp. 2449–2464.

[32] Witkin, H.A. (1962) *Psychological Differentiation: Studies of Development*. New York: Wiley.

[33] Wolf, C. (2002) iWeaver: Towards an Interactive Web-Based Adaptive Learning Environment to Address Individual Learning Styles. *European Journal of Open, Distance and E-Learning*. Available at: `http://www.eurodl.org/materials/contrib/2002/2HTML/iWeaver.htm`.

# Earth Observation Data Processing in Distributed Systems

Petcu Dana, Panica Silviu, Neagul Marian, Frîncu Marc and Zaharie Daniela,
Ciorba Radu, Diniş Adrian
Computer Science Department, West University of Timişoara, Romania
E-mail: gisheo-uvt@lists.info.uvt.ro

**Keywords:** distributed computing, earth observation, data processing

**Received:** February 15, 2010

*Earth observation systems have a continuous growth in the user and internal requirements that can be handled nowadays only using distributed systems. These requirements are shortly reviewed in this paper. Huge data-sets management and processing are of special interest, as well as the particularities of the Earth observation data. On the technological side, the focus is put on service-oriented architectures that are facilitating the linkage of data or resources and processing. As proof of concept of current distributed system capabilities, the technological solutions used to build a training platform for Earth observation data processing are exposed and discussed in details.*

*Povzetek: S pomoèjo distribuiranega sistema je realiziran opazovalni sistem Zemlje.*

## 1 Introduction

The paradigm known as Data Intensive Science [7] is currently changing the way research and innovation is being conducted. This paradigm is based on access and analysis of large amounts of existing or new data that were or are created not only by scientific instruments and computers, but also by processing and collating existing archived data. Earth observation systems, in particular, are gathering daily large amounts of information about the planet and are nowadays intensively used to monitor and assess the status of the natural and built environments.

Earth observation (EO) is most often referring to satellite imagery or satellite remote sensing, the result of sensing process being an image or a map. Remote sensing refers to receiving and measuring reflected or emitted radiation from different parts of the electromagnetic spectrum. Remote sensing systems involve not only the collection of the data, but also their processing and distribution. The rate of increase in the remote sensing data volume is continuously growing. Moreover, the number of users and applications is also increasing and the data and resource sharing became a key issue in remote sensing systems. Furthermore, EO scientists are often hindered by difficulties locating and accessing the data and services. These needs lead to a shift in the design of remote sensing systems from centralized environments towards wide-area distributed environments that allow a scale-out in a wide range of issues from real-time access to enormous quantities of data to experimental repeatability through the use of workflows. The underlying technologies used in service-oriented architectures, either Web, Grid or Cloud based, are facilitating this transition as well the linkage of data, resources, and processing.

In this context, the paper starts with a survey of the current requests imposed on distributed systems and coming from remote sensing application field. This survey is based on several recent research reports of EO and distributed systems communities and it is an extended version of [19]. A deeper look is dedicated to the Grid usage benefits for EO through techniques like bringing computations to the data, rather than data to the computations. Furthermore, a snapshot of the requests that can be satisfied by the current technologies is provided through a case study on a newly proposed service-based system for training in EO. A short list of conclusions is provided in the last section.

## 2 EO requests on distributed environments

Satellite image processing is usually a computational and data consuming task and special techniques are required for both data storage and processing in distributed environments. In what follows we point some main topics. This section is a survey of the ideas exposed in the recent scientific reports [3, 6, 7, 8].

### 2.1 Data management

The management of the distribution of data, from storing to long-term archiving, is currently an important topic in EO systems. The first issue is the data format that is varying from image files, databases, or structured file. Usually an EO data contain metadata describing the data, such as the dimensionality or reference coordinates. Another issue is related to the user need to access remotely the EO data. Due to the size of the EO data, a distributed file system is needed. For more than three decades there are several distributed file systems enabling multiple, distributed

servers to be federated under the same file namespace. Another issue is the data discovery and this is currently done usually exploiting metadata catalogs. Replica management services are essential for EO systems, allowing to determine an optimal physical location for data access based on data destination aiming to reducing the network traffic and the response time. Data transfers secure protocols were developed to extends the traditional file transfer protocol.

In what concerns file catalogs there are no current standards, but several implementations are available in Grid environments that are using special file catalogs allowing data replications. The same situation is valid also for metadata catalogs; fortunately, in the particular case of EO this issue is pursued by the Open Geospatial Consortium (http://www.opengeospatial.org).

While for the basic needs mentioned above there are several stable and standardized solutions, the current key issue in EO data management is to make the data reachable and useful for any application through interoperability.

Interoperability is achieved through the usage of standard interfaces and protocols. Interoperable interfaces are attractive to users allowing the fast design of distributed application based on multiple components. Achieving interoperability includes also building adapted interfaces providing different front ends to basic services and bridging protocols. There are at least two layers for interoperability: for resource format and domain encoding, and semantic interoperability.

Interoperability solutions for resources structures and content are often application-field dependent. The solutions are related to different levels, like device, communication, middleware and deployment ones. At device level, the solutions are mostly standardized and are referring to the interfaces to the storage devices. At communication level, there are standardized data transfer protocols (as HTTP, HTTPS, or GridFTP), standardized protocols for Web services, and less standardized data movers for heterogeneous computing environments. At middleware level there are fewer standard solutions. For example, for data storage it is necessary a single consistent interface to different storage systems – a solution is coming from Grid community through the open standard storage resource manager, a control protocol for accessing mass storage.

In what concerns the interoperability of federated databases, a standard again proposed by the Grid community is the Open Grid Services Architecture Data Movement Interface (OGSA-DMI, http://forge.gridforum.org/sf/projects/ogsa-dmi-wg).

At deployment level, interoperability degradation is related to the event of new deployments and currently there are no automated tools or standard interfaces allowing the propagation of updates.

While resource-level interoperability is ensuring the compatibility of implementations at hardware and software levels, the semantic interoperability is enabling data and information flows to be understood at a conceptual level. Research efforts are currently devoted to the definition of generic data models for specific structured linguistic data types with the intention to represent a wide class of documents without loosing the essential characteristics of the linguistic data type.

Data provision services in EO are not satisfying the nowadays' user needs due to current application and infrastructure limitations. The process of identifying and accessing data takes up a lot of time, according [6], due to: physical discontinuity of data, diversity of metadata formats, large volume of data, unavailability of historic data, and many different actors involved.

In this context, there is a clear need for an efficient data infrastructure able to provide reliable long-term access to EO data via the Internet, and to allow the users to easily and quickly derive information and share knowledge. Recognizing these needs, the European INSPIRE Directive (http://inspire.jrc.ec.europa.eu) requires all public authorities holding spatial data to provide access to that data through common metadata, data and network service standards. OPeNDAP (http://opendap.org/) is a data transport architecture and protocol widely used in EO; it is based on HTTP and includes standards for encapsulating structured data, annotating the data with attributes, and adding semantics that describe the data. Moreover, it is widely used by governmental agencies to EO data [6].

The Committee on EO Satellites (www. ceos.org) maintains a Working Group on Information Systems and Services with the responsibility to promote the development of interoperable systems for the management of EO data internationally. This group plans to build in the next decade the Global EO System of Systems (GEOSS) targeting the development of a global, interoperable geospatial services architecture [11].

## 2.2 Data processing

To address the computational requirements introduced by time-critical satellite image applications, several research efforts have been oriented towards parallel processing strategies. According to the Top500 list of supercomputer sites, NASA, for example, is maintaining two massively parallel clusters for remote sensing applications. The recent book [23] presents the latest achievements in the field of high performance computing (HPC).

Currently ongoing research efforts are aiming also the efficient distributed processing of remote sensing data. Recent reports are related to the use of new versions of data processing algorithms developed for heterogeneous clusters as [22]. Moreover, distributed application framework specifically have been developed for remote sensed data processing, like JDAF [30]. EO applications are also good candidates for building architectures based on components encapsulating complex data processing algorithms and being exposed through standard interfaces like in [10].

As datasets grow larger, the most efficient way to perform data processing is to move the analysis functions as close to the data as possible [28]. Data processing can be

easily expressed today by a set-oriented, declarative language whose execution can benefit enormously from cost-based query optimization, automatic parallelism, and indexes. Moreover, complex class libraries written in procedural languages were developed as extension of the underlying database engine. MapReduce, for example, is nowadays a popular distributed data analysis and computing paradigm; its principle resembles the distributed grouping and aggregation capabilities existing in parallel relational database systems. However, partitioned huge datasets are making distributed queries and distributed joins difficult. While simple data-crawling strategy over massively scaled-out data partitions is adequate with MapReduce, this strategy is suboptimal: a good index can provide better performance by orders of magnitude [28]. Moreover, joins between tables of very different cardinalities are still difficult to use.

Web services technology emerged as standard for integrating applications using open standards. In EO, the Web services play a key role. A concrete example is the Web mapping implementation specification proposed by OpenGIS (http://www.opengis. org). Web technologies are allowing also the distribution of scientific data in a decentralized approach and are exposing catalogue services of dataset metadata.

Grid computing services and more recent Cloud computing services are going beyond what Web services are offering, making a step forward towards an interactive pool of processes, datasets, hardware and software resources.

# 3 Grid-based environments for Earth observation

The promise of a Grid for EO community is to be a shared environment that provides access to a wide range of resources: instrumentation, data, HPC resources, and software tools. There are at least three reasons for using Grids for EO:

1. the required computing performance is not available locally, the solution being the remote computing;

2. the required computing performance is not available in one location, the solution being cooperative computing;

3. the required services are only available in specialized centres, the solution being application specific computing.

Realizing the potential of the Grid computing for EO, several research projects were launched to make the Grid usage idea a reality. We review here the most important ones.

## 3.1 Grid-based EO initiatives

Within the DataGrid project funded by the European Commission, an experiment aiming to demonstrate the use of Grid technology for remote sensing applications has been carried out; the results can be found for example in the paper [15]. Several other international Grid projects were focused on EO, like SpaceGrid (http://www.spacegrid.org), Earth Observation Grid (http://www.e-science.clrc.ac.uk/web/projects/earthobservation), or Genesis [35]. The MediGrid project (http://www.eu-medigrid.org) aimed to integrate and homogenize data and techniques for managing multiple natural hazards. The authors of paper [1] present an overview of SARA Digital Puglia, a remote sensing environment that shows how Grid and HPC technologies can be efficiently used to build dynamic EO systems for the management of space mission data and for their on-demand processing and delivering to final users.

A frequent approach is to use the Grid as a HPC facility for processor-intensive operations. The paper [29], for example, focuses on the Grid-enabled parallelization of the computation-intensive satellite image geo-rectification problem. The aim of the proposed classification middleware on Grid from [31] is to divide jobs into several assignments and submit them to a computing pool. The parallel remote-sensing image processing software PRIPS was encapsulated into a Grid service and this achievement was reported in [33]. In the paper [34] is discussed the architecture of a spatial information Grid computing environment, based on Globus Toolkit, OpenPBS, and Condor-G; a model of the image division is proposed, which can compute the most appropriate image pieces and make the processing time shorter.

CrossGrid (http://www.crossgrid.org) aimed at developing techniques for real-time, large-scale grid-enabled simulations and visualizations, and the issues addressed included distribution of source data and the usefulness of Grid in crisis scenarios.

DEGREE (http://www.eu-degree.eu) delivered a study on the challenges that the Earth Sciences are imposing on Grid infrastructure. D4Science (http://www.d4science.org) studied the data management of satellite images on Grid infrastructures. G-POD (http://eogrid.esrin.esa.int/) aims to offer a Grid-based platform for remote processing the satellite images provided by European Space Agency. The GlobAEROSOL service of BEinGRID [24] is processing data gathered from satellite sensors and generates an multi-year global aerosol information in near real time.

The GEOGrid project [26] provides an e-Science infrastructure for Earth sciences community and integrates a wide varieties of existing data sets including satellite imagery, geological data, and ground sensed data, through Grid technology, and is accessible as a set of services.

LEAD (https://portal.leadproject.org/) is creating an integrated, scalable infrastructure for meteorology research; its applications use large amounts of streaming data from sensors.

The Landsat Data Continuity Mission Grid Prototype (LGP) offers a specific example of distributed processing of remotely sensed data generating single, cloud and shadow scenes from the composite of multiple input scenes [8].

GENESI-DR (http://genesi-dr.eu) intends to prove reliable long-term access to Earth Science data allowing scientists to locate, access, combine and integrate data from space, airborne and in-situ sensors archived in large distributed repositories; its discovery service allows to query information about data existing in heterogeneous catalogues, and can be accessed by users via a Web portal, or by external applications via open standardized interfaces (OpenSearch-based) exposed by the system [6].

Several other smaller projects, like MedioGrid [20], were also initiated to provide Grid-based services at national levels.

### 3.2 Remote sensing Grid

A Remote Sensing Grid (RSG) is defined in [8] as a highly distributed system that includes resources that support the collection, processing, and utilization of the remote sensing data. The resources are not under a single central control. Nowadays it is possible to construct a RSG using standard, open protocols and interfaces. In the vision of [8] a RSG is made up of resources from a variety of organizations which provide specific capabilities, like observing elements, data management elements, data processing and utilization elements, communications, command, and control elements, and core infrastructure.

If a service oriented architecture is used, modular services can be discovered and used to build complex applications by clients. The services should have the following characteristics [8]: composition, communication, workflow, interaction, and advertise. These requirements are mapped into the definition of specific services for workflow management, data management and processing, resource management, infrastructure core functions, policy specification, and performance monitoring.

The services proposed in [8] are distributed in four categories: workflow management services, data management services, applications in the form of services, and core Grid services.

In the next section we describe a case study of a recent Grid-based satellite imagery system that follows the RSG concepts.

## 4 Case study: GiSHEO

The rapid evolution of the remote sensing technology is not followed at the same developing rate by the training and high education resources in this field. Currently there are only a few number of resources involved in educational activities in EO. The CEOS Working Group of Education, Training and Capacity Building is one of the few facilities that is collecting an index of free EO educational materials (http://oislab.eumetsat.org/CEOS/webapps/).

Recognizing the gap between research activities and the educational ones, we have developed recently a platform, namely GiSHEO (On Demand Grid Services for Training and High Education in EO, http://gisheo.info.uvt.ro) addressing the issue of specialized services for training in EO. Contrary to the existing platforms providing tutorials and training materials, GiSHEO intends to be a living platform where experimentation and extensibility are the key words. Moreover, special solutions were proposed for data management, image processing service deployment, workflow-based service composition, and user interaction. A particular attention is given to the basic services for image processing that are reusing free image processing tools. A special feature of the platform is the connection with the GENESI-DR catalog mentioned in the previous section.

While the Grid is usually employed to respond to the researcher requirements to consume resources for computational-intensive or data-intensive tasks, GiSHEO aims to use it for near-real time applications for short-time data-intensive tasks. The data sets that are used for each application are rather big (at least of several tens of GBs), and the tasks are specific for image processing (most of them very simple). In this particular case a scheme of instantiating a service where the data are located is required in order to obtain a response in near-real time. Grid services are a quite convenient solution in this case: a fabric service is available at the server of the platform that serves the user interface and this service instantiates the processing service where the pointed data reside. In our platform the Web services serve as interfaces for the processing algorithms. These interfaces allow the remote access and application execution on a Grid using different strategies for fast response.

The platform design concepts were shortly presented in [5, 17] and the details about the e-learning component can be found in [9]. The EO services were described in [21] and the data management is detailed in [13]. In this section we describe the technological approaches that were undertaken to solve the key issues mentioned in the previous sections.

### 4.1 System architecture

The GiSHEO Architecture is a Grid-enabled platform for satellite image processing using a service oriented architecture structured on several levels including user, security, service, processing and a data level (Figure 1).

The user level is in charge with the access to the Web user interface (built by using DHTML technologies).

The security level provides security context for both users and services. The security context defines the mechanisms used for authentication, authorization and delegation. Each user must be identified by either using a username/password pair or a canonical name provided by a digital certificate. The services are using a digital certificate for authentication, authorization, and trust delegation. The authorization is based on VOMS (Virtual Organization Management Service, http://vdt.cs.wisc.edu/VOMS-

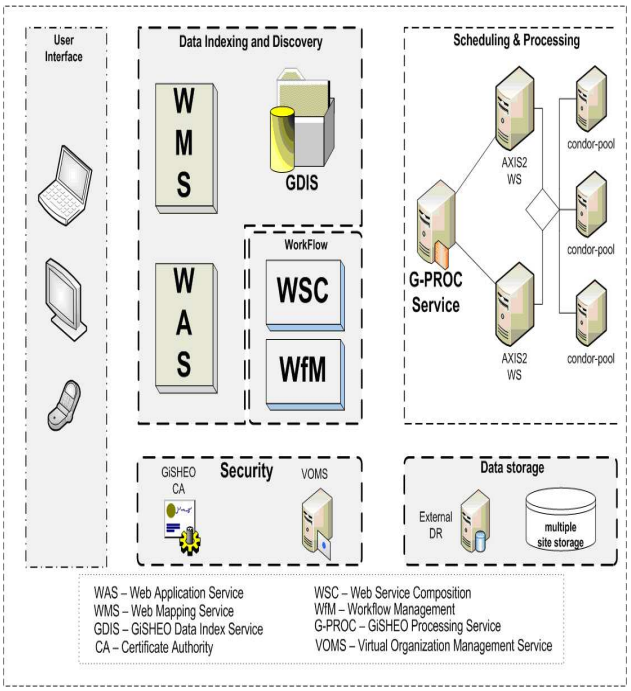



Figure 1: GiSHEO's architecture.

documentation.html) service which extends the PKI (Public Key Infrastructure) mechanism by adding support for certificate attributes. The service level exposes internal mechanisms part of the GiSHEO platform by using various Grid services technologies including:

- The Grid processing service - internal processing platform exposed as a specialized Web service and capable of connecting with an external resource management system.
- The workflow service - internal workflow engine which can be accessed by using a specialized Web service.
- The data indexing and discovery - access to the GiSHEO proposed data management mechanisms.

At processing level the GiSHEO platform uses Condor HTC (Condor High Throughput Computing, http://www.cs.wisc.edu/condor/ description.html) as processing model, task registration, scheduling and execution environment. It uses also a direct job submission interface using Condor's specific Web service interface.

At data level two different types of data are involved: database datasets which contain the satellite imagery repository and processing application datasets used by applications to manipulate satellite images.

Each of these components will be presented in further details in the following subsections.

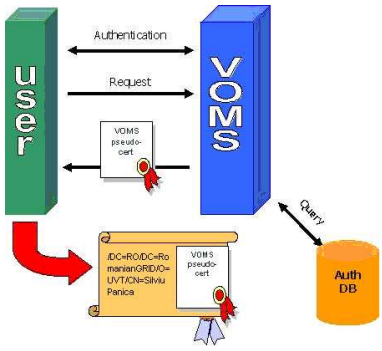



Figure 2: VOMS workflow.

## 4.2 Security mechanisms

The security level of the GiSHEO platform is divided into three categories: authentication, authorization and delegation.

The authentication is accomplished by using X.509 digitally signed certificates. Each entity either a user or a service will have a digital certificate attached. In case users choose to use only a username/password pair, a digital certificate (also a user private/public key) will be also generated and stored on the Web portal side (using a private key vault). In order to be valid, each certificate must be signed by a trusted CA (Certificate Authority), GiSHEO CA or a EuGRIDPMA (European Policy Management Authority for Grid Authentication, www.eugridpma.org) affiliated CA.

In conclusion each entity will be uniquely identified by using a CN (Canonical Name) which has the following form:

```
/DC=RO/DC=GiSHEO CA/O=UVT/CN=User1
/DC=RO/DC=GiSHEO CA/O=UVT/CN=service/GTD
```

For authorization, a VOMS service approach has been chosen. VOMS is a service providing VO membership for both users and services by using a set of attributes that are included inside the user's digital certificate. VOMS can be viewed as an extension to the simple digital certificate authorization (in which case only CA signing validation is made). As the following example will show in VOMS each entity is mapped to a set of attributes as configured by the VO administrator:

```
"/DC=RO/DC=GiSHEO CA/O=UVT/CN=User1" .gisheo
"/DC=RO/DC=GiSHEO CA/O=UVT/CN=U2" .sgmgisheo
"/gisheo/Role=ops" .gisheo
"/gisheo/Role=ops/Capability=NULL" .gisheo
"/gisheo/Role=VO-Admin" .sgmgisheo
```

In the above example `User1` is mapped to `.gisheo` group while `U2` is mapped to `.sgmgisheo` group. Each group has attached one or more attributes. For example group `.sgmgisheo` has attribute `/gisheo/Role=VO-Admin` attached to it which means that any service user belonging to group `.sgmgisheo` is a VO Administrator.

The VOMS authorization process is described by the following steps (Figure 2). First a request for validation arrives from a VOMS service. Then the VOMS service checks the user's CN membership and creates a proxy certificate (with a limited life time) with user's attributes available for GiSHEO VO. After the creation of the proxy certificate it can be used by the user to securely access any services belonging to the GiSHEO VO.

The GiSHEO Infrastructure uses a single sign-on authentication system; therefore a delegation of credentials mechanism is needed. For this to happen a MyProxy Credentials Management Service (http://grid.ncsa.uiuc.edu/myproxy/) is used. MyProxy also provides VOMS authentication mechanisms therefore can be easily integrated with the VOMS service. The goal of the MyProxy service is to provide delegation mechanism for both entities, users and services.

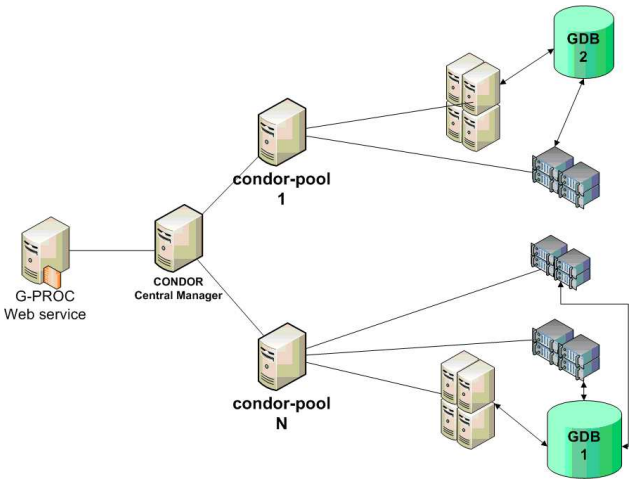



Figure 3: G-PROC with CondorWrappers.

## 4.3 Processing platform

The GiSHEO processing platform consists of three parts, the Web service interface for the processing platform, the EO processing tools connectors and a set of connectivity wrappers that describe the mechanism of connecting GiSHEO's platform to Condor HTC workload management system (WMS).

The Grid Processing Web service (G-PROC) is built using Web service technologies from Apache Axis2 (http://ws.apache.org/axis2/). It is responsible for the interaction with other internal services including the Workflow Composition Engine. Its main responsibilities are at this point to receive tasks from the workflow engine or directly from the user interface, to use a task description language (the ClassAd meta language for example in case of Condor HTC) in order to describe a job unit, to submit and check the status of jobs inside the workload management system, and to retrieve job logs for debugging purposes.

In order to access the available computational resources, G-PROC provides a set of wrappers as interface with the Condor WMS (Figure 3). This interface can be expanded to support other types of grid resources (e.g. Globus Toolkit 4, EGEE gLite Middleware).

The proposed wrapper interface is able to support the entire Grid specific life cycle: job registration (each task request must be translated into the WMS specific language - i.e the ClassAd language for Condor), job submission (each translated task becomes a Condor specific job ready to be executed), job status (at any time the status of submitted jobs can be retrieved), job logging (when requested, a logging file can be created for debugging reasons) and job retrieval (at the end the output of the job execution can be retrieved).

G-PROC's last component is represented by the processing tools' connectors. The connectors create a bridge between the processing tools and the processing infrastructure. They are required when the arguments of the processing application need to be translated so that they match arguments defined by the WMS description language. They are also linked directly with the WMS and its specific job description language.

## 4.4 Storage architecture

One of the fundamental components of the GiSHEO project is the system responsible for storing and querying the geographical data.

The GiSHEO Data Indexing and Storage Service (GDIS) provides features for data storage, indexing data using a specialized RDBMS, finding data by various conditions, querying external services and for keeping track of temporary data generated by other components. GDIS is available to other components or external parties using a specialized Grid service. This service is also responsible for enforcing data access rules based on specific Grid credentials (VO attributes, etc.).

### 4.4.1 Data storage

The Data Storage component part of GDIS is responsible for storing the data by using available storage backends such as local disk file systems (eg. ext3), local cluster storage (eg. GFS [27], GPFS [25]) or distributed file systems (eg. HDFS, KosmosFS, GlusterFS). One important requirement for the storage component is that data distributed across various storage domains (local or remote) should be exposed through a unique interface.

This is achieved by implementing a front-end GridFTP service capable of interacting with the storage domains on behalf of the clients and in a uniform way (Figure 4). The GridFTP service also enforces the security restrictions provided by other specialized services and related with data access.

The GridFTP service has native access to the Hadoop Distributed File System (http:// lucene.apache.org/hadoop) offering access to data stored inside the internal HDFS file systems and providing the required access control facilities.

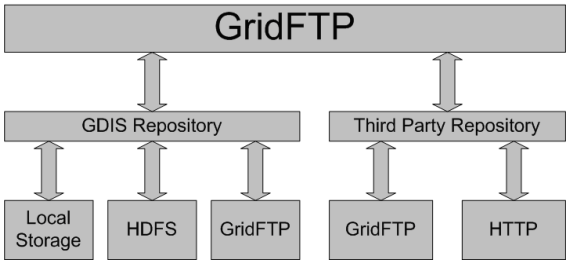



Figure 4: GridFTP interface.

Together with the data indexing components the GridFTP service provides features for manipulating the data repository by providing a basic method for managing data (upload, deletion, retrieval, etc.).

### 4.4.2 Data indexing

The data indexing components enable GDIS to provide fast access to the data offered by the storage component. Data indexing is performed by PostGIS, an OGC (Open Geospatial Consortium) compliant Spatial Database extension for the PostgreSQL RDBMS engine. The PostGIS layer indexes the metadata and location of the geographical data available in the storage layer. The metadata usually represents both the extent or bounding box and the geographical projection of the data (representing the exact geo-location).

The PostGIS Layer provides advanced geographical operations (backed by a GiST index) which allows searching the data by using various criteria including interaction with raw shapes, interaction with shapes representing geopolitical data (country, city, road, etc.) or any other type of geographical data which can be represented in PostGIS. The geo-political data is typically provided by data imported from public sources and extended using the Open Street Map (OSM) data.

The catalogue also maintains data about the type of the datasets. This data is useful not only for retrieving data from the catalogue but also for the workflow engine and execution components. It is used by these components to enforce the data types that can be used with various image processing operations or workflows.

### 4.4.3 Data query

Given the advanced data indexing capabilities of the PostGIS component, GiSHEO provides an advanced and highly flexible interface for searching the project's geographical repository. The search interface is built around a custom query language (LLQL - Lisp Like Query Language) designed to provide fine grained access to the data in the repository and to query external services (TerraServer, GENESI-DR, etc). The syntax of the query language is inspired from the syntax of the LISP language and partially by LDAP filters. The language allows querying the repository both for raster images (Figure 5) and also for various aggregated data or object properties (Figure 6).

```
(select '(url, owner)
 (and
  (or
   (ogc:interacts
    (osm:country "Romania"))
   (ogc:interacts
    (osm:country "Hungary"))
    )
  (gdis:type "RASTER/AERIAL")
 )
)
```

Figure 5: Raster query.

```
; Find cities in Romania
; filter by bbox
(select-place '(name)
 (and
  (ogc:interacts
   (ogc:bbox 16.69 43.97 24.8 48.48))
  (osm:country "Romania")
  (osm:type "city")
 )
)
```

Figure 6: Feature query.

Figures 5 and 6 show in detail how the LLQL syntax looks like. The PostGIS related queries are translated directly to PostGreSQL queries, while the external lookups are resolved prior to submitting the query's to PostGIS.

The GDIS layer also provides a simpler query language called GDIS Query Language (GQL). GQL is suitable for search engines or direct user query. The queries are translated automatically into corresponding SQL queries (through an extensible custom made SQL generation engine).

A simple example of an GQL query is:

```
"place:Timisoara,Timis,Romania
 type:DEM vendor:NASA"
```

When invoked using this query the catalogue returns all datasets that match the criteria (when an explicit operand was not specified the default is `and`).

### 4.4.4 External services

Another set of tasks handled by GDIS are represented by the interaction with external services. In this case GDIS represents a thin middleware layer interacting with external repositories and exposes only one unique interface (similar and possibly integrated with the internal repositories). One example of external back-ends supported by GDIS is represented by the GENESI-DR catalog.

## 4.5 Image processing workflows

Processing satellite images for usage in geographical and historical analysis could require large amount of processing steps involving different image processing transformations. This scenario implies linking the processing algorithms to form a workflow either defined by the user or selected from an already existing list. These algorithms could be located on the same or on different machines spread inside a Grid. In the former case each of them could be exposed as a Web or Grid service. Distributing them across a Grid where each resource exposes several algorithms could help in balancing the resource workload.

Starting from some practical applications involving historical and geographical analysis a few usage scenarios which require more than one image transformation on the source image can be detailed.

In archaeology, for example, assuming that there is a need to identify artificial ancient sites containing human settlements and fortifications from satellite images, the following image transformations could be applied in sequence: gray level conversion, histogram equalization, quantization and thresholding. Applying the resulted image over the initial one allows users to better observe the previously described artificial structures.

Another example could involve identifying linear shapes such as roads or linear fortifications (wave like structures). In this case a workflow made up of the following sequence of operations could provide useful information as output: grayscale conversion, edge detection (e.g. Canny filter), lines detection (e.g. Hough transform).

Related with geography a scenario in which the vegetation index for some particular area needs to be detailed could result from the following operations which also need to be applied in sequence: extract red band, extract infrared band, compute by using the previously obtained images the Normalized Difference Vegetation Index (NDVI).

In the same way detecting changes in river beds or vegetation can be computed by first applying some shape recognition techniques on images taken at different moments in the past and then by overlaying them.

## 4.6 Workflow composition engine

In general workflow image processing transformations are sequential or parallel involving at some point a join or a split. There are few cases which require the use of loops.

As each of the transformation is exposed as a Web or Grid service belonging to a certain resource and due to the dynamic nature of Grids a self adaptive scenario in which tasks would be reassigned (when their corresponding resources might become unable to solve them needs) to be taken into consideration. The natural way to achieve this is by using an Event-Condition-Action (ECA) approach. ECA usually implies a rule governed scenario in which an action takes place on as a result of an event and in case one or more conditions are met. The reason for choosing this paradigm is that it allows the separation of logic represented by rules and data which is represented by objects, declarative programming which is useful for applications focused on what to do instead on how to do it, scalability, centralization of knowledge, etc.

The proposed workflow engine, namely OSyRIS (Orchestration System using a Rule based Inference Solution), detailed in this section is based on DROOLS (http://drools.org) which uses an object oriented version of the RETE [4] algorithm. A simplified workflow language has been built on top of it with the aim of offering a simple yet general rule based language. The following subsection will present in greater detail the language with emphasis on its syntax and its capability of expressing general workflows.

### 4.6.1 Rule-based language

The rule based language, namely SiLK (Simple Language for worKflow), envisioned as part of the GiSHEO project aims at providing a simple yet general language which could be used without modifications in general purpose workflows. The basis of the language are the following: tasks and relations (rules) between them. It is similar with the SCUFL [16] language, but does not rely on XML: it allows the introduction of more workflow specific issues and the ECA approach allows a greater flexibility when expressing data and task dependencies. The following paragraphs will detail these characteristics.

Each task is made up of several mandatory and optional attributes. The mandatory attributes consist of at least one input and one output port. Each task can have several such ports as it could receive input from more than one task and could produce more than one output. The optional attributes are also called meta-attributes. They are not used by the workflow engine and are simply passed over to the service handling the task under the assumption that it can decode and understand them. Meta-attributes are declared by using quotation marks both for the attribute name as well as for the attribute value.

Each task needs to be declared before actually being used. It can be noticed the lack of any mandatory attributes concerning the input or output data type and content. This is due to the fact that the compatibility issues between tasks are resolved at the moment the workflow is created by using methods which are specific to each workflow. These methods should be implemented by the platform running the workflow engine and should not be incorporated inside the workflow description. Because of the nature of the rule engine there is a need for a fictive start task which has the role of a trigger causing the first actual task in the workflow to be executed.

Rules are defined by simply mentioning the events and conditions which should take place in order to trigger the execution of right hand side tasks. Each event is being seen as a completed task and is placed on the left hand side of the rule. Linking the output of left hand tasks with the input of right hand side tasks is accomplished by using variables.

For example the rule:

```
A[a=o1] -> B[i1=a]
```

links the output of task A with the input of task B through variable a. Conditions are placed at the end of the rule and can involve numerical or string variables:

```
A[d=o1] -> B[i1=d] | d<1.
```

In the same way splits and joins made of, but not restricted to, two tasks could be expressed in the same way as the following rules:

```
# synchronized join
A[b=o1],B[c=o1] -> C[i1=b#i2=c]
# parallel split
A[a=o1] -> B[i1=a],C[i1=a],D[i1=a]
```

Loops can be also modeled as in the following example:

```
A[d=o1],B[e=o1] -> A[i1=d#i2=e] | d<1
and
A[d=o1],B[e=o1] -> C[i1=d#i2=e] | d>=1.
```

The former rule expresses the condition to reiterate the loop while the latter expresses the condition to exit the loop. As a remark it should be noticed that when several right hand side tasks need to be activated their execution will take place in parallel. Synchronization between several tasks can also be achieved by adding them into the left hand side of the rule:

```
A[b=o1],B -> C[i1=b].
```

The previous example shows how task A is synchronized with task B and cannot execute until the latter one is completed. Tasks can also have multiple instances. For instance, a rule could produce 5 instances of a task:

```
B[a=o1] -> C[i1=a#instances=5]
```

with the default number of instances being one. Instances of left hand tasks are usually consumed when the rule fires

```
B[a=o1#consume=true] -> C[i1=a].
```

However this feature is optional with the default behaviour being consume. It is useful for example when there are $n$ rules with the same left hand task which can fire but only $k$ rules are needed. In this case having $k$ instances of that task could prove useful. Another utility for this feature could be the case when a rule is needed to be fired several times. Multiple task instances allow users to express workflows related to natural processes such as those involved in chemical reactions [14] and cellular membranes [18].

Several meta-attributes including datagroup, dataset, processing and argument-list need to be introduced in order to fully define a GiSHEO workflow. The meta-attributes are used to identify the image to be processed, the operation and the arguments to be used. For example the datagroup and dataset attributes identify the group and the set inside the group to which the image belongs. The processing attribute identifies the operation to be applied to the image. Its value follows a C-like prototype format with return type, operation name and argument list. The attribute-list specifies the optional attributes used by the operation. It is a list where the values are pairs in the form <name=value>. Each value is separated by a # sign. The name inside the pair must match the name given to the attribute in the processing description.

```
# Define a fictive start task.
# Needed for initial activation
A0:= [o1:output="FTP address",
     "instances"="1"];
# The following two tasks belong to the
# processing workflow
A:= [i1:input,o1:output,"datagroup"="ID",
 "dataset"="ID",
 "processing"="outIMG_1 grayscale(inIMG_1)",
 "argument-list"=""];
B:= [i1:input,o1:output,"datagroup"="ID",
 "dataset"="ID", "processing"=
 "outIMG_1 canny(inIMG_1#canny1#
  aperture_size#canny2)",
 "argument-list"=
 "<canny1=80>#<aperture_size=3>#
  <canny2=120>"];
C:=[i1:input,o1:output,"datagroup"="ID",
 "dataset"="ID","processing"=
 "outIMG_1 hough_lines(inIMG_1#min_line#
  min_gap#hough_treshold)",
 "argument-list"=
 "<min_line=20>#<min_gap=10>#
 <hough_treshold=100>",
 "isLast"="true"];

# Start rule: compute grayscale
# from the initial image
A0[a=o1] -> A[i1=a];
# Compute Canny from the grayscale image
# A[a=o1] -> B[i1=a];
# Compute a Hough transform from
# the Canny image
B[a=o1] -> C[i1=a];
```

Figure 7: Workflow example using SiLK

Figure 7 shows a complex example, for detecting linear structures by combining several basic operations as: grayscale conversion, edge detection with Canny filter and lines detection with Hough transform. It also presents the required attribute values.

As previously mentioned using an ECA approach allows for creating adaptive workflows which can react to changes either in the configuration of the Grid or inside the workflow itself. Changes inside the Grid are handled by creating specific rules which allow resource selection based on various task scheduling criteria. Modifications of the workflow are usually accomplished either by inserting or retracting at runtime rules belonging to it or by modifying the executor of the task in case a better one is found. It is very hard or almost impossible to express adaptivity by using classic workflow languages such as WS-

BPEL (http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html), SCUFL [16], JSDL (Job Submission Description Language (JSDL), http://www.gridforum.org/documents/GFD.56.pdf), DAGs (Directed Acyclic Graphs) or Petri Nets. Most of the previously listed languages are either too verbose and strict due to the use of XML or their structure is predetermined by the corresponding DAG or Petri Net which cannot be changed at runtime.

Different from the previous classic approaches are the ones based on ECA. The nature inspired approach based on chemical reactions and described in paper [14] falls into this category. Moreover it allows self adaptation to changes by using the High Order Chemical Language also described in the paper. Other approaches include the AGENTWork [12], a workflow engine based on a language supporting rule based adaptation and which is used in medical practices. The paper [32] presents a Condition-Action (CA) language and platform called FARAO, based on an Adaptive Service Oriented Architecture (ASOA), which has the aim of supporting the development of adaptable service orchestrations.

### 4.6.2 Particularities of the workflow engine

Each of the non ECA workflow languages listed in the previous subsection has a workflow enactment engine built on top of it. WS-BPEL relies for example on Active-BPEL (http://www.active-endpoints.com/active-bpel-engine-overview.htm), SCUFL uses Taverna [16] while Condor and Pegasus [2] use DAGman to execute workflows expressed as DAGs.

Not all of the previously mentioned workflow systems provide however sufficient functionalities to cover system and logical failures that could occur when running the workflow. To cope with this issue a workflow management platform has been built on top of the workflow language described in the previous subsection. Its role is to execute the workflow and to handle system and logical failures by inserting additional rules inside the initial rule database. These new rules will be expressed by using the same language and will be interpreted by using DROOLS.

The workflow platform is also responsible for checking the integrity of the workflow syntax. However it does not check the compatibility between different nodes. This problem is left to the specialized system which will use the platform.

Given existing ECA based workflow engines such as AGENTWork [12] which already deal with adaptiveness issues, the aim of this platform is not only to provide a simpler yet general language but to offer solutions in form of workflows to different problems such as archaeological and geographical image related transformations. In this direction the goal is to build a workflow backwards from the desired result to the input requirements given by the user, and to provide the user with one or more solutions from which he/she can use the most appropriate one.

## 4.7 Example for visualising linear shapes using the GiSHEO interface

GiSHEO offers a web portal from which the user can select images, apply various operations on them and visualize the results. In what follows we present an example that shows all the required steps starting with the selection of the image and until the final image is obtained. The processing we selected follows the workflow example described in Figure 7 from previous subsection.

First the desired geographical area needs to be selected by introducing a GQL expression (see subsection 4.4.3 for details) as shown in Figure 8. After the area is selected the user can choose the area of interest (see Figure 9). This is done by simply using the mouse to select a rectangular region that will be used when image processing operations are invoked.

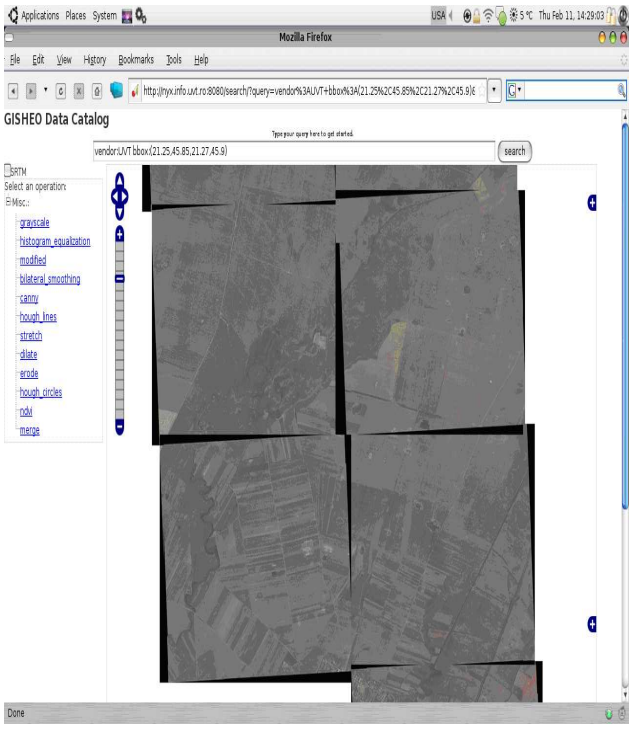



Figure 8: Selecting the images by geographical position.

Once the desired area has been selected the user can proceed by choosing the operations he/she would like to perform on the images. As our example follows the workflow specified in Figure 7, the first operation the user will select is the grayscale conversion (see Figure 10). Once a name for the result image is set the request is sent to the server which triggers the execution. The result can be visualized as soon as it is available (as seen in Figure 11).

The other requests are for the Canny edge detector (see Figures 12 and 13) and for the Hough transform for lines detection (see Figures 14 and 15).

Once the final operation is completed the user can clearly see the linear shapes indicated by red lines (see Figure 15). By comparing this final image with the initial one (see Fig-

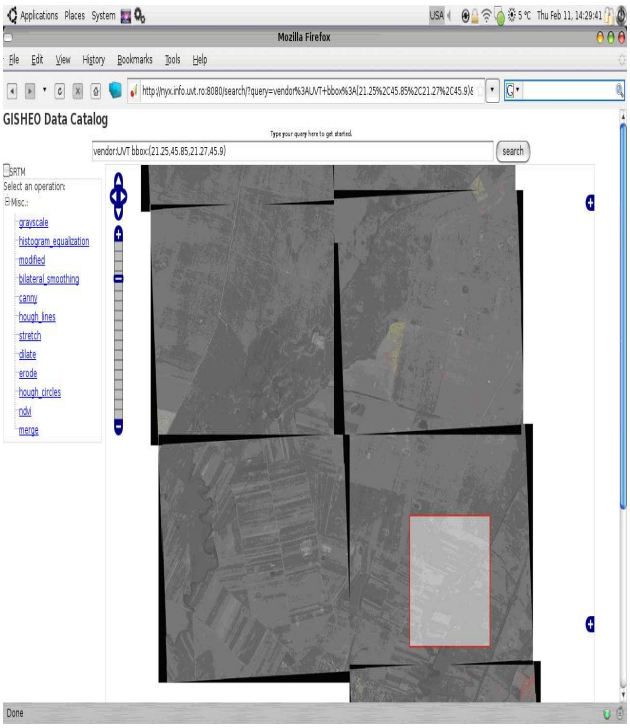



Figure 9: Selecting the area of interest inside the selected images.

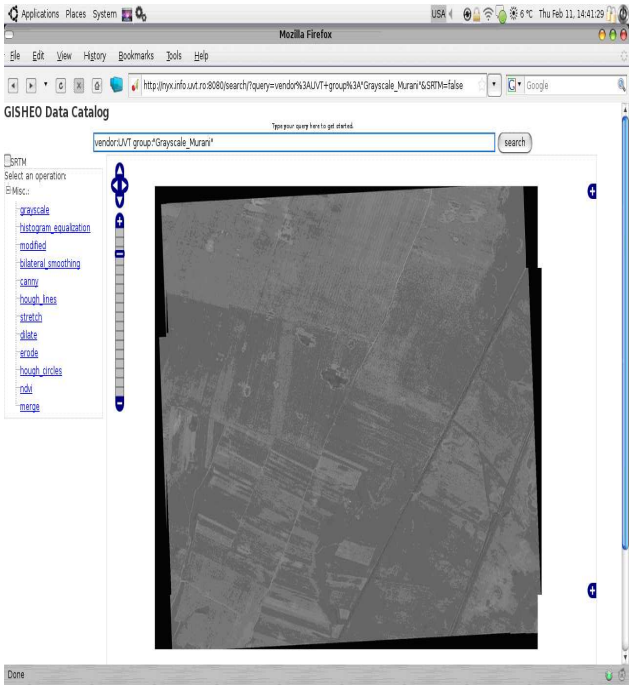



Figure 11: Visualizing the result of the grayscale conversion.

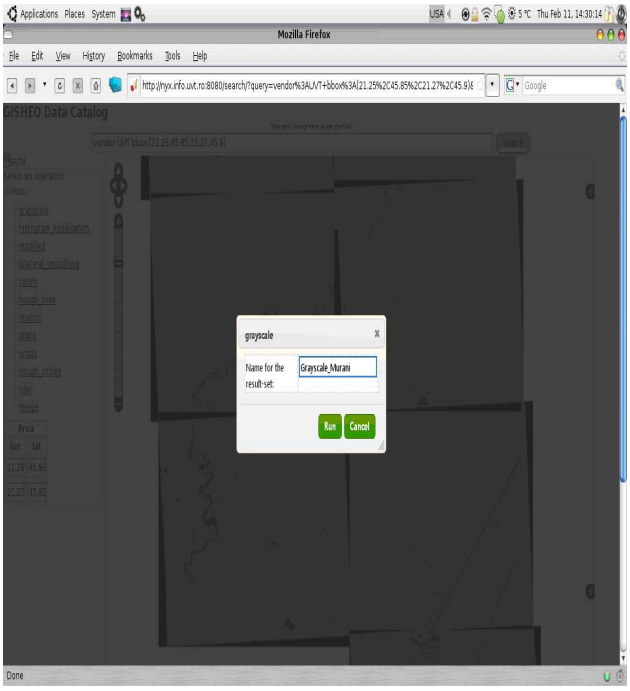



Figure 10: Applying the grayscale conversion.

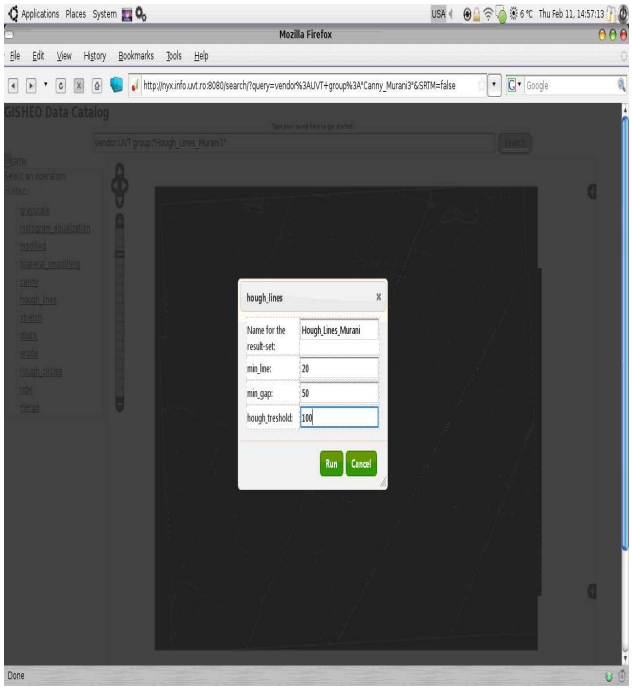



Figure 12: Applying the Canny edge detector on the grayscale image.

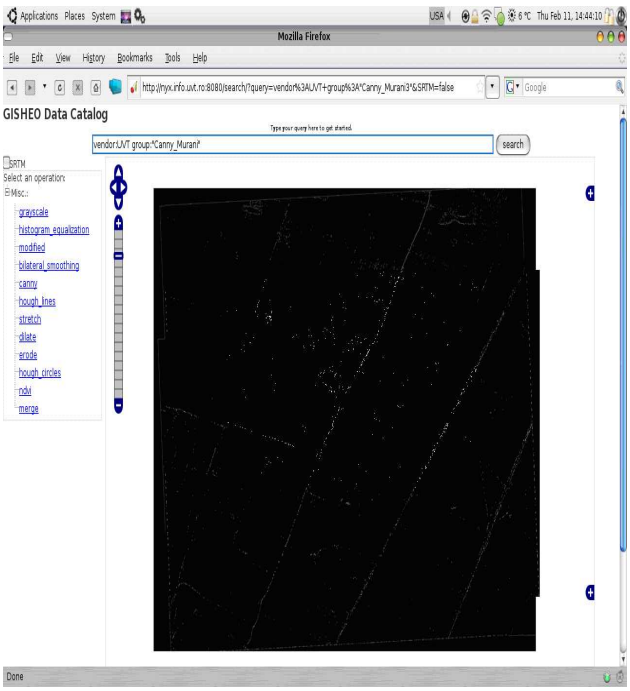



Figure 13: Visualizing the result of the Canny edge detector.

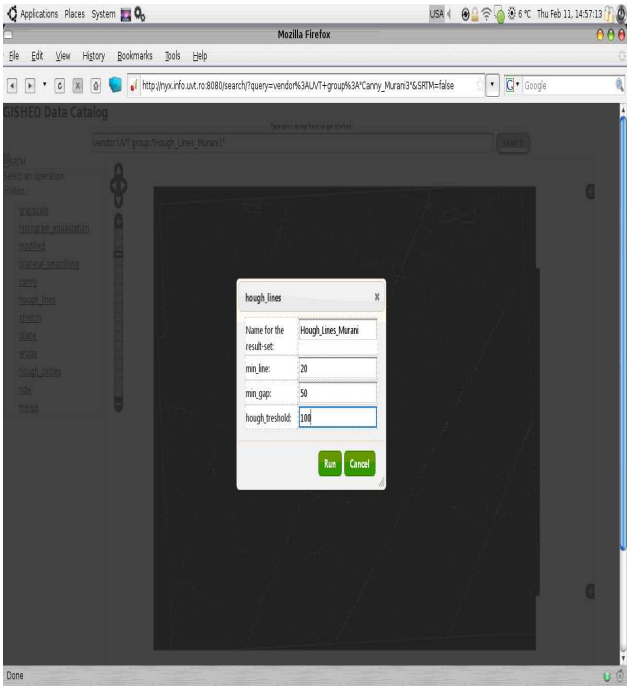



Figure 14: Applying the Hough transform on the Canny image.

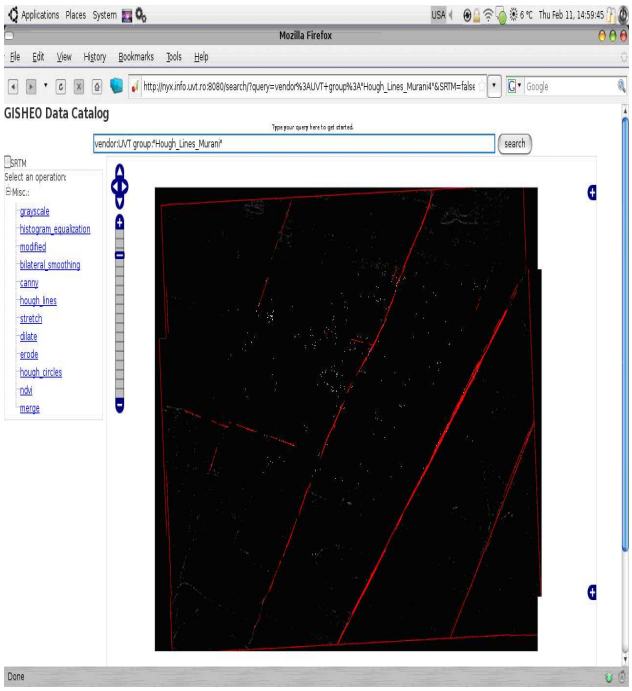



Figure 15: Visualising the result of the lines detection operation.

ure 8) the importance of selecting a proper sequence of processing steps when a specific study (e.g. the identification of important lines inside an image) is undertaken becomes obvious.

The user has two choices in what concerns the steps required to obtain a final image. By using the interface he/she can either build a workflow made of several basic operations or send basic operations to the server one at a time. In either case the requests are processed by the workflow engine as it treats even basic operations as workflows composed of a single task.

If several data are selected by the user, the actions will be undertaken on all of them in parallel.

# 5 Conclusions

Starting from a debate of the current issues in satisfying the user and system requirements in Earth observation systems, as well as from the need of training in Earth observation field, a service oriented platform was proposed and described in this paper. It uses the latest technologies both in distributed systems as well as in Earth observation systems and therefore can be considered as a proof of concept of what is currently possible to be done with available technologies.

During the platform design and implementation, innovative solutions were proposed like the custom query language or the specific rule-based language for the workflow engine. While these solutions are not in line with the current standards, they proved to be more efficient in the implementation, as well as to respond better to the require-

ments of the Earth observation system and to obtain a fast response from the distributed platform.

## Acknowledgement

This research is supported by ESA PECS Contract no. 98061 GiSHEO – On Demand Grid Services for High Education and Training in Earth Observation.

# References

[1] Aloisio, G., Cafaro, M. (2003). A dynamic Earth observation system. *Parallel Computing* 29 (10), pp. 1357-1362.

[2] Deelman, E., Singh, G., Su, M-H., Blythe, J., Gil, Y., Kesselman, K., Mehta, G., Vahi, K., Berriman, G.B., Good, J., Laity, A., Jacob, J.C., Katz, D.S. (2005). Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming* 13, pp. 219–237.

[3] eIRG Data Management Task Force (2009). *e-IRG report on data management*. http://www.e-irg.eu.

[4] Forgy, C. (1982). Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem, *Artificial Intelligence* 19, pp. 17–37.

[5] Frincu, M.E., Panica, S., Neagul, M., Petcu, D. (2009). Gisheo: On demand Grid service based platform for EO data processing, *Procs. HiperGrid'09*, Politehnica Press, pp. 415-422.

[6] Fusco, L., Cossu, R, Retscher, C. (2008). Open Grid services for Envisat and Earth observation applications. *High Performance Computing in Remote Sensing*, Plaza, A., Chang, C. (Eds.), pp. 237–280.

[7] Hey, T., Tansley, S., Tolle, K. (2009). *The fourth paradigm. Data-itensive scientific discovery*. Microsoft research.

[8] Gasster, S.D., Lee, C.A., Palko, J.W. (2008). Remote sensing Grids: architecture and implementation. *High Performance Computing in Remote Sensing*, Plaza, A., Chang, C. (Eds.), pp. 203–236.

[9] Gorgan D., Stefanut T., Bacu V. (2009). Grid based training environment for Earth observation. *LNCS* 5529, pp. 98–109.

[10] Larson, J.W., Norris, B., Ong, E.T., Bernholdt, D.E., Drake, J.B., Elwasif W.E., Ham, M.W., Rasmussen, C.E., Kumfert, G., Katz, D., Zhou, S., DeLuca, C., Collins, N.S. (2004). Components, the common component architecture, and the climate/weather/ocean community, *Procs. 84th AMS Annual Meeting*, Seattle.

[11] Lee, C, A. (2008). An introduction to Grids for remote sensing applications. *High Performance Computing in Remote Sensing*, Plaza A., Chang C.(Eds.), pp. 183–202.

[12] Muller, R., Greiner, U., Rahm, E. (2004). AGENTWORK: a workflow system supporting rule-based workflow adaptation. *Data & Knowledge Engineering*, 51 (2), pp. 223–256.

[13] Neagul, M., Panica, S., Petcu, D., Zaharie, D., Gorgan, D. (2009). Web and Grid services for training in Earth observation, *Procs. IEEE IDAACS'09*, IEEE Computer Society Press, pp. 241-246

[14] Nemeth, Z., Perez, C., Priol, T. (2005). Workflow enactment based on a chemical metaphor. *Procs. SEFM 2005*, IEEE Computer Society Press, pp. 127-136.

[15] Nico, G., Fusco, L., Linford, J. (2003). Grid technology for the storage and processing of remote sensing data: description of an application. *SPIE* 4881, pp. 677–685.

[16] Oinn, T., Greenwood, M., Addis, M., Apldemir, M.N., Ferris, J., Glover, K., Globe, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Popock, M.R., Senger, M., Stevens, R., Wipat, A., Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience* 18 (10), pp. 1067–1100.

[17] Panica, S., Neagul, M,, Petcu, D., Stefanut, T., Gorgan, D. (2009). Desiging a Grid-based training platform for Earth observation. *Procs. SYNASC 08*, IEEE Computer Society Press, pp. 394–397.

[18] Paun, G., Grzegorz, R. (2002). A guide to membrane computing. *Theoretical Computer Science* 287 (1), pp. 73–100.

[19] Petcu, D., Challenges of data processing for Earth observation in distributed environments. *SCI* 237, Springer, pp. 9–19.

[20] Petcu, D., Gorgan, D., Pop, F., Tudor, D., Zaharie, D. (2008). Satellite image processing on a Grid-based platform. *International Scientific Journal of Computing* 7 (2), pp. 51–58.

[21] Petcu, D., Zaharie, D., Neagul, M., Panica, S., Frincu, M., Gorgan, D., Stefanut, T., Bacu, V. (2009). Remote sensed image processing on Grids for training in Earth observation. *Image Processing*, Chen Yung-Sheng (ed.), In-Teh, pp. 115–140.

[22] Plaza, A., Plaza, J., Valencia, D. (2006). Ameepar: Parallel morphological algorithm for hyperspectral image classification in heterogeneous NoW. *LNCS* 3391, pp. 888–891.

[23] Plaza, A., Chang, C. (Eds.) (2008), *High Performance Computing in Remote Sensing*. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton.

[24] Portela, O., Tabasco, A., Brito, F., Goncalves, P. (2008). A Grid enabled infrastructure for Earth observation. *Geophysical Research Abstracts* 10, pp. 1.

[25] Schmuck, F., Haskin, R. (2002). GPFS: A shared-disk file system for large computing clusters, *Procs FAST'02*, pp. 231–244.

[26] Sekiguchi, S. Tanaka, Y. Kojima, I. Yamamoto, N. Yokoyama, S. Tanimura, Y. Nakamura, R. Iwao, K. Tsuchida, S. (2008). Design principles and IT overviews of the GEOGrid. *IEEE Systems Journal* 2 (3), pp. 374–389.

[27] Soltis, S.R., Erickson, G.M., Preslan, K.W., O'Keefe, M.T., Ruwart, T.M. (1997). The Global File System: a file system for shared disk storage. *IEEE Transactions on parallel and distributed systems*.

[28] Szalay, A.S., Blakeley, J.A. (2009). Gray's laws: database-centric computing in science. *The fourth paradigm. Data-itensive scientific discovery*, Hey, T., et al. (eds.), Microsoft research, pp. 5–12.

[29] Teo, Y.M., Tay, S.C., Gozali, J.P. (2003). Distributed geo-rectification of satellite images using Grid computing. *Procs. IPDPS'03*, IEEE Computer Society Press, pp. 152–157.

[30] Votava, P., Nemani, R., Golden, K., Cooke, D., Hernandez, H. (2002). Parallel distributed application framework for Earth science data processing, *Procs. IGARSS 02*, IEEE Press, pp. 717–719.

[31] Wang, J., Sun, X., Xue, Y., Hu, Y., Luo, Y., Wang, Y., Zhong, S., Zhang, A., Tang, J., Cai, G. (2004). Preliminary study on unsupervised classification of remotely sensed images on the Grid. *LNCS* 3039, pp. 981–988.

[32] Weigand, H., Heuvel, W.J.A.M., Hiel, M. (2008). Rule-based service composition and service-oriented business rule management. *Procs. REMOD*, vol. 342, CEUR, pp. 1–12.

[33] Yang, X.J., Chang, Z.M., Zhou, H., Qu, X., Li, C.J. (2004). Services for parallel remote-sensing image processing based on computational Grid. *LNCS* 3252, pp. 689–696.

[34] Yang, C., Guo, D., Ren, Y., Luo, X., Men, J. (2005). The architecture of SIG computing environment and its application to image processing. *LNCS* 3795, pp. 566–572.

[35] Yunck, T., Wilson, B., Braverman, A., Dobinson, E., Fetzer, E. (2008). GENESIS: the general Earth science investigation suite. *Procs. ESTC'04*, http://sciflo.jpl.nasa.gov/.

# Dynamic Process Integration Framework: Toward Efficient Information Processing in Complex Distributed Systems

Gregor Pavlin and Michiel Kamermans Thales Nederland B.V., D-CIS Lab Delft, Nederland
E-mail: gregor.pavlin@d-cis.nl, michiel.kamermans@d-cis.nl

Mihnea Scafeş
University of Craiova, Romania
scafes_mihnea@software.ucv.ro

**Keywords:** service oriented architectures, distributed reasoning, negotiation, multi agent systems

**Received:** February 22, 2010

*The Dynamic Process Integration Framework (DPIF) is a service oriented approach which supports efficient creation of distributed systems for collaborative reasoning. The DPIF is relevant for an important class of contemporary applications requiring efficient and reliable processing of large quantities of heterogeneous information. An example of such an application is situation assessment in complex decision making processes in dynamic environments. The DPIF supports (i) a systematic encapsulation of heterogeneous processes and (ii) negotiation-based self configuration mechanisms which automate creation of meaningful workflows implementing complex collaborative reasoning processes. The resulting systems support processing based on rich domain knowledge while, at the same time, the collaboration between heterogeneous services requires minimal ontological commitments.*

*Povzetek: Opisan je nov pristop v procesiranju informacij v kompleksnih porazdeljenih sistemih.*

## 1 Introduction

This paper introduces a service oriented architecture supporting complex collaborative processing in distributed systems. The presented approach is relevant for many contemporary applications that require reasoning about complex processes and phenomena in real world domains. For example, in crisis management advanced information processing is required for (i) identification of critical situations, (ii) impact assessment which takes into account possible evolution of physical processes, (iii) planning and evaluation of countermeasures and (iv) decision making. This can be achieved only through adequate processing of large quantities of very heterogeneous information, based on rich expertise about different aspects of the physical world. Such processing requirements typically exceed the cognitive capabilities of a single human expert; an expert typically does not have knowledge of all the relevant mechanisms in the domain and cannot process the huge amounts of available information. On the other hand, full automation of decision making processes in such settings is not feasible, since the creation of the required domain models as well as the inference are intractable problems. Specifically, automated inference processes involve many variables and relations with accompanying representation and inference mechanisms.

Such settings require collaborative processing based on a combination of automated reasoning processes and cognitive capabilities of multiple human experts, each contributing specific expertise and processing resources. Key to effective combination of human-based expertise and automated reasoning processes is a framework which allows that each piece of the relevant information is adequately considered in the final processing outcome. The main elements of such a framework are:

1. Standardized formats that facilitate sharing of heterogeneous information.
2. Filtering services which provide stakeholders in a decision making process with the right information at the right moment in time. In principle, filtering services must transform heterogeneous data to more abstract information types and route the information to the consumers who can make use of it.

In this paper we focus on the second element, which is tackled with the help of the Dynamic Process Integration Framework (DPIF). The DPIF supports seamless integration of heterogeneous domain knowledge and processing capabilities into coherent collaborative processes. Processes are encapsulated by software agents, each using identical communication and collaboration mechanisms. The DPIF combines Multi Agent Systems (MAS) and a service oriented paradigm in new ways which facilitate implementation of hybrid collaborative reasoning systems with emergent problem solving capabilities. In contrast to traditional MAS approaches [29, 11], the DPIF facilitates integration of human cognitive capabilities right into problem

solving processes in workflows; humans are not mere users of an automated system, but contribute the processing resources. From the problem solving perspective, the humans can be viewed as a specific type of processing modules, integrated into the overall processing system via assistant agents.

In general, a key to efficient collaborative processing in complex domains are *workflows*, in which peers with different processing capabilities exchange relevant information [25, 4, 2, 3, 27]. Often such workflows are created through dynamic composition of services [5, 25]. In this way the systems can adapt at runtime and deliver tailored solutions to specific problems. Creation of workflows is often based on centralized planning and ontologies describing relations between different services. Approaches exploiting centralized service composition have been successfully used in many relevant applications, such as business process modeling [15, 22], scientific querying [4], planning/booking systems [22] and simulation and scientific grid computing [14, 8].

For the challenges addressed in this paper, however, centralized approaches to composition of workflows and central description of relations between services are neither practical nor necessary. Namely, we are dealing with systems in which artificial agents and experts collaboratively process large quantities of heterogeneous information. In such settings construction of centralized ontologies describing services as well as all relations between the handled information types is likely to be very hard or even intractable. Similarly, centralized construction of workflows might not be practical, since the constellations of available services (i.e. automated processes or experts) change frequently at runtime. Given the time constraints, communication of all changes and system states to a central workflow composition process might not be feasible. Thus, the resulting, centrally composed workflows are likely to incorporate only a subset of all services relevant for a problem at hand.

It turns out that efficient solutions to service composition can be obtained if we explicitly take into account the characteristics of the problem. In particular, many of the challenges associated with centralized approaches to service composition and definition can be avoided if the resulting systems are used in organizations that can be characterized as Professional Bureaucracy [26]. In such organizations the skills are standardized, the control is decentralized to a great extent and the experts and/or automated processes do not have to share domain and processing knowledge. In other words, complex problems can be efficiently solved with the help of systems of *loosely coupled* experts and automated processes without a centralized control. Collaboration in such systems can be achieved through service discovery based on local domain knowledge. Therefore, we introduce an approach which does not require centralized service ontologies and centralized service composition methods. Moreover, fully decentralized configuration of meaningful processing workflows can be achieved by using *local knowledge of relations* between different services. The approach requires simple ontologies which serve primarily for the alignment of the semantics and syntax of messages exchanged between the processes in workflows. In addition, the relations between types of services are captured by local functions, dispersed throughout a system of modules providing the services. We show that meaningful workflows supporting globally coherent processing can be created by using only local domain knowledge. In this way we obtain systems which support processing based on rich domain knowledge while, at the same time, the collaboration between heterogeneous services requires minimal ontological commitments [9].

Overall, by using the DPIF encapsulation techniques and methods, arbitrary processing services can easily be made composable and negotiable.

The paper is organized as follows. In section 2 a rationale for decentralized collaborative reasoning in workflows is provided and the basic features of the DPIF are introduced. Section 3 explains how meaningful workflows between heterogeneous processes can be dynamically implemented, without centralized knowledge of relations between the variables in the reasoning processes. In particular, we emphasize a combination of service composition, decentralized validation methods and advanced negotiation mechanisms, which allow a systematic incorporation of various criteria into the workflow creation processes. Section 4 introduces basic DPIF architecture principles while section 5 introduces an approach to efficient construction of service ontologies by exploiting the local domain knowledge captured by different DPIF agents. Section 6 provides conclusions and plans for the future work.

## 2 Collaborative processing

Reasoning about domains requires knowledge about typical dependencies (i.e. relations) between relevant phenomena in these domains. By using (i) domain models capturing the relations between relevant phenomena and (ii) evidence based on observations of certain phenomena, we can assess (i.e. estimate) the states of the domain that cannot be observed directly. In addition, with the help of models, the future evolution in domains can be predicted. However, in complex domains reliable reasoning can be achieved only by relating large quantities of information of very heterogeneous types with very different semantics. Such dependencies can be explained only through complex models.

Irrespectively of the used models, it is unlikely that in complex domains a single model designer or an expert understands all the relevant phenomena and relations between them. Instead, a set of relatively simple domain models will exist, with each model capturing a small subset of the relevant variables and the corresponding relations. Thus, reasoning based on the relations between the entire available evidence can be achieved only by combining simpler processes, each using a limited domain model. The out-

puts of simple processes are used as inputs of other simple processes. In other words, the reasoning is based on data-driven workflows established between heterogeneous processes. In such workflows difficult problems can be solved through collaboration of heterogeneous processes, each focusing on a relatively small subset of relevant aspects in the targeted domain.

We illustrate such processing by using an example from the environmental management domain. In a chemical incident a leaking chemical starts burning which results in harmful fumes. The impact of a resulting fumes is mitigated through a collaboration of experts captured by figure 1. We assume that the factory staff (FS) at the incident have an overview of the current state of the damaged system; FS can estimate the quantity of the escaping chemical and its type. This information can be used by a chemical expert at the incident location (CE1) to estimate the type and quantity of toxic fumes resulting from the fire. By knowing the location of the fire, the meteorological conditions, and the quantity and type of the produced fumes, chemical expert (CE2) can estimate the zones in which the concentration of the toxic gases have exceeded critical levels and identify areas which are likely to be critical after a certain period of time. The CE2 makes use of the domain knowledge about the physical properties of the gases and their propagation mechanisms. In addition, it guides fire fighter teams (MT) which can measure gas concentrations at specific locations in order to provide feedback for a more accurate estimation of the critical area. A map showing the critical area is supplied to a health expert (HE) who uses the information on population obtained from the municipality to estimate the impact of the toxic fumes on the human population in case of exposure. Finally, the estimated impact on the population is supplied to decision makers, who choose between no action, evacuation and sheltering. This decision also considers estimated time and costs in case of an evacuation of people from the danger zone as well as the estimated costs and duration of a preventive evacuation. The former estimate is provided by the fire brigade representatives while the latter estimate is supplied by the police department. In other words, in such a system, each expert can be viewed as a module providing predefined services which in turn require services from other experts. Thus, the situation analysis in the presented example can be viewed as a workflow between different, weakly coupled processing services, each specialized in specific aspects of the domain. Moreover, a processing service can be provided by a human (e.g. a chemical expert analyzing the extent of the contamination) or by an automated reasoning process (e.g. detection of gases based on automatic fusion of sensor data). Note that, for the sake of clarity, the used example is a significant abstraction of real crisis management processes.

Moreover, the example can be seen as a class of problems where we have to reason about a situation which can be viewed as a specific combination of known types of events and processes, each understood by a human expert or modeled by an artificial agent. For example, the way chemicals burn and react, the effects of exposure to toxic fumes, evacuation approaches in hospitals and schools, etc. are independent of the location and time. Therefore, we can obtain general knowledge about such processes and phenomena which can be used for the analysis in any situation involving such phenomena. In other words, a mapping between experts and artificial agents, on the one hand, and event types, on the other hand, can be made a priori; we can assign roles to different experts and artificial agents based on their domain knowledge and models.

Since each situation (e.g. chemical incident) is a unique combination of known types of events, a specific workflow consisting of a particular combination of processing nodes is required for adequate situation assessment. In addition, due to unpredictable sequences of events it is impossible to specify an adequate workflow a priori. For example, given the wind direction, experts for the evacuation of hospitals and schools might be needed. However, if the gas is blown to the open sea instead, no evacuation experts are needed in the situation assessment process.

Clearly, a major challenge is creation of adequate workflows which correctly integrate the relevant processes and support globally coherent processing in decentralized collaborative systems. In the following text we explain how this can be achieved in an efficient and sound way.

## 2.1 Dynamic process integration framework

The Dynamic Process Integration Framework (DPIF) supports decentralized creation of workflows that facilitate collaborative problem solving. The DPIF is a service-oriented approach (SOA) which supports efficient composition of very heterogeneous processing services provided by different experts and automated reasoning processes. In the context of the DPIF, the information processing is abstracted from human or machine instances; a reasoning process is either provided by a human expert or an automated system implemented by a software agent. Each process provides a well defined reasoning service in the form of an estimate, prediction, cost estimate, etc. The inputs for each of such processes are provided by other processes or by direct observations (i.e. sensor measurements and reports from humans).

A human expert or an automated inference process is represented in the system by a software agent, a functional (i.e. processing) module which (i) supports standardized collaboration protocols and (ii) allows incorporation of arbitrary reasoning approaches. In other words, the agents provide a uniform communication/collaboration infrastructure allowing seamless combination of heterogeneous processes provided by human experts or implemented through AI techniques. Each agent registers in the DPIF-based system (i) the services supported by its local processing capabilities and (ii) the required inputs, i.e. types of services that should be provided by other agents in the system.

By using the registered services, agents distributed

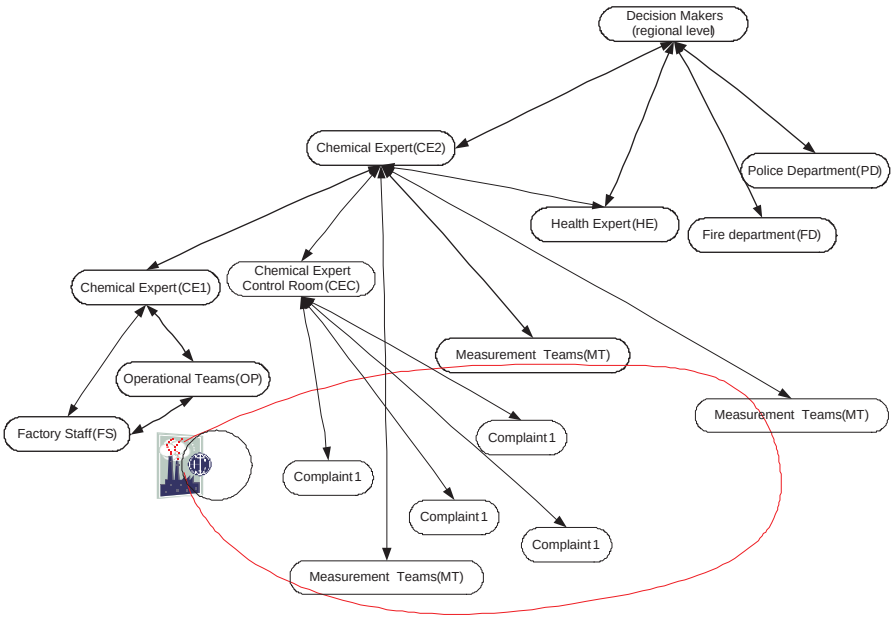



Figure 1: A workflow in a decision making process. Arrows denote information flow between different experts, each processing relevant information of different types. The circled region denotes the initial estimate of the area where concentration is likely to be critical.

throughout different networked devices can autonomously form workflows in which heterogeneous processes introduce collaborative reasoning. The configuration of workflows is based on the relations between services captured by local models; each agent *knows* what service it can provide and what it needs to do this. This local knowledge is captured by the relations between the variables in partial domain models. Thus, *no centralized ontology describing relations* between different services of various agents is required, the creation of which is likely to be intractable.

*In other words, globally coherent collaborative processing is possible by combining local processes, without any global description of relations between inputs and outputs.*

In the following discussion we focus on (i) principles for the creation of valid workflows based on the local processing capabilities of different agents and (ii) describe the basic elements of the DPIF architecture.

# 3 Processing workflows

A basic workflow element in the DPIF is a local process. Moreover, in the following discussion the term local process refers to a reasoning process provided either by a human expert or an automated system implemented by a software agent. Each local process corresponds to a function $F : \{X_1, ..., X_n\} \rightarrow Y$, mapping values in a domain $\{X_1, ..., X_n\}$ to values of some variable of interest $Y$. The value of $Y$ for particular values of arguments is given by $y = f_y(x_1, ..., x_n)$.

Such functions can be either explicit, based on some rigorous theory, or implicit, when they are provided by humans or sub-symbolic processes, such as for example neural networks. An example of a mathematically rigorous mapping is the function $x_{CE1} = f_{x_{CE1}}(x_{FS})$, an explicit formula describing the relations between the fume volume per time unit represented by $X_{CE1}$ and the escape rate of chemicals denoted by $X_{FS}$. This function is used by the Chemical Expert CE1 in figure 1. An implicit mapping, on the other hand, is performed by the health expert (HE) who estimates the critical regions with respect to the impact on the residents. HE interprets information about critical concentration $X_{CE2}$ in combination with information on population distribution $X_{POP}$ by using an implicit function $x_{HE} = f_{x_{HE}}(x_{CE2}, x_{POP})$.

## 3.1 From local to global processes

An expert or an artificial agent often cannot observe values of certain variables; i.e. variables cannot be instantiated. Instead, the inputs to the local function are supplied by other processes forming a collaborative workflow (see section 2). Thus, the inputs to one function are outputs of other functions used by the information suppliers. From a global perspective this can be seen as a function composition; in a function, each variable which cannot be instantiated is replaced by a function. This process continues until a function is obtained in which all variables are instantiated, i.e. all free variables in the resulting nested function have been reduced to direct observations. In this way, a global function emerges as different processes are connected in a workflow. The resulting function is a com-

posite mapping between directly observable variable states and hidden variables of interest.

In other words, a workflow in a DPIF system corresponds to a full composition of functions, in which each variable replaced by a function corresponds to a required service. This yields the value of the variable of interest. Let's assume an example with six service suppliers shown in figure 2(a), using the following functions:

$$\begin{aligned} x_a &= f_a(x_b, x_c), x_b = f_b(x_d), x_c = f_c(x_e, x_f), \\ x_d &= f_d(x_g), x_e = f_e(x_h), x_f = f_f(x_i). \end{aligned}$$

then the workflow supporting collaborative computation of the value for $x_a$ corresponds to the composite function

$$f_a(f_b(f_d(x_g)), f_c(f_e(x_h), f_f(x_i))) \quad (1)$$

*It is important to bear in mind that in DPIF no explicit function composition takes place in any of the agents. Instead, the sharing of function outputs in a workflow corresponds to such a composite function; i.e. a workflow models a (globally emergent) function, mapping all observations of the phenomena of interest (i.e. evidence) to a description of some unknown state of interest.*

Each workflow corresponds to a system of systems, in which exclusively local processing leads to a globally emergent behavior that is equivalent to processing the fully composed mapping from direct observations to the state of the variable of interest.

## 3.2 Decentralized validation of workflow structures

The functions in a workflow can be described through different ad-hoc or theoretically sound frameworks (such as for example Bayesian reasoning). However, the relation between workflows and function composition has several interesting properties regardless of the theory used to describe the functions. Workflows, in general, can be represented by directed graphs[1] whose topologies have important implications regarding the reasoning (e.g. see figure 2(b)). Particularly important concepts in graphs are loops and cycles. Loops occur when there is more than one way to travel between two nodes by following the directed links in a graph, i.e. there exist multiple directed paths between two nodes (e.g. see figure 3(b)). Cycles occur if a node can be revisited by following a directed path (see figure 4(c)).

Loops and cycles provide an important insight into the dependencies between different services (i.e. processes) and thus the use of information in a distributed system.

If a process uses multiple inputs obtained from services belonging to a *loop*, then these inputs might have been generated by distributed processes using the same information. In other words, these inputs may not be independent, which might lead to data incest [7], if the inputs are not properly treated. In case of data incest, the same information is reused multiple times which is likely to lead to misleading conclusions. In case of rigorous reasoning approaches, such as Bayesian networks, this problem can be avoided by clustering of variables [12], which allows correct probabilistic reasoning even if the graphs contain loops. In any case, we should be aware of loops in systems implementing distributed functions.

While loops may be permissible, as there are various ways to deal with them if detectable, *cycles* are not permitted in workflows that implement inference, as they would lead to one or more components in a workflow processing misleading data. That is, the system is generating outputs without adding new information to the system. In a data-driven approach this leads to a self perpetuating system which is likely to produce outputs which do not reflect the reality. Figure 4(b) shows an example of a system with a directed cycle, where agent $A$ keeps supplying inputs to agent $C$, which in turn produces new inputs for $A$. By looking at the composed function represented by a workflow, cyclical workflows would lead to infinite composition sequences: if some function is directly or indirectly dependent upon itself, then this would lead to an infinitely recursive composition of the full function which is likely to result in misleading outputs.

Therefore, an integral part of the DPIF is a cycle detection mechanism based on the cycle detection principles introduced in [13]. The used approach allows a fully decentralized detection of cycles based on peer to peer communication. These principles were initially used in modular systems using Bayesian networks [13], but the technique discussed is easily carried over to a more generic domain.

In order to effect cycle-free workflows, each DPIF agent must be aware of the outputs of other agents that influence the inputs of this agent. In [13] it was shown that this information can be obtained during the creation of workflows by passing simple information between the peers in a workflow; as agents link up to form workflows, the composition of the network influencing the inputs of each agent can be recorded. By using such knowledge, agents can locally determine whether extension of a DPIF workflow with a new peer would introduce cycles; i.e. each agent can find out whether or not it is setting up an (in)direct dependency on itself, without any central authority.

While in this paper we cannot describe all interesting aspects of this approach, we illustrate the basic cycle detection principles from [13] with the help of an example. In figure 4(a), agent $A$ supplying $x_a$ forms a workflow with agent $B$ supplying $x_b$. Agent $A$ knows the output variable set $S_A = \{x_a, x_b\}$, and agent $B$ knows the output variable set $S_B = \{x_b\}$. In figure 4(b), an agent $C$, able to supply $x_c$, joins the workflow. Agent $C$ knows its current set of output variables; $S_C = \{x_c\}$. Before joining, agent $A$ verifies whether the proposed connection does not introduce cycles. This is the case if a simple intersection test yields

[1] A directed graph representing a workflow consists of nodes, each representing a process (i.e. a function), and directed links which capture direct dependencies between the processes (i.e. supplier-consumer relations).

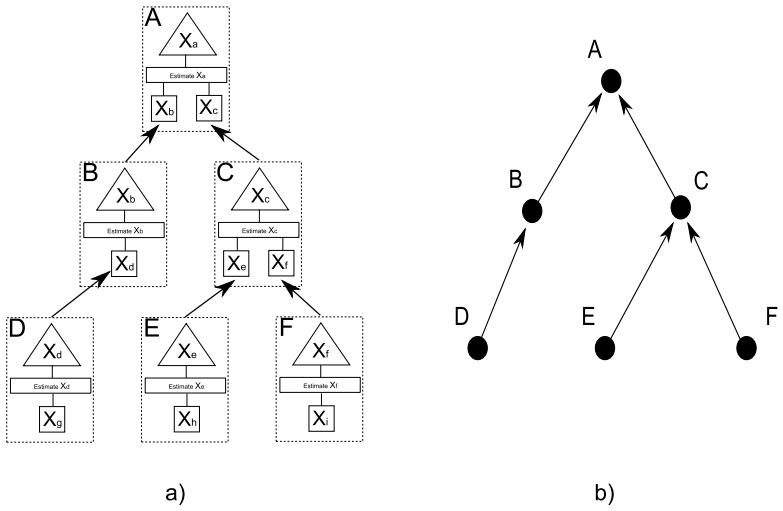



Figure 2: a) A self-organized system of agents. Each agent supplies information concerning a particular variable of interest in the domain. These outputs are based on other inferred or directly observed variables. b) A directed graph capturing the workflow between the agents from a).

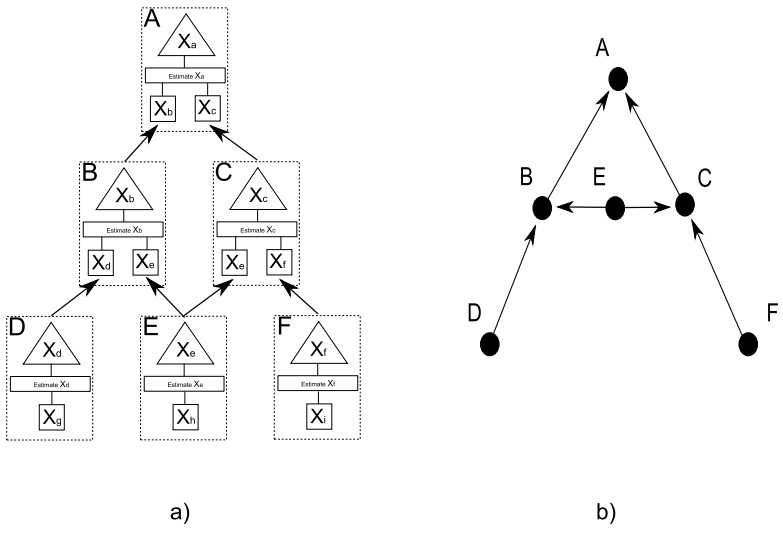



Figure 3: a) A self-organized system of agents forming a workflow corresponding to a multiply connected graph. b) A directed graph capturing the workflow between the agents from a).

$S_A \cap S_C = \emptyset$. This holds, so the connection is allowed and Agent C joins the workflow. As it does so, Agent $A$ updates its variable set to $S_A = \{x_a, x_b, x_c\}$; i.e. $A$ obtains all the relevant information about the services that influence its function by peer to peer sharing of local information.

However, when agent $C$ looks for suppliers, the only available agent supplying $x_a$ is $A$, the one to which $C$ is already connected. As $C$ conducts a verification step, in which the variable sets $S_C = \{x_c\}$ and $S_A = \{x_a, x_b, x_c\}$ are tested for empty set intersection, the intersection $S_A \cap S_C \neq \emptyset$, and so C knows that a cycle would be introduced if the service $x_A$ were supplied to it by $A$.

In fact, in [13] it was shown that cycles, as well as loops, can be detected in workflows in completely decentralized manner by collaboration of peers exchanging asynchronous messages. Peers check the intersections of dynamically assembled variable sets at different levels of the workflow, and as new agents join the workflow the new network layout needs to be reflected in all agents whose downstream network has changed by new connections. Thus, we can view the task of loop and cycle detection as a combination of (i) checks which *travel* upstream (i.e. toward the top agent) until the top agent of the network is reached, (ii) messages conveying the updated topology, and (iii) control messages which lock/unlock the agents for local checks.

In general, this approach allows for an intelligent handling of loops and cycles in workflows, where the choice on whether to allow a connection or not is dependent on the function performed by an agent that is responsible for expanding a workflow. There exist functions which require that all inputs are provided, in order to yield an output. In such cases, an agent modeling a function may decide to abandon a workflow when one or more of its inputs would lead to a cycle (or loop). On the other hand, there are also functions which yield output even when some inputs are left unknown, such as for example marginal conditional probabilities expressed with the help of Bayesian networks. In these cases, an agent modeling such a function may keep participating in the workflow, provided it can ensure that the inputs otherwise responsible for introducing cycles are not supplied to any other agent; i.e. the inputs are ignored in the evaluation of the function.

## 3.3 Negotiation

In the DPIF, communication links between local processes in agents are facilitated firstly using service discovery: whenever an agent supplying some service (we will call this service the *parent service*, and the agent implementing it the *parent*, or *manager agent*) in a workflow requires data relating to some other service (we will call this required service the *child service*, and the agent implementing it the *child*, or *contractor agent*), a communication link needs to be established between the parent agent and the child agent. However, there are two important aspects that affect whether and why links are established: i) we might have several agents in the system that provide the same service, i.e. that are able to realize the same task, and ii) we cannot always assume that a service providing agent will automatically agree to supply the service asked for by a requesting agent. For example, the providing agent might be overloaded, or it might even consider that establishing a link is inappropriate, given the current context.

In addition, on its own, service discovery can only offer links between agents based on a broad level of service matching, while for the system to solve a particular problem, a finer level of control is required to match services on whatever additional parameters may be of importance to particular links. For this we use negotiation. Rather than performing perfect matching at the service discovery level, negotiation allows us to filter potential links found through service discovery based on additional service parameters.

Negotiation in general consists of three elements [10]:

- protocols, i.e. sets of rules that describe the steps of negotiation processes; example protocols are Contract Net (CNET), monotonic concession protocol (MCP), Rubinstein's alternating offers and auctions [28] [24] [29].
- subject, i.e. the item being negotiated about. In service negotiation, the negotiation subject is the service with its parameters.
- strategies, i.e. the set of decisions that agents will make during the negotiation in order to reach a preferred agreement.

We have developed a conceptual framework for service negotiation that will be used in the DPIF [21]. The framework is generic and addresses negotiation protocols, negotiation subject and decision components (how agents make proposals and how they select the best proposals).

Establishing links is based on one-to-many negotiation; i.e. one agent (the manager) negotiates with multiple agents (possible contractors) about a service, with an arbitrary set of parameters (multi-issue subject) [19], [20]. The framework defines the common steps of negotiations, including starting negotiations, making proposals, deciding whether an agreement or a conflict has been reached and termination.

The negotiation subject consists of the service plus a subset of service parameters that are important decision factors during negotiation (they are taken into consideration when selecting the proper service providers). During negotiation, these parameters are considered negotiation issues. Thus, the negotiation designer defines the negotiable parameters of the service (negotiation issues) when configuring the service. When negotiating, agents need to know how to handle the issues and extra information about the issues should be added into the system. Therefore, issues have a set of properties. These properties include:

- *name*: a unique identifier of the issue in the subject.
- *data type*: the type of the value the issue is allowed to take.

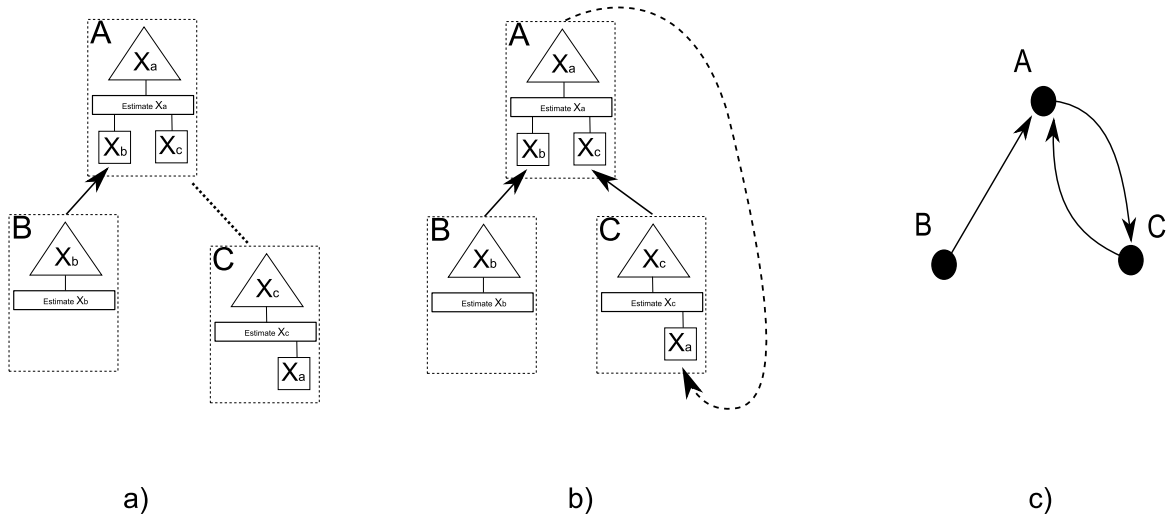



Figure 4: a) a partially formed collaboration, allowing the top agent to perform inference, leading to information about the state of variable $x_a$. b) If the potential connection represented by dashed directed link were allowed, a cycle would form in the workflow. c) A graph corresponding to a system of DPIF agent if the connection from b) were allowed.

– *value type*: specifies whether the value of the issue can be modified in proposals or the original value set by the manager should be kept.

In addition to the standard set of properties, agents (depending on their role) are allowed to assign some extra properties to the issues:

– *weight*: represents the relative importance of the issue in the subject, compared to the other issues.

– *reference value*: represents the ideal value of the issue from the point of view of the agent.

For the moment the issues are evaluated independently.

The default negotiation in DPIF is an implementation of CNET [24]. A manager starts a negotiation for a service by sending a call for proposals to all the contractors that are able to provide the service. Each contractor then tries to compute a proposal. Contractors evaluate possible proposals using a utility function that takes into consideration the effort necessary to provide the service under the conditions in the proposals. If contractors are successful in generating proposals, they send them to the manager, otherwise they withdraw from the negotiation. After receiving proposals, the manager selects the best proposals, the ones that give him the highest utility. The manager uses a weighted sum utility function to evaluate proposals.

Although the default negotiation protocol is CNET, the system allows for the use of arbitrary negotiation schemes, supporting domain specific one-to-many generic negotiation approaches where the protocol, the subject type and strategies are specified in the service configuration phase. There are two configuration levels:

– *negotiation type level*. At this level the negotiation designer defines the negotiation protocol, identifies the negotiation issues and sets default values for their properties.

– *negotiation instance level*. At this level a human expert can tune the issue properties for a certain situation.

We illustrate how negotiation takes place with such multi-issue subjects, by looking at the example described in figure 1. We emphasize the step when the CE2 decides to guide MTs to measure gas concentrations at a location *X*. In addition, CE2 decides that measurement devices *DEV_X* are the most appropriate for the measurements. Other devices can be used as well but with less precision. CE2 initiates a negotiation over the multi-issue subject (*Gas measurement, location, device*) with all MTs that are able to provide the service *Gas measurement*. MTs propose various deals for this negotiation: the locations they are currently placed at and the devices they have. Ideally, MTs would propose the best deal (*Gas measurement, X, DEV_X*). CE2 must decide what MTs to choose by taking into account various parameters: the distance between location *X* and locations where MTs are placed, the differences in precision between device *DEV_X* and devices MTs have.

The outcome of a successful negotiation between two agents is a contract, which results in the creation of a communication channel supporting peer-to-peer communication between the agents.

## 3.4 Collaborative information acquisition

A DPIF-based system of experts and automated reasoning processes typically requires acquisition of large amounts of very heterogeneous information obtained at different locations by using complex acquisition procedures.

The DPIF approach supports efficient gathering of information by distributing the search task throughout systems of collaborating agents, i.e. services. Namely, the search for the relevant information/data must be based on the domain knowledge in the system, which is used for the reasoning; only the information that is represented in implicit or explicit domain models can be processed. Since in the DPIF the domain knowledge is dispersed throughout a system y the experts or automated processes providing the reasoning services, the information gathering is carried out by the service providers as well.

Upon establishing contracts in a particular collaboration network, the service providers initiate a search for information based on their domain knowledge. Each service provider knows exactly which information is needed and under what conditions it should be obtained. The service providers either request inputs via the DPIF service discovery or they have access to direct observations (e.g. they observe sensors, databases, etc.). We say that the services with a direct access to the observations are *grounded* in the domain; they can observe real world phenomena directly. If the services are not *grounded*, they delegate the search of information and pass on only the information which *constrains* the search process carried out by service provider. Typically, an information requester knows *what* is needed and *where* and *when* it should be obtained but does not know how this information locating can be carried out. This information is passed on to the provider, that knows *how* the information can be obtained in the specified spatio-temporal context.

In this way a complex search for information is broken down into several simple search processes which guarantee that only the relevant information is inserted into the distributed reasoning system. By using the service discovery and negotiation, the judgment of what is needed and what can be provided is carried out by the information consumers and providers themselves, without the need of introducing a central authority. Note that a centralized approach to information search would require a centralized authority that would replicate some of the local knowledge and have a complete overview of the information acquisition processes and capabilities of all services. Such a central solution is not practical in the targeted domains, where new expertise/algorithms as well as new information sources are added frequently.

We illustrate the distributed approach to information acquisition with the help of the aforementioned example. Expert CE2 is requested to estimate the critical zones after a leak at a chemical plant is discovered. CE2 needs additional information to accomplish this task. Based on his local expertise, CE2 knows *what* information should be obtained and under what conditions this should happen. For example, CE2 needs the information about the type and quantity of toxic fumes escaping at location $X$ (assuming he was just informed that there was a leak at $X$). He also needs information about the weather in the larger incident area. This information should be obtained within a certain time interval and the service itself should conform to certain quality requirements, pertaining to such things as the experience of the information providers, minimum precision of the estimates, etc. The requested types of information as well as the additional constraints are used by the DPIF service discovery and negotiation mechanisms which establish contacts between the DPIF agents of the requester CE2 and suitable information providers, such as expert CE1. The negotiation ensures that the requested types of information pertain to the right location and time interval and that the experts with the right quality of service are involved. For example, a request from CE2 is processed in the DPIF, which finds agents of two experts that can provide service of type CE1. However, during the negotiation, it may turn out that only one expert can provide the estimate of the toxic fumes at location $X$ within the given time interval. The chosen CE1 can get to the location on time and supports the required quality of the service. CE1 gets to the incident location and obtains from the factory staff the information about the leaking chemical, the pressure and the leak location; i.e. CE1 knows what information should be obtained about the incident site and how this can be done. This information is used by CE1 for the estimation of the quantities of the escaping fumes. This estimate is routed to CE2 by using the peer to peer communication channel. After the first estimate of the critical areas is provided, CE2 requests additional concentration measurements at certain locations, in order to improve the initial estimate of the critical zone. CE2 specifies the types of measurements, the measurement locations as well as the maximum time interval in which the measurements must be provided. The requests are routed via the DPIF to the measurement teams (MT) who are able to go to the specified locations and carry out concentration measurements. The subsequent negotiation results in the creation of contracts between the CE2's agent and the agents representing the MTs which can then get to the specified locations in the given time and carry out the requested measurements.

Note that in this simplified example the search of information requires very different types of requests and acquisition methods. In other words, a lot of domain knowledge as well as procedural knowledge is required.

The requests to CE1 and MT reflect CE2's knowledge of the gas propagation processes, his current knowledge of the situation as well as his processing capabilities. The bids from MT and CE1, on the other hand, reflect their capabilities to provide the requested information. As a result of this approach, from a global perspective, the information acquisition process implements very complex patterns in non trivial spatio-temporal context.

# 4 Processing modules: architecture

Each local process (human or machine-based) is encapsulated by a software agent. In the DPIF, agents provide a uniform interface for the collaboration between local pro-

cesses in different agents. A key feature of DPIF agents is asynchronous, data-driven processing in complex workflows. This is achieved through a combination of weakly coupled processes inside individual agents, and service based coupling of processes between distinct agents.

Each agent consists of at least two processes implemented through asynchronous threads communicating via a local blackboard (see figure 5). The "Communication Engine" process is a thread that allows for message-based inter-agent communication, facilitating collaboration and making negotiation capabilities known to other agents. Through their Communication Engines, workflows between local processes in different agents can be established, by executing service discovery and negotiation (see section 3.3). The "Processing Engine" process, on the other hand, is a thread which encapsulates arbitrary automated or human based inference processes. The Processing Engine is responsible for the local processes that implement transformations between different types of information, based on the interpretation of complex cues. Moreover, each Processing Engine can keep track of one or more of these local processes simultaneously. The Communication Engine supplies the Processing Engine with inputs that are obtained through inter-agent messaging, by posting these on the agent's local blackboard for the Processing Engine to see. The Processing Engine then places the results of local inference processes on the local blackboard for the Communication Engine to pick up and relay via normal inter-agent messaging to interested agents.

The Communication and Processing engines must be able to execute simultaneously. Reasoning can be computationally expensive which requires a certain amount of time, but during this time an agent should be able to negotiate about possible collaboration with other agents, asynchronously collect their outputs for use in local processes and so on.

Both the Communication Engine and Processing Engine threads communicate through a limited set of messages via the local blackboard. New externally received inputs are transformed and placed by the Communication Engine on the internal blackboard, which triggers callback of adequate functions in the Processing Engine. The Processing Engine transforms these inputs into local output and places this output on the blackboard, which triggers callback at the Communication Engine for relaying this information via normal messaging to agents interested in this output. Through cascaded processing in the resulting system of collaborating agents, each piece of information influences the outcome of the collective reasoning processes, such as estimates/predictions or evaluations (i.e. reasoning result); with each contributing process, new observations and a particular expertise are incorporated into the global reasoning process.

Note that, irrespective of the type of the local processes, the Communication Engine and the Processing Engine in each agent use the same mechanisms for the creation and maintenance of correct workflows between local processes in different agents. The uniformity of configuration mechanisms can be used for a seamless integration of human experts into complex processing workflows. A DPIF agent representing an expert has a very simple Processing Engine which delivers the required inputs to a human expert via a suitable GUI. The expert's conclusions, i.e. his/her service, are also formulated with the help of this GUI and routed to interested agents via the Communication Engine. Thus, agents used by human experts merely provide automated routing of information between experts, and take care of the automated creation of collaboration connections.

# 5 Dynamic service ontologies

In order to be able to automatically compose heterogeneous services provided by different developers or experts, the definitions of service interfaces have to be standardized, which is achieved with the help of explicit service ontologies. Moreover, the locality of the domain knowledge in the DPIF approach can be exploited for efficient creation of rigorous service descriptions.

Services declared in the DPIF are typically provided by many stakeholders from different organizations whose capabilities evolve with time. Large systems of service descriptions have to be maintained and it is very difficult to specify a complete set of services prior to the operation. In other words, traditional approaches based on rigorous centralized ontologies, such as for example [4, 22], which capture service descriptions and relationships between information provided by different types of services are not practical; we simply do not know which relevant services will be available in the future and maintenance of large ontologies is likely to be very expensive or even intractable.

Fortunately, the locality of domain knowledge in the DPIF approach supports efficient creation of service ontologies. Because self organization and processing are based on domain knowledge encoded in local functions, we can avoid traditional approaches to constructing centralized ontologies, which describe domains in terms of complex relations between the concepts corresponding to the processing services. In the targeted domains, adequate maintenance of such ontologies is likely to be an intractable challenge. Instead, the services and relations between them are described by using two types of light weight ontologies:

- *The global service ontology* merely captures service descriptions, the semantics and syntax of messages used for (i) service invocation and (ii) dissemination of service results. This ontology is used for the alignment of the semantics and syntax of service descriptions at design time.

- *Local task ontologies* coarsely describe relations between different types of services supplying different types of information. In principle, they describe which types of services provide inputs to the function used by a specific service. These relations reflect the

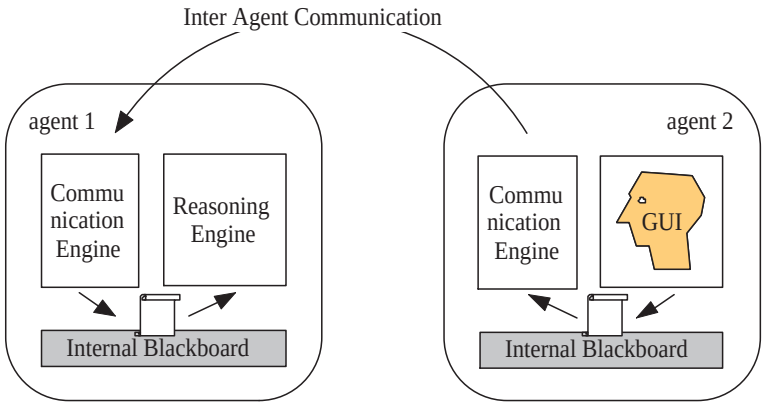



Figure 5: Interaction between agents providing heterogeneous processing services. Both agents use identical communication engine. However, agent 1 encapsulates automated processing while agent 2 introduces human-based processing.

local knowledge of each processing module. Moreover, the local ontology supports runtime creation of workflows based on service discovery.

The global ontology is a central element of the service description procedure. In order to make sure that all agents speak the same language, the global ontology captures three types of elements, namely (i) a verbal description of the service to be provided, (ii) conditions under which this service can be invoked, and (iii) a collection of representational elements resulting from the information gathered by this service. While the vocabulary with which these descriptions can be specified is rigidly formalized, it is rich enough to allow the description of arbitrarily complex services. The global ontology is used by a matching process in which service suppliers are provided with a list of existing service descriptions, based on keywords and free text. The descriptions retrieved from the global ontology are displayed in a form that facilitates inspection of the relevant subset of existing services. If an existing service description corresponds to the new service, it is adopted. Otherwise a service definition editor allows the experts to provide a new service description, which is then added to the global ontology. By making experts responsible for deciding whether they perform a role similar to another domain participant or a genuinely new role, we overcome the problem of an a priori defined ontology that is likely to be unable to account for all aspects of the domain and expert capabilities.

The *local task ontologies*, on the other hand, are created with the help of a task definition tool which supports specification of the required inputs (provided by other services) for each provided service. In this way different services are related locally, based on the local domain knowledge. The task ontologies are stored with agents of participating experts. These relations captured by local task ontologies are central to the service discovery, which is typically initiated from within the local services. Consequently, if each expert is made responsible for the description of relations between the provided and the needed services, systems using complex relations between services can be built in a collaborative way, without any centralized configuration/administration authority.

In other words, the introduced division into a global service ontology and local task ontologies dispersed throughout the system of agents allows collaborative definition of services by experts.

Construction of such ontologies is facilitated by a collection of tools that (i) help the user discover the services in the global ontology by using keywords and written language, (ii) provide an interface facilitating inspection of the human readable descriptions and (iii) editors for defining local task ontologies. By using these tools, the experts define elements of the global service ontology and the local task ontologies without using any formal language. At the same time, the tools automatically translate expert inputs to rigorous ontologies captured in the OWL format. In other words, by deploying the two types of ontologies in combination with simple construction procedures, rigorous, machine understandable service descriptions can be created without any formal knowledge of the underlying ontology techniques.

Similarly to the DPIF approach, the OpenKnowledge framework [23] avoids creation of centralized heavy weight ontologies describing all aspects of the domain. However, while the DPIF requires a mere specification of the provided and supplied services, the OpenKnowledge framework also requires specification of *interaction models* shared by the collaborating peers. Such interaction models define workflows for each processing task a priory; the OpenKnowledge approach assumes that collaborating peers understand interaction protocols and the processing sequences of collaborating peers. This can introduce additional complexity to the system configuration in which services and processes are specified. Since the DPIF is targeting Professional Bureaucracy systems [26], it is assumed that experts do not have to share knowledge about their local processes.

# 6 Conclusions and future work

The DPIF supports uniform encapsulation and combination of heterogeneous processing capabilities which are required for collaborative reasoning about complex domains. The processing capabilities can be provided by human experts or automated reasoning processes. In the DPIF context, human expertise and automated processes are abstracted to functions with well defined outputs and inputs; each function provides a particular reasoning service given certain inputs.

The DPIF provides *function wrappers*, software agents which standardize function interfacing. The interfaces are based on standardized service descriptions as well as uniform self-configuration, negotiation and logical routing protocols. With the help of the DPIF encapsulation methods very heterogeneous services can be made composable and negotiable.

The DPIF agents support automatic formation of workflows in which heterogeneous functions correspond to suppliers and consumers; outputs of some functions are inputs to other functions and so on. In other words, a workflow corresponds to a set of nested functions that captures dependencies between very heterogeneous variables. Creation of workflows and routing of information is based on the relations between different types of information. These relations are captured by *local functions* wrapped by different modules. The DPIF approach assumes that each expert or an automated process can declare the inputs and outputs of the contributed local functions, which is sufficient for automated creation of globally meaningful workflows by using service discovery. Thus, in contrast to traditional approaches to processing in workflows, *neither centralized configuration of workflows nor centralized knowledge of the combination or routing rules are needed*. The resulting systems support processing based on rich domain knowledge while, at the same time, the collaboration between heterogeneous services requires minimal ontological commitments.

Decentralized creation of emergent processing workflows, based on local domain knowledge and negotiation, is useful in dynamic domains, such as crisis management, where it is difficult to maintain a centralized overview of the resources. However, if applications require optimization of workflows, centralized approaches to workflow construction might be necessary. The DPIF approach supports also centralized approaches in several ways. Namely, the DPIF facilitates creation of processing modules whose services can easily be composed by centralized algorithms; the local task ontologies provide the information on the compatibility of services, i.e. possible service combinations in workflows.

In principle, arbitrary automated reasoning techniques can be integrated into the DPIF. However, globally coherent reasoning in such workflows can be achieved only by using rigorous approaches to designing local models and combining partial processing results. Globally coherent and theoretically sound collaborative reasoning is in general very challenging and it has not been discussed in this paper due to the limited space. An example of a theoretically sound collaborative inference system based on the DPIF is the Distributed Perception Networks (DPN), a modular approach to Bayesian inference [16]. The DPN is a fully automated DPIF variant that supports exact decentralized inference through sharing of partial inference results obtained by running inference processes on local Bayesian networks [18] in different collaborating DPN agents. If the local Bayesian networks are designed according to the rules introduced in [16], it can be shown that a collaboratively computed posterior distribution for any variable in the distributed system correctly captures all evidence. The DPN framework has been used for the implementation of robust distributed gas detection and leak localization systems based on Hidden Markov Models [17].

In general, however, it can be difficult to prove that a collaborative processing approach is sound (i.e. globally coherent), especially if algorithms and the used models are not based on rigorous theory, which is especially the case with human based reasoning. Recent research on structuring of human collaborative processing in DPIF-based systems indicated that some processing rigor can be introduced if the experts implement causal reasoning [6]. In such settings the reasoning could be structured with the help of qualitative models, such as causal graphs which allow exploitation of d-separation and concepts from Hidden Markov Models (HMM) [1], such as observation and dynamic process models defined over discrete time-slices. Such structuring has several advantages. Firstly, we can prevent processing that involves the so called data-incest and recycling of information in infinite, self perpetuating reasoning cycles. Secondly, we can introduce efficient control of discrete reasoning phases throughout time slices, which supports sound temporal reasoning.

In addition, several challenges remain with decentralized creation and control of workflows. It turns out that the complexity of workflow control depends on the problem itself and the used processing paradigm. In cases in which the reasoning can be structured as HMMs, valid workflows can easily be created and time-slices can be controlled via emission of simple reset messages within the system. In such settings no workflow deadlocks can occur and information can unambiguously be associated with subsequent time-slices.

A basic version of the DPIF as well as a prototype of the service configuration tool have been implemented and are currently being enhanced in the context of the FP7 DIADEM project. In this project we are investigating incorporation of advanced negotiation techniques as well as integration of Multi Criteria Decision Analysis and Scenario Based Reasoning methods facilitating human-based processing in workflows.

# 7 Acknowledgments

The presented work is partially funded by the European Union under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D, ref. no: 224318.

# References

[1] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[2] Shawn Bowers and Bertram Ludäscher. An ontology-driven framework for data transformation in scientific workflows. In *DILS*, pages 1–16, 2004.

[3] Jorge Cardoso and Amit P. Sheth. Semantic e-workflow composition. *J. Intell. Inf. Syst.*, 21(3):191–225, 2003.

[4] David Chiu and Gagan Agrawal. Enabling ad hoc queries over low-level scientific data sets. In *SSDBM*, pages 218–236, 2009.

[5] David Chiu, Sagar Deshpande, Gagan Agrawal, and Rongxing Li. A dynamic approach toward qos-aware service workflow composition. In *ICWS*, pages 655–662, 2009.

[6] Tina Comes, Claudine Conrado, Michiel Kamermans, Gregor Pavlin, Niek Wijngaards, and Michale Hiete. An intelligent decision support system for decision making under uncertainty in distributed reasoning frameworks. In *7th International Conference on Information Systems for Crisis Response and Management*, Seattle, USA, May 2-5 2010.

[7] Subrata Das. *High-Level Data Fusion*. Artech House, Inc., Norwood, MA, USA, 2008.

[8] Ewa Deelman, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Karan Vahi, Kent Blackburn, Albert Lazzarini, Adam Arbree, Richard Cavanaugh, and Scott Koranda. Mapping abstract complex workflows onto grid environments. *J. Grid Comput.*, 1(1):25–39, 2003.

[9] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

[10] N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra, and M. Wooldridge. Automated negotiation: Prospects, methods and challenges. *International Journal of Group Decision and Negotiation*, 10(2):199–215, 2001.

[11] Nicholas R. Jennings, Katia P. Sycara, and Michael P. Georgeff. Editorial. *Autonomous Agents and Multi-Agent Systems*, 1(1):5, 1998.

[12] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001.

[13] Michiel Kamermans. Distributed perception networks: Effecting consistent agent organisation and optimising communication volume in a distributed multi-agent network setting. *Masters Thesis, Informatics Institute, University of Amsterdam*, 2008.

[14] Philip Maechling, Hans Chalupsky, Maureen Dougherty, Ewa Deelman, Yolanda Gil, Sridhar Gullapalli, Vipin Gupta, Carl Kesselman, Jihie Kim, Gaurang Mehta, Brian Mendenhall, Thomas A. Russ, Gurmeet Singh, Marc Spraragen, Garrick Staples, and Karan Vahi. Simplifying construction of complex workflows for non-expert users of the southern california earthquake center community modeling environment. *SIGMOD Record*, 34(3):24–30, 2005.

[15] Mike P. Papazoglou, Paolo Traverso, Schahram Dustdar, and Frank Leymann. Service-oriented computing: State of the art and research challenges. *IEEE Computer*, 40(11):38–45, 2007.

[16] G. Pavlin, P. de Oude, M. Maris, J. Nunnink, and T. Hood. A multi agent systems approach to distributed Bayesian information fusion. *International Journal on Information Fusion*, 2008. To appear.

[17] Gregor Pavlin, Frans C. Groen, Patrick de Oude, and Michiel Kamermans. *A Distributed Approach to Gas Detection and Source Localization Using Heterogeneous Information*. Springer Verlag, 2010. ISBN 978-3-642-11687-2, in print.

[18] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, 1988.

[19] Mihnea Scafeş and Costin Bădică. Preliminary design of an agent-based system for human collaboration in chemical incidents response. In Ulrich Ultes-Nitsche, Daniel Moldt, and Juan C. Augusto, editors, *Proc.of MSVVEIS 2009 – 7th Int.Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems*. INSTICC Press, 2009.

[20] Mihnea Scafeş and Costin Bădică. Service negotiation mechanisms in collaborative processes for disaster management. In *Proceedings of the 4th South East European Doctoral Student Conference (DSC 2009), Research Track 2: Information and Communication Technologies*, Thessaloniki, Greece, 2009. SEERC.

[21] Mihnea Scafeş and Costin Bădică. Conceptual framework for design of service negotiation in disaster management applications. In *To appear in Proc. of 2nd International Workshop on Agent Technology for Disaster Management (ATDMâĂŹ09)*, Studies in Computational Intelligence. Springer Verlag, 2010.

[22] Quan Z. Sheng, Boualem Benatallah, Marlon Dumas, and Eileen Oi-Yan Mak. Self-serv: A platform for rapid composition of web services in a peer-to-peer environment. In *VLDB*, pages 1051–1054, 2002.

[23] Ronny Siebes, David Dupplaw, Spyros Kotoulas, Adrian Perreau de Pinninck Bas, Frank van Harmelen, and David Robertson. The openknowledge system: an interaction-centered approach to knowledge sharing. In *Proceedings of the 15th Intl. Conference on Cooperative information systems (CoopIS) in OTM 2007*, 2007.

[24] R. G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Trans. Comput.*, 29(12):1104–1113, 1980.

[25] Paolo Traverso and Marco Pistore. Automated composition of semantic web services into executable processes. In *International Semantic Web Conference*, pages 380–394, 2004.

[26] Chris J. van Aart, Bob Wielinga, and Guus Schreiber. Organizational building blocks for design of distributed intelligent system. *International Journal of Human-Computer Studies*, 61(5):567 – 599, 2004.

[27] W.M.P. van der Aalst, A.H.M. ter Hofstede ans B. Kiepuszewski, and A.P. Barros. Workflow patterns. *Distributed and Parallel Databases*, pages 5–51, 2003.

[28] José M. Vidal. *Fundamentals of Multiagent Systems: Using NetLogo Models*. 2006. http://www.multiagent.com.

[29] Michael Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2002.

# A Reflection on Some Critical Aspects of Online Reading Comprehension

Antonella Chifari, Giuseppe Chiazzese, Luciano Seta,Gianluca Merlo, Simona Ottaviano and Mario Allegra
Institute for Educational Technologies - Italian National Research Council
Via Ugo La Malfa, 153, Palermo - Italy
E-mail: gym2learn@itd.cnr.it

**Keywords:** online reading comprehension, instructional strategies, surfing and comprehension skills

**Received:** September 7, 2009

*This paper reflects on some important aspects related to online reading comprehension. In particular, it explains the Interactive REading Comprehension (IREC) model that explores the different dimensions and interactions involved in an online reading comprehension process. The components of the model and their impact on the two principal processes characterizing any reading activity on the web, surfing and comprehension, are described. The final section of the paper focuses on some critical design issues related to the development of a web based tool to support online reading comprehension in relation to the model.*

*Povzetek: Predstavljen je model IREC, povezan z razumevanjem sprotnega branja.*

## 1 The scenario

Today the reading scenario of an adolescent has changed. While on the one hand the book exists as a traditional vehicle for the dissemination and comprehension of knowledge, on the other hand, the web represents a new type of reading space.

Knowledge construction on the web requires the ability to flexibly integrate traditional reading comprehension skills with new strategic knowledge applications elicited by the new reading domain for processing, comprehending and sharing information. More precisely, the web has become an important resource that extends the traditional reading comprehension scenario into an open hypermedia and multimedia knowledge space where a set of online comprehension strategies are employed to effectively locate, comprehend, and use the informational contents. When students are engaged in Internet learning and communication activities, reading comprehension is affected by the presentation of the contents to read: mail, blogs, social networks, multimedia and hypermedia contents introduce a fundamental change in the architecture of acts of reading. In fact, reading comprehension becomes a more complex, ongoing, self-regulated, decision process which involves choosing from different possible links, possible texts, possible purposes and among different ways of interacting with information [1]. This situation highlights a rapid change in the nature of reading so that the online domain requires a different reading literacy from traditional ones and a change of perspective in the dynamics of reading comprehension. Readers influenced by the information and communication contexts of the web adopt new ways of reading, locating information, employing a more complex dimension of inferential reasoning strategies to construct meaning. In fact, Leu [2] stated that new comprehension skills, strategies, and dispositions may be required to generate questions, and to locate, evaluate, synthesize, and communicate information on the web. Thus, reading in Internet contexts requires the ability not only to construct meaning from a text, but also to construct meaning through flexible and purposeful choices of relevant hyperlinks, icons, and interactive diagrams [3].

However, faced with this situation, the International PISA assessments on reading comprehension skills of European adolescents reveal a worrying image of “poor” readers lacking in basic cognitive strategies such as locating information or creating a mental overview of the text, connecting the meaning of one sentence to the meaning of another, using previous knowledge to try to clarify and connect meanings of words and phrases. Besides, readers find difficulty in comparing, contrasting or categorising information, inferring which information in the text is relevant to their task, critically evaluating or hypothesising and drawing on specialised knowledge [4]. As result of this data there is a clear need to study more carefully technological and methodological aspects of online reading comprehension processes.

In the next sections the Interactive REading Comprehension (IREC) model and a definition of online reading comprehension are introduced. Then, the different dimensions and relationships involved in the model and its applications to support online reading comprehension are described.

## 2 The IREC model

The need to study the complex relationship between web tools and reading processes is particularly urgent in the current contexts in which new tools, often on line, are being developed, and the hypertext is becoming the main structure of many learning materials in use in the classroom, in substitution or in addition to the traditional textbook. The enhancement of the reading proficiency of students by means of specific web-based learning tools and the development of a new literacy related to the hypertextual structure of web contents are two different goals, which are often not clearly distinguished. The IREC model aims to deal with these two problems in a unitary framework, in which the text structure and the tool design are taken into account jointly and evaluated from a more general point of view. More in depth the model is inspired by theories about the design effectiveness of learning tools or devices aimed to support users' learning processes and social interactions [5,6,7], by research into instructional strategies [8,9,10] and also by novel studies of reading comprehension processes on the Internet [1,2,11,12].

It is possible to draw together the different theoretical approaches to produce the Interactive REading Comprehension (IREC) model (Figure 1) which describes a number of different situations related to the learning design in a technologically mediated environment. To evaluate interactions between the learning activities and the technology in use, it is necessary to take into account four interrelated components:

- Pedagogical component;
- Technological component;
- Content component;
- User component.

The weight of the single components and the reciprocal relations between them establishes the idiosyncratic nature of each different approach. More precisely, the model has a flexible structure depending on the type of domain of use. In fact, by selecting one of the four components it is possible to define a specific domain in which the model can be applied.

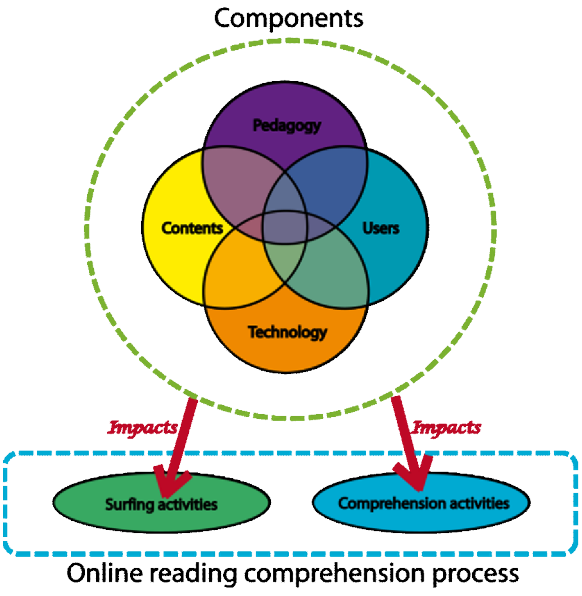



Figure 1. The IREC model.

Since this study aims to explore the design issues related to development of web based tools to support online reading comprehension, the "technological" component can be considered as fixed. In this particular representation of the model, shown in figure 2, specific assumptions are made for each component: the pedagogical component is represented by the instructional strategies supported by the web tools; the user component describes the adolescent reader's characteristics in terms of prior strategic knowledge and prior contents knowledge; the content component refers to the structure of learning materials, hypertexts and multimedia; finally, the technological component, which in this case has been extrapolated, consists of the design characteristics of a web tool and it is represented by an oval inside a dotted line including the other three components.

The IREC model stresses the relationships between these three components and their impact on the two principal processes underlying the online reading comprehension process, namely the surfing and comprehension activities.

The model is based on the most recent theories according to which skilled readers are able to balance both the demands for comprehending and for orienting themselves in hypertexts [13]. This concept must be borne in mind while providing instruction, planning the contents to study and evaluating user characteristics.

The next section focuses on the web tool features and how the features can be suitably developed according to the interaction with the characteristics of the three components.

### 2.1 Instructional strategies

This component identifies the relationship between a chosen instructional model (peer tutoring, collaborative learning, reciprocal teaching, etc.) and the technological choices/functionalities of the tool which are needed to enhance the online reading comprehension process. So, it is important to focus the design efforts on the key instructional principles which make offline and online

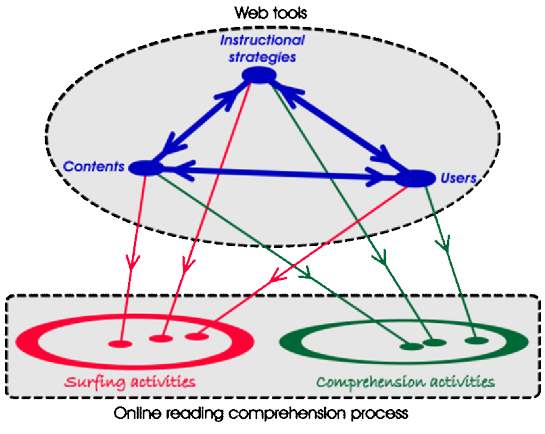



Figure 2. The IREC model for designing web tools

reading significantly different. The Texas Education Agency [10] states that instruction is effective if it is based on modelling processes and is well organized, explicit, intensive and long lasting, and if students are made aware of text organization and are motivated to read widely. In other words, the informed educational paths are much more powerful than blind ones, especially when the purpose is to stimulate metacognitive learning.

### 2.2 Users

This component guides the construction of users' profiles in terms of the proficiency level of online reading comprehension. The profile takes into consideration different aspects, such as prior strategic knowledge, prior contents knowledge, motivation, social and communication skills, planning skills, etc. Proficient online readers are able "to manage their strategic action as a part of a complex metacognitive domain". An independent reader is like a mental manager who plans his online reading strategy with awareness and implements strategic activities (e.g. asking and responding to questions, constructing meaning according to links chosen during surfing, critically evaluating the credibility of a source) [14]. The knowledge of an initial proficiency level assists in the construction of a more precise user profile and the provision of more effective metacognitive training activities.

### 2.3 Contents

This component regards two principal aspects: the first aspect is technological and related to the structure of information in terms of its level of multimediality and hypermediality. In fact, it is a central design issue and takes into account some well-known aspects such as modularity, linearity, multimediality, granularity, interactivity and the different characteristics of texts (narrative, informational, scientific, etc.). The second aspect is educational, regarding important guidelines for developing learning materials consistent with the students' level of proficiency and with their personal perception of meaningful information so as to provide a rich context for learning. According to Baker [9], if the material is essentially meaningless to the student, he will have a great deal of difficulty in retaining it. On the contrary, if the student perceives the logical structure of the material, he will be better able to learn from it.

### 2.4 Online reading comprehension process

The online reading comprehension process applied by the user is a complex set of strategies employed to construct meaning. More precisely, two levels of strategies are employed: a first level of cognitive strategies for orienting oneself in hypertext reading and for comprehending textual information, and a second level of metacognitive strategies employed for monitoring and checking the efficacy of reading comprehension and surfing processes.

Traditionally, reading comprehension is a complex active process of constructing meaning that is interactive, strategic and adaptable [10] It is interactive because it involves not just the reader but also interaction with the text in which reading takes place [15]; it is strategic since readers have goals that guide their reading and they use different cognitive strategies and skills as they construct meaning [9, 16] it is adaptable because readers change the strategies they use as they read different kinds of text or as they read for different purposes [17]. At the same time, we define the surfing process as a complex, active process of constructing paths and finding directions. It is interactive because it involves hypertext links through which browsing take place; it is strategic because surfers have information to find that orients their choice of links and they use different cognitive and metacognitive strategies and skills while they are following a path; it is adaptable since surfers change their strategies according to the design characteristics of different content structures. The combination of these two processes employed by students during the comprehension of online contents gives a new meaning to the acts of reading. This is because hypertext readers need to became competent both in constructing meaning and also in the employment of strategies for managing the different aspects of the surfing process [1].

## 3 The interaction among the studied components

The main intention of the IREC model is to stress the relationships between the components, and their impact on the two principal processes characterizing any study activity on the web: surfing and comprehension.

In particular, regarding the three components considered above, pedagogy, users and contents, while it is sufficiently clear that the choice of an instructional model must be the result of careful evaluation of the characteristics of both the users and contents in use, it is less evident that the assessment of users' proficiency should be made according to the instructional model applied and the structure and organization of didactic materials. In the same way, the content design is influenced by user characteristics: the contents can be developed to satisfy different user profile and different reading proficiency levels. So, different contents could be developed for supporting the learning of specific strategies such as locating information, creating a mental overview of the text; connecting the meaning of one sentence to the meaning of another. Moreover the instructional strategies can affect the level of interaction of the content in terms of personal, reciprocal and collaborative construction of meaning on the web. Finally, the users' characteristics, in terms of prior knowledge, motivation, cognition and metacognition strategies, mental managing, proficiency and mastery, have to be taken into account in order to establish a suitable level for the teaching topic and the structure of the contents, and to plan the learning activities.

But the model also wants to emphasize that these characteristics are not static, they evolve over time and so some specific tools are needed to keep pace with this evolution. Generally, this aspect is discussed in relation

to the rapid changes in information and communication technologies. The development of new techno-based systems appears to be a major stress factor in the educational environment, with teachers and students running a non-stop race to acquire the latest novelties. Little attention is paid to the parallel evolution of pedagogical methodologies, users' behavioural habits, multimedia and hypermedia languages. One of the reasons why the diffusion of some technology based educational tools has also been possible is due to the increasing familiarity of learners and teachers with new modalities of interaction as well as the development of pedagogical approaches based on, for example, simulation and visual knowledge management.

But the evaluation of the relationships between the components is not sufficient. It is also necessary to recognize how these components impact on surfing and on comprehension activities. For example, any decision aimed at improving the surfing process might have a negative effect on the comprehension activity, and vice versa. So, web tools designed to facilitate the storage of web pages, a typical surfing feature, might limit the students' ability to identify the main concepts of a text, an important comprehension activity; likewise, tools designed to organize the contents graphically, a useful reading comprehension activity, might hamper orientation on the web, a surfing aid for monitoring surfing behaviour.

All these relationships have different impacts on the design of web tools to support on line reading comprehension processes, so the design has to be a multi-level activity, involving different professional figures, such as teachers, pedagogues, psychologists, and technicians. But many solutions can arise from theoretically based observations of on line learning practices. In this respect the IREC model may prove to be a useful tool to distinguish the most relevant variables and dimensions involved in a web-based learning experience.

# 4 Conclusions and discussion

The rapidity of technological change and the increasingly frequent use of Internet for educational purposes have increased the learning demands for comprehension and for thoughtful navigation.

The additional value of the presented model can be ascribed to a systemic design perspective in which the characteristics of each component interact dynamically. Any variation of the intrinsic value of a component affects not only the characteristics of other components, but also the design domain. In this context the level of proficiency in reading comprehension affects not only the selection and construction of specific contents and instructional strategies but also design choices to enhance the empowerment of each component.

Focusing attention on the technological characteristics that a web tool requires for supporting online reading comprehension more effectively has the advantage both of stimulating theoretical research in this field and inviting a reflection on design and implementation issues, so that the technological solution represents an effective support, enabling students to become proficient readers during on line surfing.

From a theoretical point of view, it is necessary to investigate the processes that regulate online reading behaviour and in particular cognitive and metacognitive strategies, social competences and the effects of content structure on reading comprehension. It is also essential to find new indicators to measure levels of reading proficiency more accurately. Equally important is a reflection on comprehension instruction to promote students' cognitive scaffolding.

From a technological point of view, the IREC model suggests a reflection focusing on the following design choices: setting out a clear purpose for the intended tool; identifying a target and defining user profiles; identifying an instructional comprehension model and evaluating how it could be applied in a web-based environment; balancing surfing and comprehension features according to the established goal; including motivational features/activities to promote greater user participation [18].

Consequently, some of the following features could be implemented into a web tool: aids for monitoring all online comprehension behaviour such as reflection and annotation tools, cognitive and metacognitive prompts; aids to improve the research for information such as choosing keywords, identifying the best query results, evaluating web credibility; aids for organizing contents graphically such as conceptual maps and flow charts; aids for managing web page storage such as history, bookmarks/social bookmarks; opportunities for students to self-assess their knowledge; aids to promote a shared understanding of the goals for metacognitive activities and so on.

In conclusion, the IREC model could provide a starting point for further research discussion about the nature of online reading comprehension and the development of new online reading comprehension tools.

# References

[1] Afflerbach, P. and Cho, B. (2009). Identifying and Describing Constructively Responsive Comprehension Strategies in New Traditional Forms of Reading. *Handbook of Research on Reading Comprehension*, Routledge, New York and London, pp. 69-90.

[2] Leu, D. J., Coiro, J., Castek, J., Hartman D., K., Henry, L., A. and Reinking, D. (2008). Research on Instruction and Assessment in the New Literacies of Online Reading Comprehension. In Block, C. C. and Parris, S. R. (Eds), *Comprehension Instruction: Research-based Best Practices*, The Guildford Press, New York, pp. 321-341.

[3] Spiro, R.J. (2004) Principled pluralism for adaptive flexibility in teaching and learning. In R.B. Ruddell and N. Unrau (Eds.), *Theoretical Models and Processes of Reading.* Newak, DE: International Reading Association, pp. 654-659.

[4] OECD (2007) *PISA 2006 Science Competencies for Tomorrow's World Volume 1 – Analysis* OECD from http://www.pisa.oecd.org/dataoecd/30/17/39703267.pdf

[5] Jonassen D.H. (1994). Thinking technology. Toward a Constructivist Design Model. *Educational Technology*, pp.34-37.

[6] Bannert, M., Mildebrand, M., Mengelkamp C. (2009). Effects of a metacognitive support device in learning environments. *Computers in Human Behavior*, Elsevier, The Netherland, pp. 829-835

[7] Lin, X. D. (2001). Designing metacognitive activities. *Educational Technology Research & Development*, Springer, 49(2), pp. 23–40.

[8] Mishra, P. and Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A new framework for teacher knowledge. *Teachers College Record* 108 (6), pp. 1017-1054.

[9] Baker, L. And Brown, A. L. (1980). Metacognitive Skills and Reading. *Technical Report No. 188. Illinois Univ., Urbana, Center for the Study of Reading*, Bolt, Beranek and Newman, Inc., Cambridge, MA.

[10] Texas Education Agency (2002). Comprehension Instruction Texas Education Agency 1701 North Congress Avenue Austin, Texas.

[11] Coiro, J. and Dobler E. (2007). Exploring the online reading comprehension strategies used by sixth-grdade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly*, International Reading Association, pp. 214-257.

[12] Chiazzese G., Chifari A, Merlo G., Ottaviano S., and Seta L. Metacognition for enhancing online learning. *Technology Enhanced Learning: Best Practices*, IGI Global, pp. 135-153.

[13] Leu, D., Zawilinski, L., Castek, J., Banerjee, M., Housand, B., and Liu, Y. (2008). What is new about the new literacies of online reading comprehension? L.S. Rush, A.J. Eakle, and A. Berger (eds), *Secondary School Literacy: What Research Reveals for Classroom Practices*, NCTE/NCRLL, Urbana, IL (USA), pp. 37-68.

[14] Bimmel, P. and Van Schooten, E. (2004). The relationship between strategic reading activities and reading comprehension. *Educational Studies in Language and Literature*, 4, pp. 85-102.

[15] Heilman, A., Blair, T. R. and Rupley, W. R. (1998). *Principles and Practice of Teaching Reading.* Merril/Prentice-Hall, Upper Saddle River, NF.

[16] Wasik, B. and Turner, J. C. (1991). The development of strategic readers. In *Handbook of Reading Research, Vol. 2*, Longman, New York, pp. 609-640.

[17] Dole, F., Duffy, G., Roehler, L. and Pearson, P. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research*, 61(2), pp. 239-364. doi: DOI: 10.3102/00346543061002239.

[18] Giacoma, G. and Casali, D. (2008). Design motivazionale. Usabilità Sociale e Group Centered Design. Retrieved June, 15, 2009, from http://ibridazioni.com/2008/11/01/design-motivazionale-usabilita-sociale-e-group-centered-design/

# Enhancing DDoS Flood Attack Detection via Intelligent Fuzzy Logic

Zhengmin Xia, Songnian Lu and Jianhua Li
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
E-mail: miaomiaoxzm@sjtu.edu.cn, snlu@sjtu.edu.cn, lijh888@sjtu.edu.cn

Junhua Tang
School of Information Security Engineering, Key Lab of Information Security Integrated Management Research, Shanghai Jiao Tong University, Shanghai 200240, China
E-mail: junhuatang@sjtu.edu.cn

**Keywords:** network security, statistical detection, self-similarity, fuzzy logic

**Received:** February 1, 2010

*Distributed denial-of-service (DDoS) flood attack remains great threats to the Internet. This kind of attack consumes a large amount of network bandwidth or occupies network equipment resources by flooding them with packets from the machines distributed all over the world. To ensure the network usability and reliability, real-time and accurate detection of these attacks is critical. To date, various approaches have been proposed to detect these attacks, but with limited success when they are used in the real world. This paper presents a method that can real-time identify the occurrence of the DDoS flood attack and determine its intensity using the fuzzy logic. The proposed process consists of two stages: (i) statistical analysis of the network traffic time series using discrete wavelet transform (DWT) and Schwarz information criterion (SIC) to find out the change point of Hurst parameter resulting from DDoS flood attack, and then (ii) adaptively decide the intensity of the DDoS flood attack by using the intelligent fuzzy logic technology to analyze the Hurst parameter and its changing rate. The test results by NS2-based simulation with various network traffic characteristics and attacks intensities demonstrate that the proposed method can detect the DDoS flood attack timely, effectively and intelligently.*

*Povzetek: Opisan je postopek za prepoznavo spletnega napada DDoS s pomočjo mehke logike.*

## 1 Introduction

Distributed denial-of-service (DDoS) attack has been one of the most frequently occurring attacks that badly threaten the stability of the Internet. According to CERT Coordination Center (CERT/CC)[1], there are mainly three categories of DDoS attacks: flood attack, protocol attack and logical attack. This paper mainly focuses on flood attack. In the DDoS flood attack, an intruder bombs attack packets upon a site (victim) with a huge amount of traffic so as to actually jam its entrance and block access by legitimate users or significantly degrade its performance[2]. Therefore, a real-time and accurate detection of these attacks is critical to the Internet community.

Usually, the attack detection methods are classified into two categories. One is misuse detection and the other is anomaly detection. Misuse detection is based on a library of known signatures to match against network traffic. Hence, unknown signatures from new variants of an attack mean 100% miss. Anomaly detection does not suffer from this problem. Considering that DDoS flood attack is a process changing dynamically and frequently, anomaly-based detectors play a key role in detecting this kind of attack. As far as anomaly detection is concerned, quantitatively characterizing statistic of network traffic without attack is fundamental[2].

As shown by Leland[3] et al., and supported by a number of later research [4-5], the measurements of local and wide-area network traffic, wire-line and wireless network traffic all demonstrate that network traffic possesses self-similarity characteristic in large time-scale. Self-similarity is the property associated with the object whose structure is unchanged at different scales, and its degree can be described by the Hurst parameter.

Several studies show that DDoS flood attack can exert remarkable influence on the self-similarity of network traffic. Thus, this kind of attack can be effectively detected by monitoring the change of the Hurst parameter[6-7]. Existing flood attack detection methods based on the self-similarity nature of network traffic divide the network traffic into non-overlapping segments. The Hurst parameter of each segment is estimated, once the Hurst parameter changes beyond a pre-defined fixed threshold, the loss of self-similarity (LoSS) occurs and the DDoS flood attack is detected. However, the DDoS flood attack may take place at arbitrary moment whenever the traffic changes its self-similarity characteristic. The intensity of DDoS flood attack is also varying, which leads to changing Hurst parameter. Therefore, these existing fixed threshold detection methods lack flexibility and self-adaptability.

In this paper, we propose a DDoS flood attack detection method using discrete wavelet transform (DWT) and Schwarz information criterion (SIC) to determine the change point of self-similarity. The SIC[8] statistic is based on the maximum likelihood function for the model, and can be easily applied to change point detection by comparing the likelihood of the null hypothesis (i.e., no change in the variance series) against the alternative hypothesis (i.e., a change is present). This paper presents the SIC algorithm working with the DWT to detect the change point of self-similarity in real-time. After the change point detection, we use the fuzzy logic[9-15] to adaptively determine the intensity of the DDoS flood attack. We also design a set of decision rules for the fuzzy logic to determine the intensity of the DDoS flood attack. As a result, this proposed attack detection method can accurately detect not only the moment when the flood attack happens, but also the intensity of the attacks.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 gives a brief introduction to self-similarity and the relationship between the wavelet coefficients and the Hurst parameter. Section 4 first deduces the basic detection principle, and then presents the outline of the whole detection process. Section 5 describes the on-line self-similarity change point identification in detail. In section 6, the decision rules of the attack intensity are given. Section 7 discusses the performance of our method by NS2-based simulation with various network traffic characteristics and attack intensities. Finally, a brief summary of our work and future research are provided in section 8.

# 2 Related work

Several anomaly detection methods have been proposed against DDoS flood attack in the literature[16-18]. In these methods, the network traffic activity is captured and then a profile representing its stochastic behavior is created. This profile is mainly based on metrics such as the network traffic rate, the number of packets or bytes for each protocol, the rate of connections, the number of different IP addresses, etc. Any activity that deviates from the profile is treated as a possible attack.

There is a serious problem with these statistical anomaly detection methods. That is, it is hard to decide the appropriate metric on the global scale, because the linear superposition of these micro-based detection methods can not cope with the complex behavior of whole network. In 1993，Leland[3] et al. first found that the network traffic is self-similar and this attribute is one of the basic natures of the network traffic. Later, the work in [19] pointed out that the self-similarity of Internet traffic is attributed to a mixture of the actions of a number of individual users, and hardware and software behaviors at their originating hosts, multiplexed through an interconnection network. In other words, this self-similarity always exists regardless of the network type, topology, size, protocol, or the type of services the network is carrying.

The research done by Li[20] first mathematically proved that there is a statistically significant change in the average Hurst parameter under DDoS flood attack. Allen[21] et al. and W.Schleifer[22] et al. proposed a method using Hurst parameter to identify attack, which causes a decrease in the traffic's self-similarity. Those methods consider the normal range of Hurst parameter to be [0.5, 0.99], and there is an attack when the Hurst parameter runs out of this range. The experiment results demonstrate that the method proposed in [21-22] has an average detection rate of 60% to 84% depending on the intensity of the attack. Ren[23] et al. proposed using the wavelet analysis method to estimate the Hurst parameter, and consider there is an attack when the Hurst parameter runs out of the range [0.6, 0.9]. The cut down of normal range of Hurst parameter can be more efficient in detecting the low-rate DDoS flood attack. Nevertheless, all of these existing detection methods can only detect the presence of attack after the attack occurs, they can not identify at what time the attack happened.

Fuzzy logic is one of the most popular methods used in attack detection for it can deal with the vague and imprecise boundaries between normal traffic and different levels of attacks[10]. Wang[24] et al. proposed to use the fuzzy logic to analyze the Hurst parameter and estimate the time duration of DDoS attack. However, the work in [24] didn't consider the intensity of the attack traffic compared with the background traffic, therefore cannot accurately reflect the level of damage that is caused by the attack.

The major contributions of this paper are: (i) considering the inherent relationship between DWT and self-similarity, we propose to use SIC combined with DWT to detect the *occurrence* of the DDoS flood attack, therefore real-time DDoS attack detection is achieved; (ii) we propose a fuzzy set and its implementation to decide the intensity of DDoS flood attack *against the background traffic* dynamically and intelligently, which provides an accurate indication of the possible damage caused by the attack.

# 3 Self-similarity

## 3.1 A brief review of self-similarity

Self-similarity means that the sample paths of the process $W(t)$ and those of rescaled version $c^H W(t/c)$, obtained by simultaneously dilating the time axis $t$ by a factor $c>0$, and the amplitude axis by a factor $c^H$, can not be statistically distinguished from each other. Equivalently, it implies that an affine dilated subset of one sample path can not be distinguished from its whole. $H$ is called the self-similarity or Hurst parameter. For a general self-similar process, the parameter $H$ measures the degree of self-similarity.

Network traffic arrival process is a discrete time process, so the discrete time self-similarity definition will be used here. Let $X = \{x_i, i \in \square_+\}$ be a wide-sense stationary discrete stochastic traffic time series with constant mean $\mu$, finite variance $\sigma^2$, and autocorrelation

function $r(\tau), (\tau \in \mathbb{R}_+)$. Let $X^{(m)} = \{x_i^{(m)}, i, m \in \mathbb{R}_+\}$ be an $m$-order aggregate process of $X$, then

$$x_i^{(m)} = (x_{mi-m+1} + \cdots + x_{mi})/m . \quad (1)$$

For each $m$, $X^{(m)}$ defines a wide-sense stationary stochastic process with autocorrelation function $r^{(m)}(\tau)$.

**Definition 1**. A second-order stationary process $X$ is called exactly second-order self-similar (ESOSS) with Hurst parameter $H$=1-$\beta$/2, 0<$\beta$<1, if the autocorrelation function satisfies

$$r^{(m)}(\tau)=r(\tau), \quad (2)$$

where $r(\tau) = [(\tau+1)^{2-\beta} - 2\tau^{2-\beta} + (\tau-1)^{2-\beta}]/2$.

**Definition 2**. A second-order stationary process $X$ is called asymptotical second-order self-similar (ASOSS) with Hurst parameter $H$=1-$\beta$/2, 0<$\beta$<1, if the autocorrelation function satisfies

$$\lim_{m\to\infty} r^{(m)}(\tau) = r(\tau) . \quad (3)$$

In the field of network traffic theory, it is more practical to use ASOSS.

### 3.2 Wavelet-based Hurst parameter estimation

Currently, several methods have been proposed to estimate the Hurst parameter. Some of the most popular ones include the aggregated variance, local whittle, and the wavelet-based methods. However, the wavelet-based estimator[25] of the Hurst parameter stands out as one of the most reliable estimators in practice since it is more robust with respect to smooth polynomial trends and noise.

In this section, an on-line Hurst parameter estimation is proposed using the multiple resolutions feature of wavelet analysis. The estimation process is summarized as follows:

- **Wavelet decomposition:** For a given traffic trace time series $X$, we compute the wavelet coefficients $d(j,k)$ using a pyramidal filter bank in an on-line fashion for each scale $j$ and position $k$, as shown in Figure 1. At each level in the recursive structure, the bandpass (BP) output wavelet coefficients $d(j,\cdot)$, and the lowpass (LP) output scale coefficients $a(j,\cdot)$, occur at half rate of the input $a(j\text{-}1,\cdot)$.

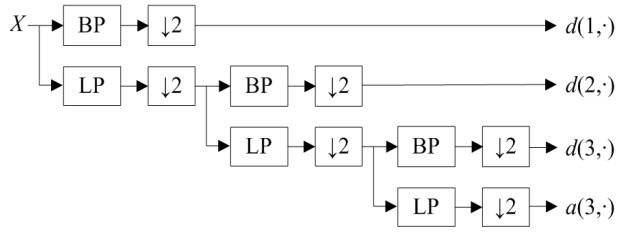



Figure 1: Pyramidal filter bank.

- **Detail variance estimation:** Let the current stored sum of squares calculated from the available wavelet coefficients at scale $j$ be

$$S_j = \sum_{k=1}^{n_j} d^2(j,k) , \quad (4)$$

where $n_j$ means the number of wavelet coefficients available at scale $j$. Assume that the arrival of a new traffic sample results in the new wavelet coefficient $d(j,n_j+1)$ at scale $j$ from the filter bank. The sum is then updated as follows:

$$\begin{aligned} n_j &\leftarrow n_j + 1 \\ S_j &\leftarrow S_j + d^2(j,n_j) \end{aligned} . \quad (5)$$

When the variance estimation at scale $j$ is required for the next step, it can be calculated as $\varepsilon_j=S_j/n_j$.

- **Analysis using the Logscale diagram:** We make a plot of $\log_2(\varepsilon_j)$ versus scale $j$ and apply a weighted linear regression over the curve region that looks linear, and then compute the slope $\alpha$. There is no need to compute the Logscale diagram every time a new traffic sample is acquired, since they may be recalculated only when needed.
- **Hurst parameter estimation:** The Hurst parameter $H$ can be estimated according to

$$H=(\alpha+1)/2 . \quad (6)$$

This on-line wavelet-based Hurst parameter estimation is performed in an accumulative way, that is, it returns the updated Hurst parameter computed over all available samples from beginning to current time. We can see that this estimation method is accumulative but not dynamic, thus can not be used directly in detecting the change point of self-similarity in the network traffic. Therefore, a change point detection method combined with DWT is proposed and discussed in detail in section 5.

## 4 Attack detection process

### 4.1 Attack detection principle

Let $X = \{x_i, i \in \mathbb{R}_+\}$ and $Y = \{y_i, i \in \mathbb{R}_+\}$ be normal and abnormal traffic respectively, and $Z = \{z_i, i \in \mathbb{R}_+\}$ be the attack traffic during transition process of attacking. $X$ and $Z$ are uncorrelated[2], so $Y$ can be abstractly expressed by $Y$=$X$+$Z$.

Figure 2 illustrates the components of normal and abnormal traffic. $x_i(p)$ represents the number of bytes sent out by node $p$ at time $i$ for normal network services, and $z_i(q)$ represents the number of bytes sent out by node $q$ at time $i$ for network attack, and $y_i$ is the total traffic the target received at time $i$.

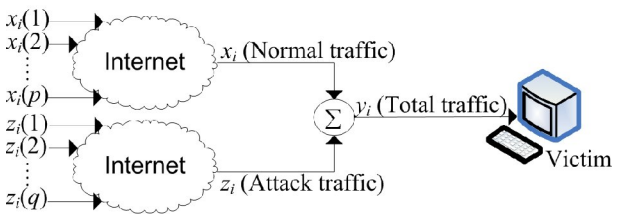



Figure 2: Composition of normal and attack traffic.

Based on the theorems in reference [26], we can get the conclusion that no matter whether $Z$ is a self-similar process, if $X$ is a second-order stationary self-similar

process, then $Y$ will still be a self-similar process, but the degree of self-similarity may be changed. Let $r_X$, $r_Z$ and $r_Y$ be the autocorrelation functions of $X$, $Z$ and $Y$ respectively. During the attack, $\|r_Y - r_X\|$ is noteworthy[2], and $r_Y = r_X + r_Z$. For each value of $H \in (0.5,1]$, there is exactly one autocorrelation function with self-similarity[27]. Thus, the consequence is that $\|H_Y - H_X\|$ is considerable, where $H_Y$ and $H_X$ are average Hurst parameters of $Y$ and $X$, respectively. Hence, $H$ is a parameter that can be used to describe the abnormality of network traffic.

### 4.2 Outline of the attack detection process

The whole process of the DDoS flood attack detection is displayed in Figure 3.

From Figure 3, we can see that the whole detection process consists of two stages: on-line attack moment identification and intelligent attack intensity decision. In the part of attack moment identification, the wavelet coefficients and SIC statistic will be updated along with the incoming of new traffic samples, then the change point detection will re-run in every scale to find out whether there is a change point. It will signal a change point of self-similarity in network traffic if change points exist in enough scales at the same moment. After the attack moment identification, we then segment the network traffic into pieces around the identified attack moment. After that, we can decide the intensity of the attack using intelligent fuzzy logic technology. According to the Hurst parameter and its changing rate (the difference between the Hurst parameters of traffic pieces prior to and after the identified attack moment), we can determine the intensity of the DDoS flood attack using the fuzzy decision rules. The next two sections present the detailed implementation of attack moment identification and attack intensity decision.

## 5 Change point estimation with SIC

### 5.1 SIC

The SIC is a powerful approach in detecting the change point of self-similarity in network traffic[8]. The principle of SIC is that a sequence with a variance change point has higher entropy than a sequence with constant variance.

Given a sequence of length $M$, and suppose there is only one change point at position $g$ ($1<g<M$). The way of simultaneously detecting the presence and location of this change point is to compute the entropy of the entire sequence and of the pairs of pieces ($f_1=1,\ldots, g$ and $f_2=g+1,\ldots, M$), compare their values and then decide if there is a change point at position $g$ according to whether the entropy of the pieces is significantly lower than the entropy of the entire sequence.

We test the null hypothesis $A_0$ (no change is present) against the alternative $A_1$ (a single change is present). Assuming gaussianity and independence[8], the SIC statistic for the two hypotheses is given by

$$A_0 : \mathrm{SIC}(M) = M\log 2\pi + M\log\hat{\sigma}^2 + M + \log M,$$

$$A_1 : \mathrm{SIC}(g) = M\log 2\pi + g\log\hat{\sigma}_1^2 + (M-g)\log\hat{\sigma}_2^2 + M + 2\log M,$$

where $\hat{\sigma}^2$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the unbiased maximum likelihood estimators (MLEs) of the variances of the entire sequence and of the first and second pieces, respectively.

Our decision will follow the principle of minimum information, that is, $A_0$ will not be rejected if

$$\mathrm{SIC}(M) \le \min_g \mathrm{SIC}(g),$$

otherwise $A_0$ will be rejected if

$$\mathrm{SIC}(M) > \mathrm{SIC}(g),$$

for some $g$. The change point at position $g$ can be estimated according to

$$\mathrm{SIC}(g) = \min_{1\le g<M} \mathrm{SIC}(g). \quad (7)$$

Reference [28] gives a proof that $g$ is a consistent estimator of the true change point, and it also gives the expression for computing the signification level of SIC statistic. The analytic study by Tian[29] et al. shows that the Hurst parameter has close relationship with variance structure of wavelet coefficients. The SIC has the merit of detecting the change point in the variance structure of the sequence, so the combination of DWT and SIC can be used to detect the change point of Hurst parameter in the network traffic.

### 5.2 Connecting DWT and SIC

In this section, we apply the SIC change point detection to the wavelet coefficients $d(j,k)$ at each scale $j$. It will signal a change point of Hurst parameter if we find the same variance change position across all or a significant number of scales. A change in variance at a single scale only tells us of non-stationary in the variance at that scale

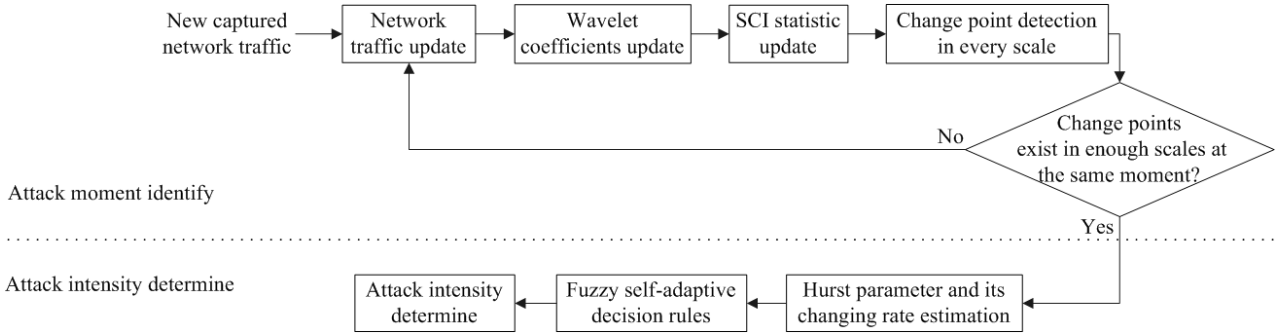



Figure 3: Diagram of DDoS flood attack detection

For the DWT, the number of wavelet coefficients at each scale is not the same, since each branch of the filter bank suffers a different number of decimations. Let $n_j$ be the number of available wavelet coefficients at scale $j$ and $n_{j+1}$ be the number of wavelet coefficients at scale $j$+1, then the relationship between adjacent scales of $n_j$ and $n_{j+1}$ satisfies $n_j=2n_{j+1}$. Figure 4 (a) shows the temporal relationship between the wavelet coefficients of each scale. We can see that the wavelet coefficients of each scale in the dotted line correspond to certain length of network traffic. In order to provide a good estimation of the change point at all available scales, a phase correction should be applied to higher scales, with a scale dependent delay as illustrated in Figure 4 (b). This phase correction aligns the position of the wavelet coefficients with a scale-dependent delay and with their zone of influence to the change point detection.

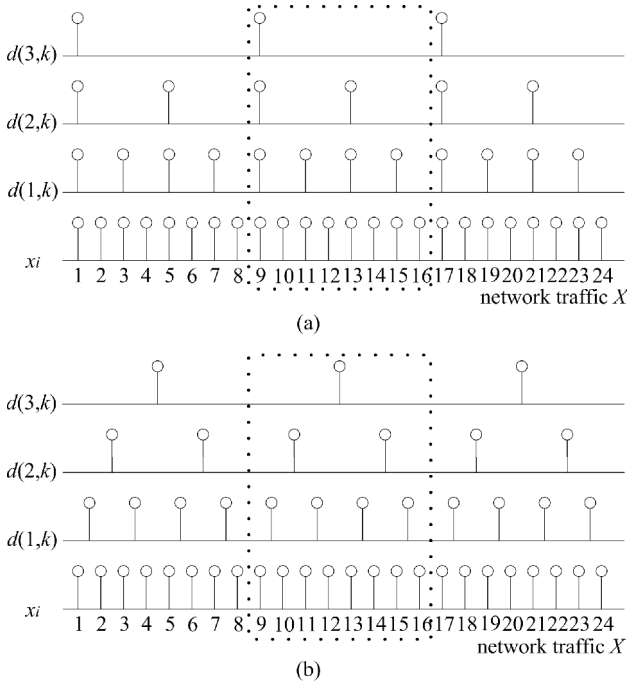



Figure 4: Temporal relationship between: (a) wavelet coefficients; (b) phase corrected wavelet coefficients.

There are two main requirements for the attack detection performance, one is accurate detection and the other is real-time detection. In the subsequent section, a real-time detection method is emphasized, which performs DWT and SIC statistics in sequential fashion, and with a slide-window to detect the change point of self-similarity on-line.

## 5.3 On-line change point estimation

The on-line detection scheme of the change point of self-similarity in network traffic comprises three steps: network traffic update using slide-window, wavelet coefficients update using pyramidal filter bank and SIC statistic update with new wavelet coefficients available at each scale.

- **Network traffic update:** Let $l$ represent the original number of network traffic time series used to detect the change point of self-similarity, and $h$ represent the size of slide-window. The process of updating network traffic is displayed in Figure 5. When $h$ samples of new traffic are acquired, we add these new traffic samples into the tail of original traffic and discard the same number of old samples at the head of original traffic simultaneously, so the reserved network traffic for the next detection is $l$-$h$. This process is iterated whenever new samples are acquired. By doing so, we can guarantee that there is only one change point of self-similarity in this finite length of traffic before we perform the detection.

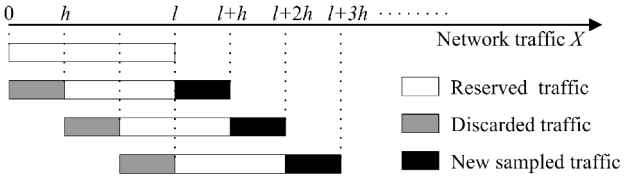



Figure 5: Network traffic update.

- **Wavelet coefficients update:** For the new acquired $h$ traffic samples, we feed them into the pyramidal filter bank displayed in Figure 1. The transforms give the new wavelet coefficients corresponding to the convolution of the new acquired $h$ samples and the memory of the filter. For example, let $h$ be 64 ($2^6$), then only wavelet coefficients at scales 1~5 need to be updated. Therefore, the change points in the variance structure will appear progressively in the higher scales as new samples are acquired. For the discarded traffic, we only need to discard the wavelet coefficients of each scale related to the discarded traffic, as we can see in Figure 4 (b).
- **SIC statistic update:** The update of SIC statistic is relatively simple to implement, since it only requires the variance of the new wavelet coefficients at each scale to be added and the variance of the discarded wavelet coefficients at each scale to be discarded. We then re-test the null hypothesis $A_0$ against the alternative $A_1$ at each scale to find out whether there are change points in the variance structure.

A decision that there is a change point of self-similarity in network traffic can be made if change points in the wavelet coefficients variance structure exist in enough scales at the same moment. A change in the wavelet coefficients variance structure in one scale or a few scales only tells us the non-stationary at that scale. After the change point of self-similarity detection in network traffic, we then segment the network traffic into pieces around the identified change moment. After that, we can determine the intensity of DDoS flood attack using intelligent fuzzy logic technology, which takes the Hurst parameter and its changing rate as decision-making basis.

# 6 Attack intensity decision with intelligent fuzzy logic

## 6.1 Intelligent fuzzy logic

Intelligent fuzzy logic decision disposes information based on fuzzy or non-fuzzy reasoning rules[10-12]. It makes self-adaptive decision in light of mature experience. The general fuzzy decision process consists of three parts: fuzzy quantitative disposal, fuzzy decision rules and fuzzy decision. The fuzzy quantitative disposal makes the real input parameter as a fuzzy set, and then the fuzzy decision carries out the output calculation based on the fuzzy set and fuzzy operators defined at fuzzy decision rules. This section will describe in detail how fuzzy logic can be utilized in DDoS flood attack intensity decision.

## 6.2 Attack intensity decision

Based on the basic theory and method of fuzzy mathematics, we propose an intelligent DDoS flood attack intensity decision system. DDoS flood attack intensity itself includes fuzziness, because the boundary between the light attack, moderate attack and severe attack is not well defined. So when judging the intensity of attack, one should take the intensity of background traffic into consideration. For example, a DDoS flood attack is considered as light attack if it causes slight decline of the network performance when the traffic load is high, but is considered as severe attack if it causes serious decline of the network performance when the network load is light.

In the proposed decision system, the DDoS flood attack intensity decision rules and operations are expressed by fuzzy sets, and then we feed these fuzzy decision rules and related information into knowledge repository. The network elements take the dynamic process of actual attack into consideration, and then use fuzzy reasoning to determine the intensity of attack dynamically and intelligently.

In this paper, the structure of fuzzy decision is two-dimensional input and one-dimensional output. The two inputs are the Hurst parameter and its changing rate. The Hurst parameter reflects the influence of dynamic normal traffic on attack intensity and the changing rate reflects the influence of attack on normal traffic. The output is the intensity of the attack. As shown in Figure 6, the fuzzy decision process of the intensity of the attack consists of three parts: Hurst parameter and its changing rate fuzzification, fuzzy decision rules of attack intensity and fuzzy reasoning of attack intensity.

The description of each part of the fuzzy decision process is as follows:

- **Hurst parameter and its changing rate fuzzification:** Fuzzification makes the real input parameters of Hurst parameter and its changing rate as fuzzy sets. According to the change scope of Hurst parameter and its changing rate, we define the universe of discourse of the Hurst parameter as $UH$={0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1}; the universe of discourse of the changing rate as $UHC$={0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}. The fuzzy sets of $UH$ and $UHC$ are $H'$={$S$, $M$, $B$} and $HC'$={$S$, $M$, $B$}, where "$S$" stands for small, "$M$" the moderate, and "$B$" the big. The variable's membership degree function of each fuzzy language satisfies normality assumption
$$\mu(x)=\exp\left[-(x-v)^2/b^2\right], \quad (8)$$
where $v$ and $b^2$ are the mean and variance of the membership degree function. Through Eq. (8), we can obtain fuzzy judgment model of every parameter as well as the membership degree assignments of every fuzzy subset.
- **Fuzzy decision rules of attack intensity:** The decision rules take note of the relationship between input fuzzy sets and output fuzzy sets. Define the fuzzy decision result of DDoS flood attack intensity as a variable $L$, and the fuzzy set of $L$ as $L'$={$LA$, $MA$, $SA$}, where "$LA$", "$MA$" and "$SA$" represent light DDoS flood attack, moderate DDoS flood attack and severe DDoS flood attack, respectively. Considering the relationship between Hurst parameter, its changing rate and DDoS flood attack intensity, we can get the fuzzy decision rules displayed in Table 1.

Table 1. The fuzzy decision rules of the DDoS flood attack.

| $HC'$ | $H'$ | | |
|---|---|---|---|
| | $S$ | $M$ | $B$ |
| $S$ | $MA$ | $LA$ | $LA$ |
| $M$ | $SA$ | $MA$ | $LA$ |
| $B$ | $SA$ | $SA$ | $MA$ |

- **Fuzzy reasoning of attack intensity:** After fuzzifying the input parameters Hurst parameter and its changing rate, we can reason the intensity of attack according to decision rules presented in Table 1. For example, when the Hurst parameter is considered moderate, we infer there is a light DDoS flood attack if the changing rate of the Hurst parameter is considered small. In a similar way, there is a moderate DDoS flood attack if the changing rate of the Hurst parameter is moderate, and severe DDoS flood attack if the changing rate of the Hurst parameter is big.

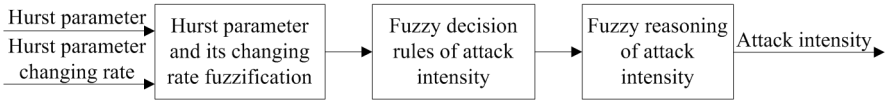



Figure 6: Intelligent fuzzy decision process.

# 7 Experiments and analysis

## 7.1 Simulation environment

The test traffic time series is constructed using NS2 simulator with varying parameters (e.g. the self-similarity degree of normal traffic and the attack intensity). The simulated traffic model is shown in Figure 2, and we let the number of nodes that provide normal service is $p$=100, and the number of nodes that implement the attack is $q$=100.

We conducted our simulation in two steps: First we generate the normal traffic using fractional Gaussian noise (=fGn) model with Hurst parameter $H$={0.6, 0.7, 0.8, 0.9}. The FGN model was first introduced by Mandelbrot and Van Ness[30], and now it is widely used in network traffic modeling for its simplicity and mathematically attractive. Other traffic models also can be used in this simulation in the same way. Second we inject constant rate attack traffic at time 600 (second) with maximum attack intensity varying from 100KBps to 500KBps. The constant rate attack achieves its maximum rate immediately and lasts for about 400 seconds. The normal traffic will persist for another 200 seconds after the attack stops.

The simulated abnormal traffic trace with different self-similarity degree and attack intensity is displayed in Figure 7. The merging time scale is 100 ms.

In Figure 7, every piece of traffic in the same column has same self-similarity degree with Hurst parameter displayed at the top of the column, and every piece of traffic in the same row suffered from same attack intensity. From the first row to the last row, the attack intensity is 100KBps, 200KBps, 300KBps, 400KBps, and 500KBps, respectively.

## 7.2 Test results and analysis

We use Daubechies(3) as mother wavelet and set the decomposition level to 6. In the experiment, the size of slide-window is $h$=64, and $l$=8$h$. Under the condition of significance equal to $10^{-5}$, we can identify the change points of Hurst parameter at points 6000 and 10000 as shown in Figure 8. In the simulation, point 6000 is when attack happens and point 10000 is when attack stops.

Figure 8 shows the results of phase corrected 6-level DWT-SIC statistics analysis of the simulated traffic trace. For normal traffic with same degree of self-similarity, we can see that: (i) At lower scales, the change points caused by both light DDoS flood attack and severe DDoS flood attack are clearly identified; (ii) At higher scales, the change points caused by light DDoS flood attack disappear because the DWT lacks temporal accuracy. Under the condition of same attack intensity, we can see that more change points appear at higher scales when self-similarity degree is higher, which means the higher degree of traffic self-similarity, the more sensitive the network is to DDoS flood attack.

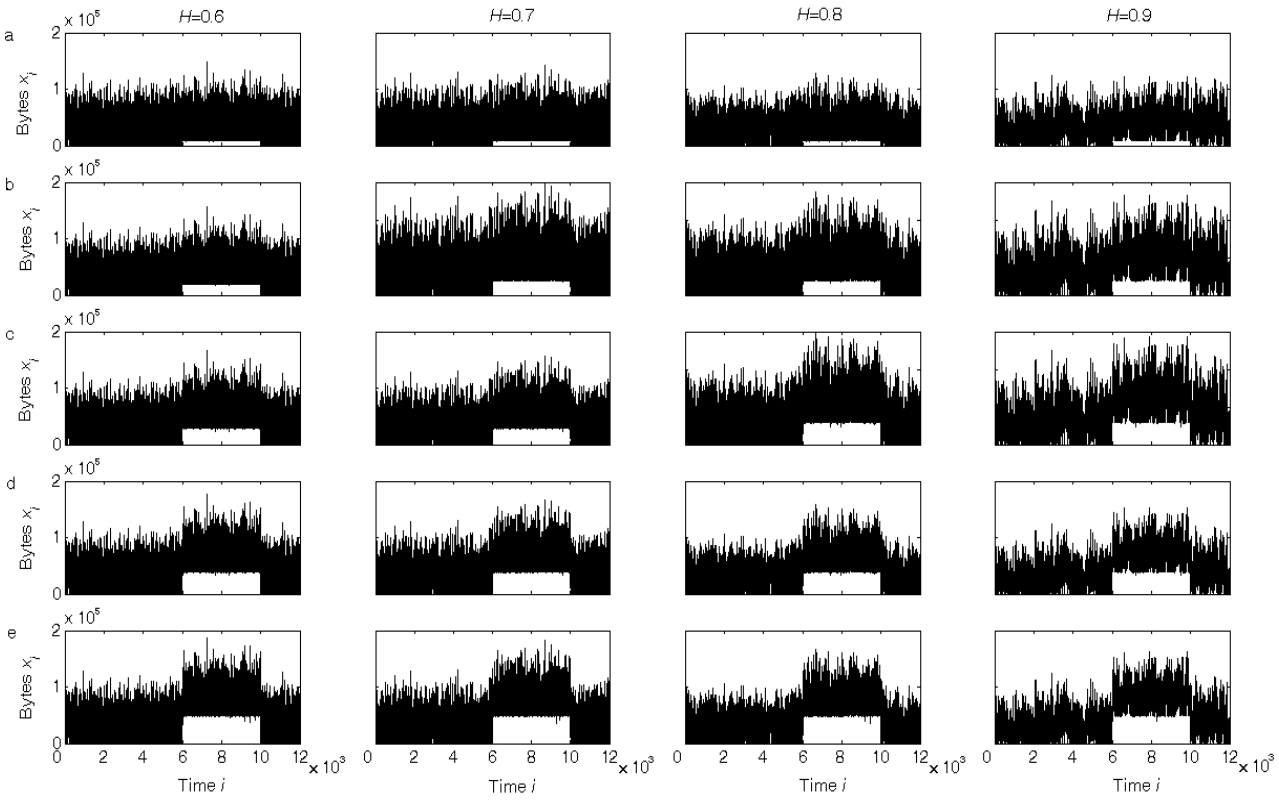



Figure 7: Simulated abnormal traffic trace: (a) attack intensity-100KBps; (b) attack intensity-200KBps; (c) attack intensity-300KBps; (d) attack intensity-400KBps; (e) attack intensity-500KBps.

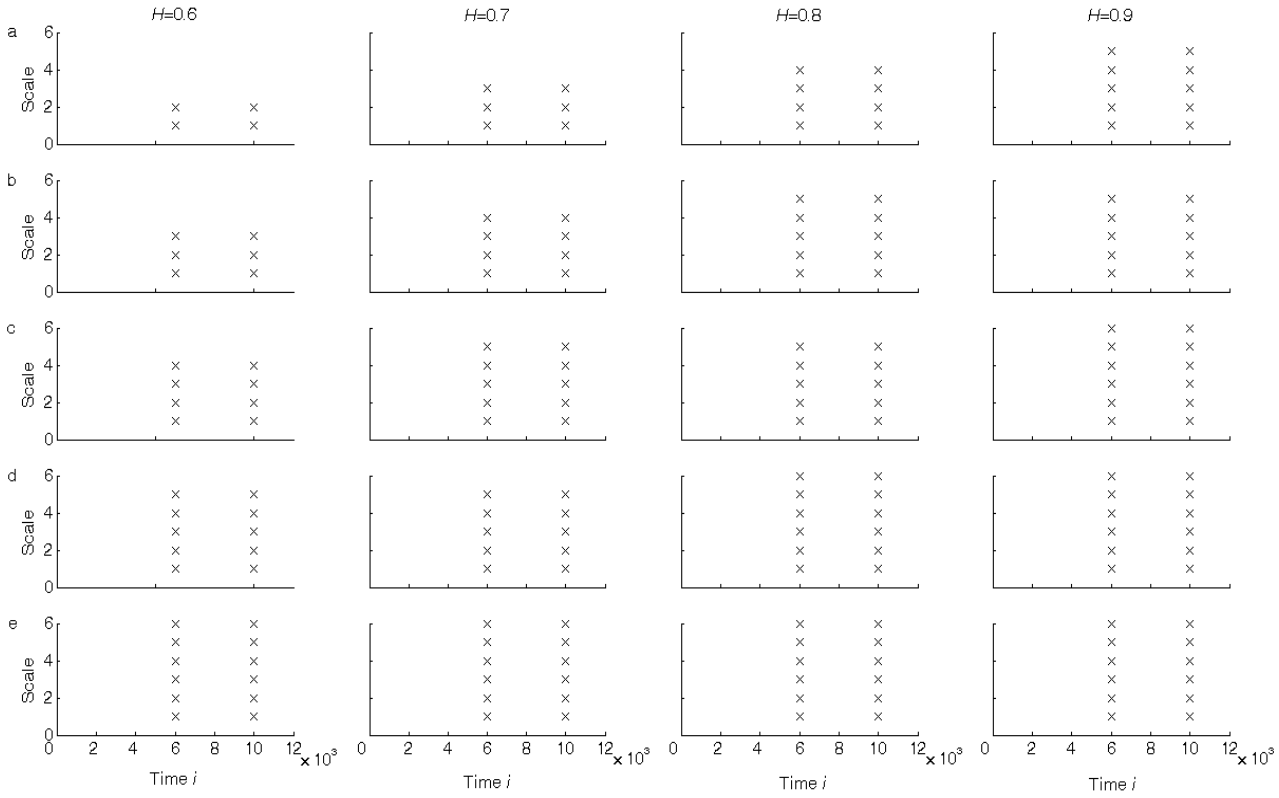



Figure 8: Change point candidates at each scale. The test traffic trace: (a) attack intensity-100KBps; (b) attack intensity-200KBps; (c) attack intensity-300KBps; (d) attack intensity-400KBps; (e) attack intensity-500KBps.

We segment the simulated abnormal traffic around the identified change points 6000 and 10000, and estimate the Hurst parameters of first and second traffic pieces. The changing rate of Hurst parameter can be computed by calculating the difference of these two Hurst parameters. Table 2 displays the changing rate of Hurst parameter under different degree of network traffic self-similarity and DDoS flood attack intensity.

Table 2. Input parameters in fuzzy decision of DDoS flood attack.

| Attack intensity | *H* | | | |
|---|---|---|---|---|
| | 0.6 | 0.7 | 0.8 | 0.9 |
| 100KB | 0.011 | 0.018 | 0.026 | 0.035 |
| 200KB | 0.030 | 0.043 | 0.061 | 0.088 |
| 300KB | 0.053 | 0.074 | 0.105 | 0.152 |
| 400KB | 0.082 | 0.119 | 0.170 | 0.248 |
| 500KB | 0.120 | 0.187 | 0.272 | 0.397 |

We select the input parameters *H* and *HC* in Table 2, and put them into the fuzzy logic decision process shown in Figure 6. We first fuzzify *H* and *HC* as fuzzy sets based on the membership degree function defined in Eq. (8). The mean and variance of the membership degree function is 0 and 1. According to the decision rules in Table 1, we can get the decision results of DDoS flood attack intensity shown in Table 3.

Table 3. Fuzzy decision results of DDoS flood attack.

| *HC* | *H* | | | |
|---|---|---|---|---|
| | 0.6 | 0.7 | 0.8 | 0.9 |
| 0.011 | *LA* | *LA* | *LA* | *LA* |
| 0.018 | *LA* | *LA* | *LA* | *LA* |
| 0.026 | *LA* | *LA* | *LA* | *LA* |
| 0.030 | *LA* | *LA* | *LA* | *LA* |
| 0.035 | *MA* | *LA* | *LA* | *LA* |
| 0.043 | *MA* | *LA* | *LA* | *LA* |
| 0.053 | *MA* | *MA* | *LA* | *LA* |
| 0.061 | *MA* | *MA* | *LA* | *LA* |
| 0.074 | *MA* | *MA* | *MA* | *LA* |
| 0.082 | *MA* | *MA* | *MA* | *LA* |
| 0.088 | *SA* | *MA* | *MA* | *LA* |
| 0.105 | *SA* | *MA* | *MA* | *MA* |
| 0.119 | *SA* | *MA* | *MA* | *MA* |
| 0.120 | *SA* | *MA* | *MA* | *MA* |
| 0.152 | *SA* | *SA* | *MA* | *MA* |
| 0.170 | *SA* | *SA* | *MA* | *MA* |
| 0.187 | *SA* | *SA* | *MA* | *MA* |
| 0.248 | *SA* | *SA* | *SA* | *MA* |
| 0.272 | *SA* | *SA* | *SA* | *MA* |
| 0.397 | *SA* | *SA* | *SA* | *SA* |

### 7.3 Comparison with existing detection method

In order to compare our proposed detection method with the existing self-similarity based detection methods, we carry out the following experiments. Firstly, we divide each network traffic with different degree of self-similarity into non-overlapping sections of length $l$=8$h$, where $h$=64. Secondly, we estimate the Hurst parameter of each section using Abry-Veitch wavelet-based estimator[31], and the Matlab source code for this estimator is available at [32]. The Hurst parameter of each section under different attacks intensity is displayed in Figure 9.

In Figure 9, we can see that the Hurst parameters of section 12 (corresponds to points 5633-6144) and section 20 (corresponds to points 9729-10240) are larger than the Hurst parameter before the attack happens. This is because at the beginning and ending moments of attack, the network traffic becomes more bursty and non-stationary, which leads to increase of the Hurst parameter. But during the attack, the normal traffic is overwhelmed by the attack traffic, and the Hurst parameter decreases (from section 13 to section 19, corresponding to points 6155-9728). Using the detection threshold proposed by Allen[21], we find that those attacks with severe intensity can be detected properly, but attacks with light or moderate intensity will be missed. Using the detection threshold proposed by Ren[23], we find that those attacks with severe or moderate intensity can be detected properly, but attacks with light intensity are missed. To make things even worse, in Ren's method, normal traffic with light degree of self-similarity and high degree of self-similarity are taken as attack behaviors. Our detection method takes the self-similarity degree of normal traffic into account, and study the influence of different attack intensities to the network traffic self-similarity. So our detection method can detect light, moderate and severe intensity of attacks accurately and intelligently.

In addition, in methods proposed by Allen and Ren, it is important to choose a proper length of the section($l$), because short section can not guarantee the amount of data required for estimating the Hurst parameter, and long section will result in prolonged detection latency. But our detection method does not suffer from this problem, because if there is a change point of self-similarity in network traffic, we only need to sample a few more data after the change point, then we can detect this change point timely by detecting the changes of wavelet coefficients variance structure at several scales. For example, if the wavelet decomposition level is 5, then we only need to sample another 64($2^6$) sample data to find out this change point.

## 8 Conclusion

In this paper, we proposed a method to detect the occurrence and intensity of DDoS flood attack based on the change of self-similarity in network traffic. To identify the DDoS flood attack, we adopt a kind of Schwarz information criterion that can not only find out the presence of attack in network traffic, but also its occurring moment. After the attack identification, we further propose a method to determine the intensity of attack based on intelligent fuzzy logic technology. To verify the effectiveness of our method, we conducted experiments using traffic trace constructed by NS2 simulator. The results demonstrate that the proposed method can detect the DDoS flood attack timely, effectively and intelligently. The future work will focus

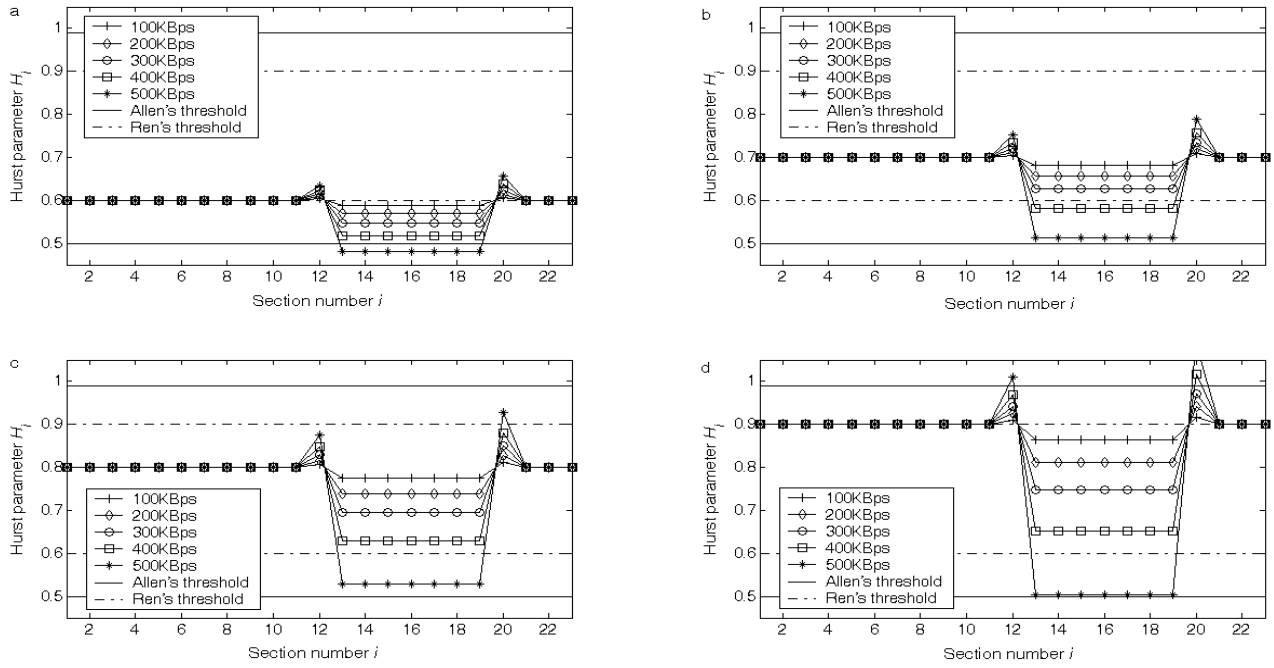



Figure 9: Change trend of Hurst parameter under different DDoS flood attack intensity. The degree of normal traffic self-similarity is: (a) 0.6; (b) 0.7; (c) 0.8; (d) 0.9.

on constructing fuzzy rule base by some learning techniques and testing this method on traffic trace from live networks.

## Acknowledgement

This work was supported in part by the National High Technology Research and Development Program of China under Grant No.2007AA01Z473; the National Natural Science Foundation of China under Grant No.60605019 and No. 60702047. The authors would also like to thank Patrice and Darryl for providing the Matlab source code and other members of Lab of Information Security Integrated Management Research for their valuable suggestions that have considerably increased the quality of this paper.

# References

[1] http://www.cert.org

[2] M. Li. An approach to reliably identifying signs of DDOS flood attacks based on LRD traffic pattern recognition. *Computers & Security*, 23(7): 549-558, 2004.

[3] W.E. Leland, M.S. Taqqu and W. Willinger. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1): 1-15, 1994.

[4] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3): 226-244, 1995.

[5] O. Tickoo and B. Sikdar. On the impact of IEEE 802.11 MAC on traffic characteristics. *IEEE Journal on Selected Areas in Communications*, 21(2): 189-203, 2003.

[6] C.S. Sastry, S. Rawat and A.K. Pujari. Network traffic analysis using singular value decomposition and multiscale transforms. *Information Sciences*, 177(23): 5275-5291, 2007.

[7] M.F. Rohani, M.A. Maarof and A. Selamat. Continuous LoSS detection using iterative window based on SOSS model and MLS approach. *In Proceedings of the International Conference on Computer and Communication Engineering*, Kuala Lumpur, Malaysia, May 2008.

[8] D. Rincón and S. Sallent. On-line segmentation of non-stationary fractal network traffic with wavelet transforms and Log-likelihood-based statistics. *LNCS*, 3375: 110-123, 2005.

[9] N. K. Swain. A survey of application of fuzzy logic in intelligent transportation systems (ITS) and rural ITS. *In Proceedings of the IEEE Southeast Conference,* March 2006.

[10] S.X. Wu and W. Banzhaf. The use of computational intelligence in intrusion detection systems: A review. *Applied Soft Computing*, 10: 1-35, 2010.

[11] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3): 338-353, 1965.

[12] A. Abraham. Neuro-fuzzy systems: State-of-the-art modeling techniques, connectionist models of neurons, learning processes, and artificial intelligence. *LNCS*, 2084: 269-276, 2001.

[13] A. Meier and N. Werro. A fuzzy classification model for online customers. *Informatica,* 31(2): 175-182, 2007.

[14] J. Abonyi and B. Feil. Computational intelligence in data mining. *Informatica*, 29(1): 3-12, 2005.

[15] S. Avdoshin and V. Serdiouk. Some approaches to information security of communication networks. *Informatica*, 26(1): 1-10, 2002.

[16] C. Douligeris and A. Mitrokotsa. DDoS attacks and defense mechanisms: classification and state-of-the-art. *Computer Networks*, 44(5): 643-666, 2004.

[17] A. Patcha and J. M. Park. An overview of anomaly detection techniques: existing solutions and latest technological trends. *Computer Networks*, 51(12): 3448-3470, 2007.

[18] P. García-Teodoro, J. Díaz-Verdejo and G. Maciá-Fernández. Anomaly-based network intrusion detection: techniques, systems and challenges. *Computers & Security*, 28(1-2): 18-28, 2009.

[19] W.B. Gong, Y. Liu and V. Misra. Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications. *Computer Networks*, 48(3): 377-399, 2005.

[20] M. Li. Change trend of averaged Hurst parameter traffic under DDOS flood attacks. *Computers & Security*, 25(3): 213-220, 2006.

[21] W.H. Allen and G.A. Marin. The LoSS technique for detecting new denial of service attacks. *In Proceedings of IEEE South East Conference*, Greensboro, NC, March 2004.

[22] W. Schleifer and M. Männle. Online error detection through observation of traffic self-similarity. *In Proceedings of the IEE Communications Conference*, 148(1): 38-42, 2001.

[23] X.Y. Ren, R.C. Wang and H.Y. Wang. Wavelet analysis method for detection of DDoS attack on the basis of self-similarity. *Frontiers of Electrical and Electronic Engineering in China*, 2(1): 73-77, 2007.

[24] J.T. Wang and G. Yang. An intelligent method for real-time detection of DDoS attack based on fuzzy logic. *Journal of Electronics (China)*, 25(4): 511-518, 2008.

[25] S. Stoev, M. Taqqu and C. Park. On the wavelet spectrum diagnostic for Hurst parameter estimation in the analysis of Internet traffic. *Computer Networks*, 48(3): 423-445, 2005.

[26] S. Song and K. Joseph. Some results on the self-similarity property in communication networks. *IEEE Transactions on Communications*, 52(10): 1636-1641, 2004.

[27] J. Beran. Statistics for long-memory processes. *Chapman & Hall*, New York, 1994.

[28] J. Chen and A. Gupta. Testing and Locating Variance Change points with Application to Stock Prices. *Journal of the* American *Statistical Association*, 92: 739-747, 1997.

[29] X. Tian, J. Wu and C. Ji. A unified framework for understanding network traffic using independent wavelet models. *In Proceedings of IEEE INFOCOM'*, June 2002.
[30] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4): 422-437, 1968.
[31] A. Patrice and V. Darryl. Wavelet analysis of Long-Range-Dependence traffic. *IEEE Transactions on Information Theory*, 44(1): 2-15, 1998.
[32] http://www.cubinlab.ee.unimelb.edu.au/%7Edarryl

# Multisignature Scheme Based on Discrete Logarithms in the Plain Public Key Model

Zuhua Shao
Zhejiang University of Science and Technology, P. R. of China
E-mail: zhshao_98@yahoo.com

**Keywords:** discrete logarithm, random oracle model, group oriented cryptography, multisignature.

**Received:** November 18, 2008

*In this paper, we propose a new multisignature scheme based on discrete logarithms. We show that this new scheme can resist existential forgeries against adaptive chosen-message attacks in the random oracle model. The main contribution is that our security model gets rid of the special security requirement on the generation of the signers' public keys. Adversaries are not required to reveal private keys corresponding to the public keys of its choice to the challenger in attack games. Thus the new multisignature scheme does not suffer from the problem identified by Micali et al., which is shared by many current multisignature schemes. Moreover, if the joint public key of a group of signers in this multisignature scheme is precomputed, the proposed multisignature scheme is optimal.*

*Povzetek: Opisana je shema podpisov za zaščito javnih ključev.*

## 1 Introduction

Society oriented cryptography is a notion introduced by Desmedt [1]. A society oriented signature is essentially like a single signature except that is generated by plural individuals simultaneously.

A multisignature scheme is one kind of society oriented signature scheme, which allows multiple signers to sign the same message in a collaborative and simultaneous manner. A trivial solution is that every signer signs the message using a normal signature scheme respectively. Obviously, this simple solution will meet the security requirements for the multisignature scheme if the underlying signature scheme is secure. Its main drawback, however, is that both the data expansion and the computation costs for verification increase linearly with the number of signers in the group. Harn [2] submitted two additional properties that need to be achieved in the design of an optimal multisignature scheme:

1. The size of a multisignature should be identical to that of an individual signature.
2. The verification process of a multisignature should be almost identical to that of an individual signature.

Hence, in an optimal multisignature scheme, not only the size of signatures is independent of the number of signers participating in signing, but also the computation costs for verification.

Since the notion of multisignatures was introduced by Itakura and Nakamura [3], there have been many multisignature schemes proposed in literatures. However, most importantly, until the works of Ohta and Okamoto [4] and of Micali et al. [5], there were no formal security models for multisignatures. This lack of formalism has led not only to some confusion as to the precise security requirements for multisignatures, but also to some multisignature schemes having been subsequently broken [6, 7].

In group oriented cryptosystems, we must consider the possibility that an adversary could corrupt some fraction of participants, and thereby comes into possession of their private keys. We even allow the adversary to specify the public keys of the corrupted participants. In so-called rogue-key attacks, the adversary would register public keys created as a function of public keys of other honest participants. This kind of attacks could be extremely danger to break some multisignature schemes. The security notion of Ohta and Okamoto [4] is not strong enough to withstand such rogue-key attacks in the key generation.

Micali et al. [5] gave the first strong security notion for multisignature schemes in the plain public key model. They discussed a series of more sophisticated approaches based on zero-knowledge proofs, by which the private keys corresponding to the public keys can be extracted. Their scheme requires, as a pre-processing step, that the set of potential signers engage in an interactive key generation protocol to generate their key pairs. Besides expensive and resulting in complex public keys, this dedicated key generation enforces the set of potential signers to be static.

Boldyrva [8] proposed an efficient multisignature scheme based on the Gap-Diffie-Hellman group. Lu et al. [9] proposed the first multisignature scheme from pairings, provably secure without random oracles. Their security models allow an adversary to create arbitrary public keys for the corrupted signers possibly dependent on the public keys of the honest signers. But they require the adversary to prove the knowledge of the

corresponding secret keys (*KOSK*) during the public key registration. For simplicity, it has the adversary to hand over the secret keys of the corrupted signers in key generation algorithm. However, this *KOSK* assumption is not realized by existing public key infrastructure (PKI). Key registration protocols specified by the most widely used standards, PKCS#10 [10] - used by VeriSign and RFC 4210 [11, 12] do not specify that the Certification Authority (*CA*) should require proofs of knowledge, instead, specify that the *CA* should require proofs of possession (*POP*). That is, applicant is required to hand over the *CA* a signature, under the public key it is attempted to get certified, of some message that includes the public key and the identity of the applicant.

While such requirement might intuitively appear to stop adversaries from picking rogue keys, it does not suffice to realize the security models of [8, 9]. Ristenpart and Yilek [13] analyzed these schemes when key registration requires *POP*s. They showed that the standardized *POP* mechanism does not lead to secure multisignatures. Both schemes fall to rogue-key attacks despite the use of standardized *POP*. They presented a straightforward and natural fix for this problem: simple use separate hash function for *POP*s and multisignatures at the cost of upgrading existing PKI.

In 2006, Bellare and Neven [14] proposed a new multisignature scheme in the plain public key model, requiring nothing more than each signer has a (certified) public key in GF($p$), which means neither *KOSK* or *POP* is required in key registration protocols. They provided a security proof in the random oracle model. However, their scheme is less efficient than the original Schnorr signature since the computation of verification increases linearly with the number of signers in the group.

In this paper, we also propose a new multisignature scheme based on Discrete Logarithms (DL) in the plain public key model. We show that this new scheme can resist existential forgeries against adaptive chosen-message attacks in the random oracle model. The main contribution is that our security model gets rid of the special security requirement on the generation of participants' public keys. Namely, like [14], our security model not only allows so-called rogue-key attacks in the key generation, but also gives the adversary complete freedom in specifying the public keys of the corrupted signers. The adversary is no longer enforced to prove either knowledge or possession of the private keys corresponding to the public keys of its choice. The second contribution is that our multisignature scheme can provide sequentially accountability, which means that not only individual signers can be identified from the multisignatures, but also the order of accountability. The main technique of this paper is the joint public key composed of the public keys of a group of users, which has been used in self-certified signatures [15, 16] and joint encryption scheme [17] to achieve provably secure.

Moreover, if the joint public key of a group of signers is precomputed, the proposed multisignature scheme is optimal, since the size of multisignatures and the verification costs are the same as those for the single-signer Schnorr signature scheme, regardless of the number of signers.

# 2 The new multisignature scheme

In this section, we first present a formal definition for the multisignature scheme, and then provide an implementation of Multisignature Scheme based on Discrete Logarithms (MSDL). Let $U = \{U_1, U_2, \ldots, U_n\}$ be a group of $n$ signers.

## 2.1 Definition for multisignature scheme

**Definition 1**. A multisignature scheme is specified as four randomized algorithms: ParaGen, KeyGen, Sign and Verify:

**ParaGen**: takes a security parameter $1^k$ as input and returns a system parameter $P$, including some cryptographic hash functions.

**KeyGen**: takes the system parameter $P$ as input, each signer $U_i$ of the group $U$ chooses its keypair $(x_i, y_i)$ respectively.

**Sign**: takes as input $P$, the signers of any subset $S$ of the group $U$ (without loss of generality $S = \{U_1, U_2, \ldots, U_t\}$) cooperatively generate a multisignature $\sigma$ for a message $M$ by using their keypairs $(x_i, y_i)$. The joint public key $Y_S$ of the subset $S$ is composed of the individual public keys $\{y_1, \ldots, y_t\}$.

**Verify**: takes as input $P$, $M$, the joint public key $Y_S$ and a multisignature $\sigma$, it returns *invalid* or *valid*, with the property that if $(P) \leftarrow$ **ParaGen**$(1^k)$, $(\{x_1, \ldots, x_t\}, Y_S) \leftarrow$ **KeyGen**$(P)$ and $\sigma \leftarrow$ **Sign**$(P, M, \{x_1, \ldots, x_t\}, Y_S)$, then **Verify**$(P, M, \sigma, Y_S) = valid$.

## 2.2 An Implementation of Multisignature Scheme based on Discrete Logarithms (MSDL)

We use the Schnorr signature [18] as the underlying signature, which has been proved to be secure in the random oracle model [19].

**ParaGen:** A trusted party takes a security parameter $1^k$ as input and returns the system parameter $P$, which includes a subgroup $G_{g,p} = \{g^0, g^1, \ldots, g^{q-1}\}$ of a prime order $q$ in the multiplicative group $Z_p^*$, where $g$ is a generator with the prime order $q$, and two (ideal) hash functions $H$ and $F$, where

$$H: G_{g,p} \times \ldots \times G_{g,p} \rightarrow Z_q^* \text{ and } F: \{0,1\}^* \times Z_p^* \times Z_q^* \rightarrow Z_q^*$$

**KeyGen**: takes the system parameter $P$ as input, each signer $U_i$ of a group $U$ chooses its private key $x_i \in Z_q^*$ and computes its public key $y_i = g^{x_i}$ respectively.

**Sign**: takes as input $P$, the signers of a subset $S$ of the group $U$ (without loss of generality $S = \{U_1, U_2, \ldots, U_t\}$) cooperatively generate a multisignature $\sigma$ for a message $M$ by using their keypairs $(x_i, y_i)$ as follows:

1. Each signer $U_i$ of the subset $S$ chooses a random number $k_i \in Z_q^*$, computes $r_i = g^{k_i}$ and broadcasts $(y_i, r_i)$.
2. After receiving $(y_j, r_j)$, $(j = 1, 2, \ldots, t)$, each signer

computes $R = r_1r_2\ldots r_t$, $h = H(y_1, y_2,\ldots, y_t)$ and $f = F(M, R, h)$.

3. The first signer $U_1$ computes $s_1 = k_1 - fx_1 \pmod q$ and sends it to the second signer $U_2$.
4. After receiving $s_1$ from the first signer $U_1$, the second signer $U_2$ first verifies $g^{s_1} y_1^{f} = r_1$ and then computes $s_2 = s_1 + k_2 - hfx_2 \pmod q$, and sends it to the third signer $U_3$.
5. After receiving $s_{t-1}$ from the $(t-1)$th signer $U_{t-1}$, the last signer $U_t$ first verifies $g^{s_{t-1}}(y_1 y_2^{h}\ldots y_{t-1}^{h^{t-2}})^{f} = r_1r_2\ldots r_{t-1}$, and then computes $s = s_{t-1} + k_t - h^{t-1}fx_t \pmod q$. The multisignature for the message $M$ is $\sigma = (f, s)$.

**Verify**: The verifier first computes the joint public key of the subset $Y_S = y_1 y_2^{h}\ldots y_t^{h^{t-1}}$, where $h = H(y_1, y_2,\ldots, y_t)$, and then checks the verification equation of the multisignature $f = F(M, g^{s} Y_S^{f}, H(y_1, y_2,\ldots, y_t))$.

*Completeness*: Because $s_1 = k_1 - fx_1 \pmod q$ implies $g^{s_1} y_1^{f} = r_1$, $s_2 = s_1 + k_2 - hfx_2 \pmod q$ implies $g^{s_1}(y_1 y_2^{h})^{f} = r_1r_2$. By the same reason, $g^{s_{t-1}}(y_1 y_2^{h}\ldots y_{t-1}^{h^{t-2}})^{f} = r_1r_2\ldots r_{t-1}$ and $s = s_{t-1} + k_t - h^{t-1}fx_t \pmod q$ imply the verification equation. Hence, the signature $\sigma = (s, f)$ produced by the algorithm **Sign** is always *valid*.

Notice that our results can also be carried over to other groups, such as those built on elliptic curves.

Notice that the algorithm **Verify** requires that a verifier computes the joint public key $Y_S = y_1 y_2^{h}\ldots y_t^{h^{t-1}}$ from the individual public keys $\langle y_1, y_2,\ldots, y_t\rangle$ of a subset of signers. However, this time-consuming computation is independent of messages to be signed, and hence can be done once for all. Once the joint public key $Y_S$ of a subset of signers is precomputed, the performance of the multisignature scheme is optimal.

# 3 Security model and security proof

In this section, we first define a new security model for multisignature schemes, which gives the adversary complete freedom in specifying the public keys of the corrupted signers without handing over the corresponding private keys. Then we provide the security proof in this strong security model.

## 3.1 Security model for multisignature schemes

*Security model*

Existential unforgeability against adaptive chosen message attacks (EUF-CMA) [20] is the well-accepted security model for signature schemes, where the adversary is allowed to ask the challenger to sign any message of its choice adaptively, i.e. he can adapt its queries according to previous answers. Finally, the adversary could not provide a new message-signature pair with a non-negligible advantage. Hence, it is natural to require that the multisignatures also satisfy this strong security notion.

Accordingly, existential unforgeability for group oriented setting means that the adversary attempts to generate a new multisignature without the knowledge of all private keys. We formalize this intuition as a chosen key model. In this model, the adversary is given a single public key, while the adversary is allowed to choose the key pairs of other signers of the group, and to ask the sign query for any multisignature under any joint public key. His goal is to generate a new multisignature under the joint public keys of the group composed of the given public key and the public keys of its choice.

We say that a multisignature scheme is existential unforgeable against adaptive chosen message attacks if no polynomial bounded adversary $A$ has a non-negligible advantage against the challenger in the following game:

**Setup**: The challenger takes a security parameter $1^k$ as input and runs the randomized system parameter generation algorithm and the key generation algorithm to generate the system parameter $P$ and a public key $y$. The challenger gives them to the adversary $A$.

**Queries**: Processing adaptively, the adversary $A$ requests multisignatures queries $(M_i, Y_i)$ under any joint public key $Y_i$ on any message $M_i$ of its choice, where the challenged public key $y$ might be included in the joint public key $Y_i$.

**Response**: Finally, the adversary $A$ outputs a new multisignature $\sigma$ for a message $M$ under a joint public key $Y$.

The adversary $A$ wins the game if the output multisignature $\sigma$ is nontrivial, i.e. it is not an answer of a sign query $(M, Y)$ for the message $M$ under a joint public key $Y$, and the joint public key $Y$ is composed of the individual public keys $\{y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t\}$, where the individual public keys $\{y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_t\}$, $(j \in \{1, \ldots, t\})$, are chosen by the adversary $A$ and $y$ is the given public key.

The probability is over the random bits used by the challenger and the adversary.

Notice that our security model does not suffer from the same special limitation as the multisignature schemes proposed before. The adversary is given complete freedom in specifying the public keys but the given public key and is not enforced to disclose any knowledge of the corresponding private keys.

Notice that the Schnorr signature generation is not deterministic, there may be several signatures corresponding to a given message. Hence, our security model actually adopts the more liberal rule, which makes the adversary successful when it outputs a fresh signature of a given message different from previously obtained

signatures of the same message. Thus, our security model achieves non-malleability (NM) [21].

## 3.2 Security proof of the MSDL scheme

The security of the proposed MSDL scheme is based on the DL assumption.

**Definition** 2 (DL assumption)

A probabilistic algorithm $A$ is said to $(t, \varepsilon)$-break DL in a group $G_{g,p}$, if on input $(g, p, q, y = g^x)$ and after running in time at most $t$, $A$ solves the discrete logarithm problem $x = log_{g,p}y$ with probability at least $\varepsilon$, where the probability is over the uniform random choice of $g$ from the group $G_{g,p}$, of $x$ from $Z_q^*$, and the coin tosses of $A$. The $(t, \varepsilon)$-DL assumption on the group $G_{g,p}$ is that if no algorithm $(t, \varepsilon)$-breaks DL in $G_{g,p}$.

We have the following theorem about the security of the MSDL scheme.

**Theorem**. Let the hash functions $H$, $F$ be random oracles. Then the Multisignature Signature scheme based on DL is existentially unforgeable against adaptive chosen message attacks (EUF-CMA) under the DL assumption.

Concretely, suppose that there is a EUF-CMA adversary $A$, which has an advantage $\varepsilon^{MSDL}$ against the MSDL scheme of $t$ signers and $A$ runs in time at most $t^{MSDL}$. Suppose that $A$ makes at most $q_S$ sign queries, and at most $q_H$, $q_F$ queries to the hash functions $H$, $F$, respectively. Then there is a DL algorithm $B$ that has an advantage $\varepsilon^{DL}$ in $G_{g,p}$ with running time $t^{DL}$, where:

$$\varepsilon^{MSDL} \le (4q_Fq_H)(\varepsilon^{DL})^{1/(3t+1)} + 1/q + q_S(q_F + q_S)/p \quad (1)$$

$$t^{MSDL} \ge t^{DL}/(2t) - 2q_SC_{exp}(G_{g,p}) \quad (2)$$

Here $C_{exp}(G_{g,p})$ denotes the computation cost of a long exponentiation in the group $G_{g,p}$.

*Proof*: We use the random oracle model to show that the proposed multisignature scheme is secure. Concretely, suppose that there is a EUF-CMA adversary $A$ that has an advantage $\varepsilon^{MSDL}$ against the MSDL scheme and $A$ runs in time at most $t^{MSDL}$. Suppose that $A$ makes at most $q_H$, $q_F$ queries to the hash functions $H$ and $F$ respectively, and at most $q_S$ queries to the sign oracle. Then there is a DL algorithm $B$ that has an advantage $\varepsilon^{DL}$ in $G_{g,p}$ with running time $t^{DL}$.

We show how to construct a DL algorithm $B$ that uses $A$ as a computer program to gain an advantage $\varepsilon^{DL}$ for a DL problem with running time $t^{DL}$. The challenger takes a security parameter $1^k$ and runs the system parameter generation algorithm and the key generation algorithm to obtain the group $G_{p,g}$ and $y$. Its goal to output $x = \log_{,g,p}y \in Z_q^*$.

Algorithm $B$ simulates the challenger and interacts with the adversary $A$ in the following attack games:

Algorithm $B$ gives the adversary $A$ the resulting system parameter $P$, and $y$ as the public key of an honest signer.

At any time, the adversary $A$ can query hash oracles $H$ or $F$. To response to these queries, $B$ maintains two lists of tuples for the hash oracles $H$ and $F$ respectively. We refer to these lists as $H$-list and $F$-list. The contents of the two lists are "dynamic" during the attack games. Namely, when the games start, they are initially empty, but at the end of the games, they record all pairs of queries/answers.

**Answering *H*-oracle Queries**. When $A$ queries the oracle $H$ with some message $<y_1, y_2, \ldots, y_t>$, algorithm $B$ responds as follows:

1. If the query $<y_1, y_2, \ldots, y_t>$ already appears on the $H$-list in some tuple $<<y_1, y_2, \ldots, y_t>, h>$, then algorithm $B$ responds with $h = H(y_1, y_2, \ldots, y_t)$.
2. Otherwise, algorithm $B$ picks a random $h$ in $Z_q^*$, and responds with $h = H(y_1, y_2, \ldots, y_t)$ and adds the tuple $<< y_1, y_2, \ldots, y_t >, h>$ to the $H$-list.

**Answering *F*-oracle Queries**. When $A$ queries the oracle $F$ with some message $<M, R, h>$, algorithm $B$ responds as follows:

1. If the query $<M, R, h>$ already appears on the $F$-list in some tuple $<<M, R, h>, f>$, then algorithm $B$ responds with $f = F(M, r, h)$.
2. Otherwise, $B$ checks if $h$ is in the $H$-list, then generates a random $f \in Z_q^*$, responds with $f = F(M, R, h)$, and adds the tuple $<<M, R, h >, f>$ to the $F$-list.

Obviously, in two ways, $h$ and $f$ are uniform in $Z_q^*$, and they are independent of $A$'s current view as required.

**Answering sign queries**. When the adversary $A$ requests a signature for $<M, Y>$ under a joint public key $Y$, algorithm $B$ responds to this query as follows:

1. $B$ checks if $Y$ is a valid joint public key: $Y = y_1y_2^{\,h}\ldots y_t^{\,h^{t-1}}$, where $h = H(y_1, y_2,\ldots, y_t)$.
2. Algorithm $B$ chooses at random $s$, $f \in Z_q^*$, and computes $R = g^sY^f$.
3. If there exists a tuple $<<M, R, h>, f'>$ in the $F$-list with $f \ne f'$, $B$ reports failure and terminates. (The probability of this unfortunate coincidence is at most $(q_F + q_S)/p$).
4. Otherwise, $B$ responds with $(s, f)$ to the adversary $A$, and adds the tuple $<<M, R, h>, f>$ to the $F$-list.

Obviously, the outputs of the simulated oracles are indistinguishable from those in the real attacks.

Finally, the adversary $A$ returns a new valid message $M$ and its multisignature $(s, f)$ under the joint public key $Y$ composed of public keys $\{y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t\}$, where $y$ is the challenged public key and others are chosen by the adversary $A$ such that

$$f = F(M, g^s(y_1\ldots y_{j-1}^{\,h^{j-2}}\, y^{h^{j-1}}\, y_{j+1}^{\,h^j}\ldots y_t^{\,h^{t-1}})^f, h),$$

where $h = H(y_1, y_2,\ldots, y_t)$

If the adversary $A$ has not queried $F(M, R, h)$ or $H(y_1, y_2,\ldots, y_t)$, the probability

$\Pr[f = F(M, g^s(y_1...y_{j-1}^{h^{j-2}} y^{h^{j-1}} y_{j+1}^{h^j}...y_t^{h^{t-1}})^f, h)$, where $h = H(y_1, y_2,\ldots, y_t)] \le 1/q$,

since both the responses $F(M, R, H(y_1, y_2,\ldots, y_t))$ and $H(y_1, y_2,\ldots, y_t)$ are picked randomly.

Hence, with the probability $(1- 1/q)(\varepsilon^{\mathrm{MSDL}} - q_S(q_F + q_S)/p)$ the adversary $A$ returns a new multisignature $(s, f)$ such that

$$f = F(M, g^s(y_1...y_{j-1}^{h^{j-2}} y^{h^{j-1}} y_{j+1}^{h^j}...y_t^{h^{t-1}})^f, h),$$

where $h = H(y_1, y_2,\ldots, y_t)$

and the responses $F(M, R, H(y_1, y_2,\ldots, y_t))$ and $H(y_1, y_2,\ldots, y_t)$ are in the $F$-list and the $H$-list.

The verification equation is equivalent to the equation

$$g^s(y_1...y_{j-1}^{h^{j-2}} y^{h^{j-1}} y_{j+1}^{h^j}...y_t^{h^{t-1}})^{F(M,R,h)} = R,$$

where $h = H(y_1, y_2,\ldots, y_t)$,

where $y$ is the challenged public key and other public keys $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_t$ are chosen by the adversary $A$.

Since Pointcheval and Stern proposed the forking reduction proof [19], oracle replay techniques have been used to provide formal security proofs for ElGamal–like triplet signature schemes. In this proof, we are required to find $x = \log_{g,p} y$. It is a knowledge extraction problem. Hence, we try to use the oracle replay techniques to solve this DL problem.

We use $2t$ copies of the adversary $A$. In the attack games, the adversary $A$ would ask $H$-query for $<y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t>$. We first guess a fixed index $1 \le u \le q_H$ and hope that $(y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t)$ happens to be $u$th $H$-query used in the forged multisignature. Then we guess a fixed index $1 \le v \le q_F$ and hope that $<M, R, h>$ happens to be $v$th $F$-query used in the forged multisignature. Note that $A$ must ask for $H(y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t)$ before for $F(M, R, h)$.

Suppose that we make two good guesses by chance, denoted by the event GoodGuess. The probability of the event GoodGuess is

$$\Pr[\mathrm{GoodGuess}] = 1/(q_H q_F).$$

Hence, with the probability

$$\varepsilon = (1 - 1/q)(\varepsilon^{\mathrm{MSDL}} - q_S(q_F + q_S)/p)/(q_F q_H)$$
$$\ge (\varepsilon^{\mathrm{MSDL}} - 1/q - q_S(q_F + q_S)/p)/(q_F q_H)$$

the adversary $A$ generates a new multisignature.

$B$ gives the same system parameter, the same public key $y$ and same sequence of random bits to the $2t$ copies of the adversary $A$, and responds with the same random answers to their queries for oracles until they at the same time ask the $H$-oracle query for $< y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t>$. This is the first forking point. At that point, $B$ gives $t$ independent random answers $h_1$, $h_2$ and $h_t$ to the hash queries $H$ in the $2t$ runs, the first two, gives $h_1$, the second two, gives $h_2$, and the last two, gives $h_t$.

Then $B$ gives the first two copies of the adversary $A$ same sequence of random bits, and the same random answers to their oracle queries until they both ask for $F(M_1, R_1, h_1)$. This is the second forking point. At that point, $B$ gives two independent random answers $f_{11}$ and $f_{12}$ to the hash queries $F(M_1, R_1, h_1)$ in the first two runs. Similarly, $B$ gives two independent random answers $f_{21}$ and $f_{22}$ to the hash queries $F(M_2, R_2, h_2)$ (the third forking point) in the second two runs, $f_{t1}$ and $f_{t2}$ to the hash queries $F(M_t, R_t, h_t)$ (the $(t + 1)$th forking point) in the last two runs. Thus, we would obtain $2t$ multisignatures, satisfying the following equations:

$$g^{s_{11}}(y_1 y_2^{h_1}...y_t^{h_1^{t-1}})^{f_{11}} = R_1 \quad (3)$$

$$g^{s_{12}}(y_1 y_2^{h_1}...y_t^{h_1^{t-1}})^{f_{12}} = R_1 \quad (4)$$

$$g^{s_{21}}(y_1 y_2^{h_2}...y_t^{h_2^{t-1}})^{f_{21}} = R_2$$

$$g^{s_{22}}(y_1 y_2^{h_2}...y_t^{h_2^{t-1}})^{f_{22}} = R_2$$

……

$$g^{s_{t1}}(y_1 y_2^{h_t}...y_t^{h_t^{t-1}})^{f_{t1}} = R_t$$

$$g^{s_{t2}}(y_1 y_2^{h_t}...y_t^{h_t^{t-1}})^{f_{t2}} = R_t$$

From these equations, we can derive $\log_{g,p} y$ as follows:

From eqn.(3) and eqn.(4), we can derive $a_1 = (s_{11} - s_{12})/(f_{12} - f_{11}) \pmod q$ such that

$$(y_1 y_2^{h_1}...y_t^{h_1^{t-1}}) = g^{a_1} \quad (t+1)$$

By the same way, we can derive $a_2, \ldots, a_t$, such that

$$(y_1 y_2^{h_2}...y_t^{h_2^{t-1}}) = g^{a_2} \quad (t+2)$$

……

$$(y_1 y_2^{h_t}...y_t^{h_t^{t-1}}) = g^{a_t} \quad (t+t)$$

Then from eqn.$(t+1)$– eqn.$(t+t)$, we have a system of equations

$$x_1 + h_1 x_2 + \ldots + h_1^{t-1} x_t = a_1 \pmod q$$

$$x_1 + h_2 x_2 + \ldots + h_2^{t-1} x_t = a_2 \pmod q$$

……

$$x_1 + h_t x_2 + \ldots + h_t^{t-1} x_t = a_t \pmod q$$

We can derive $x_j = \log_{g,p} y$ since $h_1$, $h_2$ and $h_t$ are different from each other.

We use the "splitting lemma" [19] to compute the probability that $A$ works as hoped. Let $X$ be the set of possible sequences of random bits and random function values that take the adversary up to the first forking point where $A$ asks for $H(y_1, \ldots, y_{j-1}, y, y_{j+1}, \ldots, y_t)$; let $Y_1$ be the set of possible sequences of random bits and random function values from the first forking point to the second forking point; let $Z_1$ be the set of possible sequences of random bits and random function values from the second forking point. By assumption, for any $x \in X$, $y \in Y_1$, $z \in Z_1$, the probability that $A$, supplied the sequences of random bits and random values $(x||y||z)$, generates a new multisignature is $\varepsilon$.

Suppose that the sequences of random bits and random function values supplied up to the first forking point in the simulations is $a$. By "splitting lemma", the probability that $\Pr\{a \in$ "good" subset $\Omega\} \ge \varepsilon/2$, and whenever $a \in \Omega$, $y \in Y_1$, $z \in Z_1$, the probability that $A$, supplied the sequences of random bits and random values $(a||y||z)$, produces a forgery is at least $\varepsilon/2$.

Suppose that the sequences of random bits and random function values supplied from the first forking point up to the second forking point in the simulations is $b$. Thus, the probability that $\Pr\{b \in$ "good" subset $\Omega'\} \geq \varepsilon/4$, and whenever $a \in \Omega$, $b \in \Omega'$, $z \in Z_1$, the probability that $A$, supplied the sequences of random bits and random values $(a||b||z)$, produces a forgery is at least $\varepsilon/4$.

By the same reason, we can compute the same probability for the other $t$ - 1 cases.

Hence the probability that $B$ solves the discrete logarithm in the $2t$ simulations is

$$\varepsilon^{\mathrm{DL}} \geq (\varepsilon)^{(3t+1)}/2^{(6t+1)} \geq ((\varepsilon^{\mathrm{MSDL}} - 1/q - q_S(q_F + q_S)/p)/(4q_Fq_H))^{(3t+1)}$$

$$\varepsilon^{\mathrm{MSDL}} \leq (4q_Fq_H)(\varepsilon^{\mathrm{DL}})^{1/(3t+1)} + 1/q + q_S(q_F + q_S)/p.$$

The time required to run one simulation is $t^{\mathrm{MSDL}} + 2q_SC_{exp}(G_{g,p})$.

The time required by the simulator $B$ to solve the discrete logarithm $\log_{g,p}y$ is

$$t^{\mathrm{DL}} \leq 2t(t^{\mathrm{MSDL}} + 2q_SC_{exp}(G_{g,p})).$$

$$t^{\mathrm{MSDL}} \geq t^{\mathrm{DL}}/(2t) - 2q_SC_{exp}(G_{g,p}).$$

Q.E.D.

# 4 Conclusion

We have proposed a Multisignature Scheme based on Discrete Logarithms (MSDL). We show that this new scheme can resist existential forgeries against adaptive chosen-message attacks in the random oracle model. The main contribution is that our security model gets rid of the special security requirement on the generation of the participants' public keys. Thus the new multisignature scheme does not suffer from the problem identified by Micali et al., which is shared by many current multisignature schemes.

The second contribution is that our multisignature scheme can provide sequentially accountability, which means that not only individual signers can be identified from the multisignature, but also the order of accountability. That is, the first signer $U_1$ is responsible for the first partial multisignature, the second signer $U_2$ is responsible for the second partial multisignature, and the last signer $U_t$ is responsible for the last multisignature. Thus, our scheme is robust. Notice that here sequentially accountability means that verifiers can demand that the signers are responsible for multisignatures according to the specified order $<U_1, U_2, \ldots, U_t>$ rather than that the signers could generate multisignatures only according to the specified order.

Furthermore, the proposed multisignature scheme is more efficient, since the size of multisignatures is the same as that of the underlying signatures, regardless of the number of participants. If the joint public key $Y$ of a group of signers is precomputed, the computation cost for verification a multisignature is identical to those of an individual's signature. Thus the proposed multisignature scheme is optimal.

However, the forking reduction proof we use makes our proof inefficient. Strictly speaking, our proof is only loosely related to the DL problem according to Micali and Reyzin [22]. Therefore, our multisignature scheme is only applicable to the group of polynomial bounded signers.

Although the Schnorr scheme provably secure by oracle replay technique is only loosely related to DL problem, there has been not any efficient forgery attack without solving DL problem first. By similar reasons, our more loosely reduction would also provide users with somewhat security confidence that there is no efficient forgery algorithm without solving DL problem first.

We have proposed a Multisignature Scheme based on Discrete Logarithms (MSDL). We show that this new scheme can resist existential forgeries against adaptive chosen-message attacks in the random oracle model. The main contribution is that our security model gets rid of the special security requirement on the generation of the participants' public keys. Thus the new multisignature scheme does not suffer from the problem identified by Micali et al., which is shared by many current multisignature schemes.

The second contribution is that our multisignature scheme can provide sequentially accountability, which means that not only individual signers can be identified from the multisignature, but also the order of accountability. That is, the first signer $U_1$ is responsible for the first partial multisignature, the second signer $U_2$ is responsible for the second partial multisignature, and the last signer $U_t$ is responsible for the last multisignature. Thus, our scheme is robust. Notice that here sequentially accountability means that verifiers can demand that the signers are responsible for multisignatures according to the specified order $<U_1, U_2, \ldots, U_t>$ rather than that the signers could generate multisignatures only according to the specified order.

Furthermore, the proposed multisignature scheme is more efficient, since the size of multisignatures is the same as that of the underlying signatures, regardless of the number of participants. If the joint public key $Y$ of a group of signers is precomputed, the computation cost for verification a multisignature is identical to those of an individual's signature. Thus the proposed multisignature scheme is optimal.

However, the forking reduction proof we use makes our proof inefficient. Strictly speaking, our proof is only loosely related to the DL problem according to Micali and Reyzin [22]. Therefore, our multisignature scheme is only applicable to the group of polynomial bounded signers.

Although the Schnorr scheme provably secure by oracle replay technique is only loosely related to DL problem, there has been not any efficient forgery attack without solving DL problem first. By similar reasons, our more loosely reduction would also provide users with somewhat security confidence that there is no efficient forgery algorithm without solving DL problem first.

## Acknowledgement

The author would like to thank the anonymous reviewers for their valuable comments and suggestions that improve the presentation of this paper significantly.

# References


[1] Y. Desmelt (1988). Society and group oriented cryptography: A new concept, *Advances in Cryptology-Crypto'87*, LNCS 293, Springer, Berlin, pp. 120-127.

[2] L. Harn (1999). Digital multisignature scheme with distinguished signing authorities, *Electron. Lett* ., 35(4), pp.294-295.

[3] K. Itakura and K. Nakamura (1983). A public key cryptosystem suitable for digital multisignatures, *NEC Research & Development*, (71): pp. 1- 8.

[4] K. Ohta and T. Okamoto (1999(. Multi-signature schemes secure against active insider attacks, *IEICE Transaction on Fundamentals of Electronics communications and computer Science*, E82-A(1), pp.21-31.

[5] S. Micali, K. Ohta, and L. Reyzin (2001). Accountable-subgroup mulitisignatures, *In ACM Conference on Computer and communications Security*, 2001.

[6] L. Harn (1994). Group-oriented (*t, n*) threshold digital signature scheme and digital multisignature, *IEE Proc.-Comput. Digit. Tech*., 141(5), pp.307-313.

[7] C.-M. Li, T. Hwang, and N.-Y. Lee (1994). Threshold-multisignature schemes where suspected forgery implies traceability of adversarial shareholders, *Advances in Cryptology – Eurocrypt 94*, LNCS 950, Springer-Verlag, pp. 194-204.

[8] A. Boldyreva (2003). Threshold signature, multisignature and blind signature schemes based on the gap-Diffie-Hellman-group signature scheme, *Proceedings of PKC 2003*, LNCS 2567, Springer-Verlag, pp. 31-46.

[9] S. Lu, R. Ostrovsky, A. Sahai, H. Shacham, and B. Waters (2006). Sequential Aggregate Signature and Multisignature without Random Oracle, In *EUROCRYPTO'06*, LNCS 4004, Springer-Verlag, Berlin, pp. 465-485.

[10] PKCS#10: Certification request syntax standard, RSA Data Security, Inc., 2000.

[11] C. Adams, S. Farrell, T. Kause, T. Monen (2005). Internet X.509 Public Key Infrastructure Certificate Management Protocol (CMP), Internet, Engineering Task Force RFC 4210.

[12] J. Schaad (2005). Internet X.509 Public Key Infrastructure Certificate, Request Message Format, Internet Engineering Task Force RFC, 4211.

[13] T. Ristenpart and S. Yilek (2007). The Power of Proofs-of-Possession: Securing Multiparty Signatures against Rogue-Key Attacks. *in Advances in Cryptology – EUROCRYPT 2007*, LNCS 4515, Springer-Verlag, pp. 228–245.

[14] M. Bellare and G. Neven (2006). Multi-signatures in the plain public-key model and a generalized forking lemma, CCS 2006, ACM, pp.390-399.

[15] Zuhua Shao (2007). Self-certified Signatures Based on Discrete Logarithms, *in Proceedings of WAIFI* 2007, LNCS 4547, Springer-Verlag, pp.252-263.

[16] Zuhua Shao (2007). Self-certified signature scheme from pairings, *Journal of Systems and Software*, 80(3), pp. 388-395.

[17] Zuhua Shao (2009). Dynamic and efficient joint encryption scheme in the plain public key model, *Computers and Electrical Engineering*, 35(1), pp. 189-196.

[18] C. P. Schnorr (1991). Efficient signature generation by smart cards, *Journal of Cryptology*, 3(3), pp.161-174.

[19] D. Pointcheval and J. Stern (2000). Security arguments for digital signatures and blind signatures, *Journal of Cryptology,* 13(3), pp. 361-396.

[20] S. Goldwasser, S. Micali, and R. Rivest (1988). A digital signature scheme secure against adaptive chosen-message attacks, *SIAM Journal on Computing*, 17(2), pp.281-308.

[21] M. Bellare, A. Desai, D. Pointcheval, and P. Rogaway (1998). Relation among notions of security for public-key encryption schemes, In *Crypto'98*, LNCS 1462, Springer-Verlag, Berlin, pp. 26-45.

[22] S. Micali and L. Reyzin (2002). Improving the exacting security of digital signature schemes, *Journal of Cryptology*, 15(1), pp.1-18.

# Ontology Extension Towards Analysis of Business News

Inna Novalija and Dunja Mladenić
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: {inna.koval, dunja.mladenic}@ijs.si
http://kt.ijs.si/

**Keywords:** ontology extension, text mining, Cyc, semantic web

**Received:** January 17, 2010

*This paper addresses the process of the ontology extension for a selected domain of interest which is defined by keywords and a glossary of relevant terms with descriptions. A new methodology for semi-automatic ontology extension, aggregating the elements of text mining and user-dialog approaches for ontology extension, is proposed and evaluated. We conduct a set of ranking, tagging and illustrative question answering experiments using Cyc ontology and business news collection. We evaluate the importance of using the textual content and structure of the ontology concept in the process of ontology extension. The experiments show that the best results are obtained with giving more to weight to ontology concept content and less weight to ontology concept structure.*

*Povzetek: Prispevek opisuje proces razširitve obstoječe ontologije konceptov.*

## 1 Introduction

This paper explores the process of the ontology extension motivated by usage of the extended ontology for business news analysis. The main contribution of this paper is in proposing a methodology for text-driven semi-automatic ontology extension using ontology content and ontology structure information. Our research also contributes to the analysis of business news by the means of semantic technologies. The new methodology for the semi-automatic ontology extension, aggregating the elements of text mining and user-dialog approaches for ontology extension, is suggested and used for inserting the new financial knowledge into Cyc [1], which maintains one of the most extensive common-sense knowledge bases worldwide.

As the ontology content of a particular concept we consider the available textual representation of the referred concept. The ontology content includes a natural language concept denotation (such as a concept label) and textual comments about the concept. As the ontology structure of a particular concept we consider the neighborhood concepts involved in the hierarchical and non-hierarchical relations with a referred concept. Ontology extension in this paper stands for: adding new concepts to the existing ontology or, augmentation of the existing textual representation of the relevant concepts with new available textual information – extension of the concept comments, changing or adding concept denotation.

The experiments on ranking, business news tagging and simple question answering show that the extended financial ontology allows for a better financial news analysis.

The evaluation of the methodology of the ontology extension shows its ability to fasten the ontology extension process.

The paper is structured as follows: Section 2 presents the information about the existing approaches of ontology extension; the new methodology of ontology extension is discussed in Section 3, Sections 4 describes the experiments and the results, the conclusion is covered in Section 5.

## 2 Existing approaches to ontology extension

The automatic and semi-automatic ontology extension processes are usually composed of several phases. Most approaches include defining the set of the relevant ontology extension sources, pre-processing the input material, ontology augmentation according to the chosen methodology and ontology evaluating and revision phases. The notable approaches of ontology extension include natural language processing based approach [2, 3], networks/graphs based approach [4, 5] and user interaction approach [6, 7].

The linguistic patterns are used by the authors of Text2Onto [7] framework for ontology learning and SPRAT [8] tool for ontology population.

Several methods of the automatic ontology extension operate with enlarging of Cyc Knowledge Base (Cyc KB). The automated population of Cyc with named entities involves the Web and a framework for validating candidate facts [9]. The semi-automatic approach for Cyc KB extension presented in [6] is based on the user-interactive dialogue system for knowledge acquisition, where, the user is engaged in a natural-language mixed-initiative dialogue. The system contains a natural language generation module, parsing module, post-processing module, dictionary assistant, user interaction agenda and salient descriptor. Medelyan and Legg [10]

describe the methodology for integrating Cyc and Wikipedia, where the concepts from Cyc are mapped onto Wikipedia articles describing correspondent concepts. Sarjant et al. [11] use Medelyan and Legg [10] method to augment Cyc ontology using pattern matching and link analysis.

In the presented research we are using a combination of top-down and bottom-up approaches to the ontology extension and apply it on Cyc ontology. The top-down part involves the user identifying the keywords for extracting relevant data from the ontology, while the bottom-up part involves automatic obtaining of the relevant information available in the ontology. Usage of text mining methods involves data preprocessing, where a chain of linguistic components, such as tokenization, stop-word removal and stemming allows normalizing the textual representation of ontology concepts and a domain relevant glossary of terms with descriptions. Text mining methods are further used for automatically determining candidate concepts in the ontology to relate to the new knowledge from the domain. A list of suggestions is provided to the user for a final decision which allows preventing the inappropriate insertions into the ontology.

# 3 Methodology

As a part of the research, we propose a new methodology for semi-automatic ontology extension, which combines text mining methods with user-oriented approach and supports the extension of multi-domain ontologies.

The proposed methodology for semi-automatic ontology extension accounts for the following phases:

1. *Domain information identification.* The user identifies the appropriate domain keywords. As well, in this module a domain relevant glossary, containing terms with descriptions is determined. We assume that the glossary terms are the candidate entry concepts for the existing ontology. Consequently, the glossary terms might be in the following relationships with the existing ontology concepts:
   - Equivalence relationship: candidate concept represented by a glossary term is equivalent to the existing ontology concept;
   - Hierarchical relationship: candidate concept represented by a glossary term is in the superclass-subclass relationship with an existing ontology concept;
   - Non-hierarchical relationship: candidate concept represented by a glossary term is the in the associative relationship with an existing ontology term. The nature of the relationship is not hierarchical;
   - No relationship: candidate concept represented by a glossary term is not related to the existing ontology concept.
2. *Extraction of the relevant domain ontology subset from multi-domain ontology.* In case of large common-sense ontologies, such as Cyc Knowledge Base, the user entering new knowledge very often needs a particular ontology subset of his domain interest. Therefore, the domain keywords are mapped to the natural language representation of the ontology domain information and a set of the relevant domains of interest is identified. Further, ontology concepts defined in these domains are extracted. By concept extracting we mean obtaining the content and structure of the ontology concept. Correspondently, we find the textual representation (natural language denotation and comments) as content for the particular ontology concept. The ontology structure of the particular concept is represented by the natural language denotations of the hierarchically and non-hierarchically connected ontology concepts. Besides that, the names of the glossary terms are mapped to the natural language denotations of the concepts from other domains and the correspondent concepts are also extracted.
3. *Domain relevant information preprocessing.* The preprocessing phase includes tokenization, stop-word removal and stemming. Textual information is represented using bag-of-words representation with TFIDF weighting and similarity between two text segments is calculated using cosine similarity between their bag-of-words representations, as commonly used in text mining [12]. For each term from the domain relevant glossary we compose bag-of-words aggregating preprocessed textual information from: (1) the glossary term name and (2) the term comment. For each concept from the extracted relevant ontology subset the following information is considered: (1) the ontology concept content consisting of the preprocessed natural language concept denotation and concept comment; (2) the ontology concept structure consisting of the preprocessed natural language concept denotation and natural language denotations of hierarchically and non-hierarchically related concepts.
4. *Composing the list of potential concepts and relationships for ontology extension.* The ranked list of the relevant concepts and possible relationships suitable for ontology extension is composed. Similarity $(SIM_{cont})$ between glossary term and ontology concept content is calculated and weighted with weight $\delta$ $(0 \leq \delta \leq 1)$ defined by the user. Similarity $(SIM_{str})$ between glossary term and ontology concept structure is calculated and weighted with weight $1-\delta$. The combined content and structure similarity $(SIM)$ is used to rank ontology concepts for each glossary term:

$$SIM = \delta * SIM_{cont} + (1-\delta) * SIM_{str} \qquad (1)$$

Ontology concepts with similarity $(SIM_c)$ larger than $SIM_{max}*(1-\beta)$ are suggested to the user, where $SIM_{max}$ represents the highest similarity value between ontology concept and a glossary term for a particular glossary term and $\beta$ is defined by the user $(0 \leq \beta \leq 1)$:

$$SIM_c > SIM_{max}*(1-\beta) \qquad (2)$$

5. *User validation.* Furthermore, the user validates the candidate entries results consisting of the glossary terms and relevant existing ontology concepts. In case of the equivalence relationship the user can extend the textual representation of the existing ontology concept by adding comment, adding or changing the natural language denotation. In case of the hierarchical

relationships the user can add subclasses to the existing ontology concepts. If the nature of the relationship is not clear, the user can create an associative relationship or choose any other relationship between a glossary term and existing ontology concept. Moreover, the list with validated entries in the relevant format is created.

6. *Ontology extension* represents adding the new concepts and relationships between concepts into the ontology.

7. *Ontology reuse.* The ontology reuse phase serves as the connection link between separate ontology extension processes.

We have adapted the methodology in order to obtain an exhaustive specific methodology for Cyc knowledge base extension. The main adaptations are based on microtheories (Mt) that Cyc is using to represent thematic subsets of the ontology. Since our motivation is in business news annotation, we have chosen Business and Finances as the domains of primary interest. Given the fact that Cyc Knowledge Base contains common sense knowledge [13], we assume that Cyc KB includes some financial knowledge – financial knowledge base (Cyc FKB).

# 4 Evaluation

## 4.1 Experiments

In order to evaluate the proposed methodology we conducted a series of ranking, news tagging and illustrative question answering experiments on the data sources, described below.

*TERM: ASSETS*
*COMMENT: A firm's productive resources.*

*TERM: INFLATION*
*COMMENT: The rate at which the general level of prices for goods and services is rising.*

*TERM: COOPERATIVE*
*COMMENT: An organization owned by its members. Examples are agriculture cooperatives that assist farmers in selling their products more efficiently and apartment buildings owned by the residents who have full control of the property.*

Figure 1: *Example Financial Glossary Entries*

For the data evaluation we have used the RSS feeds data Yahoo! Finance [14] website. The news collection used in the current research accounts for 5812 Yahoo! Finance news.

Following the first phase of the proposed methodology, domain knowledge identification should be made in the initial phase. For these purposes we have selected the Harvey [15] financial glossary which contains around 6000 hyperlinked financial terms.

Figure 1 contains the examples of the typical financial glossary entries.

Tagging experiments give a background for Cyc FKB extension displaying the level of the financial domain representation in Cyc Knowledge Base. We have used a random subset of 100 Yahoo! Finance news to identify the financial terms, occurring most frequently in the selected news, tagged the terms with Cyc Concept Tagger and checked the precision and recall of news tagging.

For the methodology evaluation, we have conducted ranking experiments on the subset of 500 random Yahoo! Finance news. The most frequent financial terms have been extracted and 100 random financial terms have been chosen. Cyc Financial Knowledge Base is then extended, using the proposed methodology, with concepts corresponding to the chosen financial terms. The efficiency of the automatic concept ranking and the importance of the ontology content and ontology structure in the ontology extension process are measured afterwards.

Illustrative question answering demonstrates the capacity of Cyc to answer simple financial questions before and after the extension of Cyc Financial Knowledge Base. Let us assume that we have a simple question and we want to get an answer using an unextended and extended Cyc Knowledge Base.

## 4.2 Results

The results of the experiments suggest that the financial ontology extension leads to better business news analysis and confirm the applicability of the suggested methodology for ontology extension to Cyc Knowledge Base augmentation.

We have found 231 financial terms in the random sample of 100 Yahoo! Finance news. The precision and recall of business news tagging with Cyc Concept Tagger accounted for 61% and 46% correspondently. This confirms our hypothesis that the Cyc ontology has still space for extension in the financial domain with terms that are relevant for financial news analysis.

Table 1 shows the quality of automatic concept ranking when using different proportions of ontology concept textual content and ontology concept structure for ranking of the related concepts. We have manually evaluated the automatically suggested ranked related Cyc concepts for every glossary term estimating the proportion of correctly suggested terms among the top 1 suggested terms.

Table 1: *Content and Structure Weighting Measures (Financial Glossary).*

| Weighting Measure | 100 Random Terms | | |
|---|---|---|---|
| | Top 1 Eqv & Hier Rels | Top 1 Assoc Rels | Top 1 All Rels |
| Names/Denotation [100%] | 18 | 10 | 28 |
| Content [0%] Structure [100%] | 31 | 30 | 61 |
| Content [10%] Structure [90%] | 32 | 30 | 62 |
| Content [20%] Structure [80%] | 29 | 31 | 60 |
| Content [30%] Structure [70%] | 30 | 31 | 61 |
| Content [40%] Structure [60%] | 35 | 33 | 68 |

| | | | |
|---|---|---|---|
| Content [50%] Structure [50%] | 35 | 36 | 71 |
| Content [60%] Structure [40%] | 35 | **37** | **72** |
| Content [70%] Structure [30%] | **36** | 35 | 71 |
| Content [80%] Structure [20%] | **36** | 34 | 70 |
| Content [90%] Structure [10%] | 35 | 33 | 68 |
| Content [100%] Structure [0%] | 32 | 33 | 65 |

For this evaluation we explore equivalent, hierarchical and associative relationships between glossary terms and the related Cyc concepts. The best performing proportions are obtained with giving more weight to the similarity calculated between glossary textual representation and Cyc concept content and less weight to the cosine similarity calculated between glossary textual representation and Cyc concept neighborhood. From the row Content [70%] Structure [30%] it is possible to notice that for 36 glossary terms the correct equivalently and hierarchically related Cyc concepts have been found among top 1 suggested concepts. For 71 glossary terms with this weighting measure any related terms have obtained the highest rank among the suggested related concepts.

It means that using the proposed methodology the user is able to compare Cyc and glossary concepts and establish the equivalent, hierarchical and other relations much faster than just using the manual search for the relevant concepts in Cyc.

The following example illustrates the relevance of the proposed Cyc ontology extension for question answering in the financial domain.

For the research purposes we have selected the following simple questions:

*What phase of the business cycle was Egypt in 2008?*
*Was Indonesia in contraction in 2008?*

*TERM: BUSINESS CYCLE*
*COMMENT: Repetitive cycles of economic expansion and recession. The official peaks and troughs of the U.S. cycle are determined by the National Bureau of Economic Research in Cambridge, MA.*

*Phases of Business Cycle:*

*TERM: CONTRACTION*
*COMMENT: A slowdown in the pace of economic activity.*

*TERM: TROUGH*
*COMMENT: The lower turning point of a business cycle, where a contraction turns into an expansion.*

*TERM: EXPANSION*
*COMMENT: A speedup in the pace of economic activity.*

*TERM: PEAK*
*COMMENT: The upper turning of a business cycle.*

Figure 2: *Business Cycle Definition.*

Using an unextended Cyc KB we get no appropriate answers because of the insufficient representation of business cycles in Cyc.

Figure 2 presents the textual definition of business cycle and its phases which we use to implement the notion of business cycles in Cyc.

Using the proposed methodology for semi-automatic ontology extension, we obtain a ranked list of related Cyc concepts for the correspondent glossary term (see Table 2).

Table 2: *Related Cyc Concepts for Glossary Term "Business Cycle".*

| Glossary Term | Ranked Related Cyc Concepts |
|---|---|
| BUSINESS CYCLE | Cycle-Situation<br>Recession-Economic<br>MacroeconomicEvent<br>Trough (a type of FluidReservoir) |

To enter new assertions into Cyc KB we use KE text format which facilitates the knowledge entry process. We select the Cyc concept *Cycle-Situation* as a superclass for glossary term *Business Cycle:*

```
KE text:
Constant: BusinessCycle.
In Mt: UniversalVocabularyMt.
isa: TemporalObjectType.
genls: Cycle-Situation.
comment: "Repetitive cycles of economic
expansion and recession. The official peaks and
troughs of the U.S. cycle are determined by the
National Bureau of Economic Research in
Cambridge, MA.".
```

Furthermore, we create a set of business cycle phases (*Contraction*, *Expansion*, *Peak* and *Trough*) as subclasses for Cyc concept *MacroeconomicEvent*. The following code displays the example of the *Contraction* phase definition:

```
KE text:
Constant: ContractionBusinessCyclePhase.
In Mt: UniversalVocabularyMt.
isa: TemporalObjectType.
genls: MacroeconomicEvent.
comment: "A slowdown in the pace of economic
activity".

In Mt: UniversalVocabularyMt.
f:(relationAllExists properSubSituations
BusinessCycle ContractionBusinessCyclePhase).
```

In addition, we create a predicate used for answering questions connected to business cycle phases of the specific countries.

```
KE text:
Constant: economyInBusinessCyclePhase.
In Mt: UniversalVocabularyMt.
isa: TernaryPredicate.
arity: 3.
arg1Isa: GeopoliticalEntity.
arg2Isa: TemporalThing.
arg3Isa: MacroeconomicEvent.
```

For the illustrative question answering example we estimate the business cycle phases by using the GDP

growth rate - the percentage increase or decrease of Gross Domestic Product (GDP) from the previous measurement cycle. We identify that a term *GDP* is already implemented in Cyc KB as *grossDomesticProduct*.

The following rule defines the conditions of being in the contraction business cycle phase for the particular country in the specified year. We assume that the contraction phase occurs when the real growth rate of GDP in the referred year $GR\ (GDP)_{Yn}$ decreases comparatively to the real growth rate of GDP in the previous year $GR\ (GDP)_{Yn-1}$ but is still higher than the real growth rate of GDP in the following year $GR\ (GDP)_{Yn+1}$:

$$GR\ (GDP)_{Yn-1} > GR\ (GDP)_{Yn} > GR\ (GDP)_{Yn+1} \quad (3)$$

```
KE text:
In Mt: UniversalVocabularyMt.
f:
(implies
(and
(evaluate ?SUCCESSOR1 (PlusFn ?Y 1))
(evaluate ?PREDECESSOR1 (DifferenceFn ?Y 1))
(evaluate ?PREDECESSOR2 (DifferenceFn
?PREDECESSOR1 1))
(grossDomesticProduct ?X (YearFn ?SUCCESSOR1)
(BillionDollars ?S1GDP))
(grossDomesticProduct ?X (YearFn ?PREDECESSOR1)
(BillionDollars ?P1GDP))
(grossDomesticProduct ?X (YearFn ?PREDECESSOR2)
(BillionDollars ?P2GDP))
(grossDomesticProduct ?X (YearFn ?Y)
(BillionDollars ?YGDP))
(evaluate ?S1GR (QuotientFn ?S1GDP ?YGDP))
(evaluate ?YGR (QuotientFn ?YGDP ?P1GDP))
(evaluate ?P1GR (QuotientFn ?P1GDP ?P2GDP))
(greaterThan ?P1GR ?YGR)
(greaterThan ?YGR ?S1GR)
(isa ?PHASE ContractionBusinessCyclePhase)
(dateOfEvent ?PHASE (YearFn ?Y)))
(economyInBusinessCyclePhase ?X (YearFn ?Y)
?PHASE)).
```

| Country | GDP Growth Rate | Year est. |
|---|---|---|
| Egypt | 7.1% | 2007 |
| Egypt | 7.2% | 2008 |
| Egypt | 4.5% | 2009 |
| Indonesia | 6.3% | 2007 |
| Indonesia | 6.1% | 2008 |
| Indonesia | 4.4% | 2009 |

The expansion, peak and trough phases occur under the following conditions:

*Expansion*:

$$GR\ (GDP)_{Yn-1} < GR\ (GDP)_{Yn} < GR\ (GDP)_{Yn+1} \quad (4)$$

*Peak*:

$$GR\ (GDP)_{Yn-1} < GR\ (GDP)_{Yn} > GR\ (GDP)_{Yn+1} \quad (5)$$

*Trough*:

$$GR\ (GDP)_{Yn-1} > GR\ (GDP)_{Yn} < GR\ (GDP)_{Yn+1} \quad (6)$$

For question answering the information from Cyc KB about the GDP levels of Egypt and Indonesia in 2006-2009 is used:

```
Cyc KB assertions:
 (grossDomesticProduct Egypt(YearFn 2009)
       (BillionDollars 470.4))
(grossDomesticProduct Egypt(YearFn 2008)
       (BillionDollars 450.1))
(grossDomesticProduct Egypt(YearFn 2007)
       (BillionDollars 419.9))
(grossDomesticProduct Egypt(YearFn 2006)
       (BillionDollars 392.1))

(grossDomesticProduct Indonesia-TheNation
       (YearFn 2009)(BillionDollars 968.5))
(grossDomesticProduct Indonesia-TheNation
       (YearFn 2008)(BillionDollars 927.7))
(grossDomesticProduct Indonesia-TheNation
       (YearFn 2007)(BillionDollars 874.4))
(grossDomesticProduct Indonesia-TheNation
       (YearFn 2006)(BillionDollars 822.6))
```

After extending Cyc KB with notion of business cycle and business cycle phases, using the information about GDP from Cyc KB, it is possible to get answers for the previously asked questions:

```
Query:
(economyInBusinessCyclePhase Egypt
(YearFn 2008) ?PHASE)
Query result:
*[Explain]     PeakBCPhase2008

Query:
(economyInBusinessCyclePhase
Indonesia-TheNation (YearFn 2008)
ContractionBCPhase2008)
Query result:
Query was proven True *[Explain]
```

According to the rules introduced into Cyc KB, Egypt was in the peak business cycle phase and Indonesia was in the contraction phase of the business cycle in 2008. `PeakBCPhase2008` and `ContractionBCPhase2008` are the correspondent instances of `PeakBusinessCyclePhase` and `ContractionBusinessCyclePhase` Cyc collections.

The results obtained in the illustrative question answering experiment are comparable with GDP growth rates in Egypt and Indonesia in 2007-2009 [16].

Table 3: *GDP Growth Rates in Egypt and Indonesia.*

Extension of Cyc Knowledge Base according to the proposed methodology allows the user to provide Cyc with new concepts and rules and perform a better question answering based on the extended ontology.

# 5 Conclusion

In this paper the aspects of ontology extension and business news analysis have been explored. The new methodology of ontology extension, combining text mining methods and user-based approach, has been proposed and exposed to the preliminary evaluation.

The evaluation of our methodology has been accomplished in the financial domain. We have tested the importance of using concept textual content and concept structure in the process of ontology extension. The best results are obtained with giving more weight to ontology concept content and less weight to ontology concept structure. In addition, we have illustrated the increase in the effectiveness of simple question answering after Cyc Knowledge Base extension with terms from Harvey [15] financial glossary.

In contrast with many other methodologies for ontology extension, our methodology deals with ontologies and knowledge bases, covering more than one domain. However, it allows restricting the area of ontology extension to a specific domain.

Unlike the developers of Text2Onto [7] and SPRAT [8] tools, we do not use lexico-syntactic patterns for the related concepts identification. The statistically driven techniques used in our methodology make the ontology extension process more language independent.

Furthermore, the user validation helps to avoid adding to the ontology irrelevant concepts and relationships.

The future work should include further extension of Cyc Knowledge Base and using it for more sophisticated news analysis. Furthermore, the proposed methodology for ontology extension should be tested on other domains. In addition, a particular attention should be given to the problem of the disambiguation of the glossary terms and terms extracted from news sources.

## Acknowledgement

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under NeOn (IST-4-027595-IP) and ACTIVE (IST-2008-215040).

# References

[1] Cycorp, Inc., *http://www.cyc.com*

[2] F. Burkhardt, J.A. Gulla, J. Liu, C. Weiss, J. Zhou: Semi Automatic Ontology Engineering in Business Applications, *Workshop Applications of Semantic Technologies*, INFORMATIK. 2008.

[3] T. Sabrina, A. Rosni, T. Enyakong: Extending Ontology Tree Using NLP Technique, In: *Proceedings of National Conference on Research & Development in Computer Science REDECS 2001*. 2001.

[4] W. Liu, A. Weichselbraun, A. Scharl, E. Chang: Semi-Automatic Ontology Extension Using Spreading Activation. *Journal of Universal Knowledge Management*, No. 1, pp. 50 – 58. 2005.

[5] J. McDonald, T. Plate, R. Schvaneveldt: Using pathfinder to extract semantic information from text, In: *Schvaneveldt*, pp. 149–164. 1990.

[6] M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, A. Vizedom: An Interactive Dialogue System for Knowledge Acquisition in Cyc, In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,* 2003.

[7] P. Cimiano, J. Vö lker: Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery, In: *Proceedings of NLDB 2005*, 2005, pp.227-238.

[8] D. Maynard, A. Funk, W. Peters: SPRAT: a tool for automatic semantic pattern-based ontology population, In: *Proceedings of International Conference for Digital Libraries and the Semantic Web*, Trento, Italy, 2009.

[9] P. Shah, D. Schneider, C. Matuszek, R. C. Kahlert, B. Aldag, D. Baxter, J. Cabral, M. Witbrock, J. Curtis: Automated population of Cyc: Extracting information about named-entities from the web, In: *Proceedings of the Nineteenth International FLAIRS Conference*, 2006, pp. 153-158.

[10] O. Medelyan, C. Legg: Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense, In: *Proceedings of Wiki-AI Workshop at the AAAI'08 Conference*, Chicago, US, 2008.

[11] S. Sarjant, C. Legg, M. Robinson, O. Medelyan: "All You Can Eat" Ontology-Building: Feeding Wikipedia to Cyc, In: *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence*, WI'09, Milan, Italy, 2009.

[12] M. Grobelnik, D. Mladenic, Knowledge Discovery for Ontology Construction, in: J. Davies, R. Studer, P. Warren (Eds.), Semantic Web Technologies: Trends and Research in Ontology-Based Systems, John Wiley & Sons, 2006, pp. 9–27.

[13] D. Lenat: Cyc: A Large-Scale Investment in Knowledge Infrastructure, *Communic. of the ACM 38 (11),* 1995.

[14] Yahoo! Finance, *http://finance.yahoo.com*

[15] C.R. Harvey: Yahoo Financial Glossary, *Fuqua School of Business, Duke University*, 2003.

[16] Central Intelligence Agency, The World Factbook: *https://www.cia.gov/library/publications/the-world-factbook*

# An Investigation and Extension of a Hyper-heuristic Framework

Prapa Rattadilok
Assumption University, Thailand
E-mail: prapa.rattadilok@gmail.com

**Keywords:** choice function, dynamic configuration, hierarchical controller, hyper-heuristic framework, low-level heuristics, timetabling.

**Received:** October 8, 2009

*Three modifications to the framework within which hyper-heuristic approaches operate are presented. The first modification automates a self learning mechanism for updating the values of parameters in the choice function used by the controller. Second, a procedure for dynamically configuring a range of low-level heuristics is described. Third, in order to effectively use this range of low-level heuristics the controller is redesigned to form a hierarchy of sub-controllers. The second and third modifications improve the inflexibility associated with having a limited number of low-level heuristics available to the controller. Experiments are used to investigate features of the hyper-heuristic framework and the three modifications including comparisons with previously published results.*

*Povzetek: Opisane so tri modifikacije hiper-hevrističnih pristopov.*

## 1 Introduction

As the complexity of optimisation problems increases methods which guarantee optimal solutions place excessive demands on computation time and computer resources. Alternative approaches have been developed including: heuristics, meta-heuristics, combinations of meta-heuristics referred to as hybrids, and more recently hyper-heuristics. Generally, these approaches do not guarantee optimal solutions but instead provide solutions of acceptable quality obtained with acceptable demands on algorithm development, tuning time, computation time, and computer resources. Surveys and comparisons among these approaches are presented in [1-9].

Heuristic approaches use rules derived from experience or intuition as opposed to those derived from mathematical formulations and they produce reasonable computational performance with conceptual simplicity. Problem specific knowledge is applied at the heuristic design phase and increases effectiveness but limits reusability for problems in other domains. Heuristic approaches have been applied successfully to a variety of specific problems including: resource investment [10]; resource usage [11]; project finance scheduling [12]; flow-shop scheduling [13]; graph colouring [14]; and train pathing [15]. Meta-heuristic approaches employ artificial intelligence methods and are different from simple heuristics in the manner in which the problem is modelled by attempting to prescribe more generic structures. Simulated annealing [16], tabu search [17], genetic algorithms [18], ant colony [19] and particle swarm optimisation [20], hill climbing and local search [21], and differential evolution [22] are well known meta-heuristic approaches. Interest in meta-heuristics has generated the development of hybrid approaches [8] and recent significant advances have combined meta-heuristics with other problem solving paradigms and improved their use in important application areas [23]. However, due to the evolutionary nature of meta-heuristic approaches the computation time may be unpredictable and there is often a need for a training period in order to tune the approach to the problem.

The aim of hyper-heuristic approaches is to be able to use the same procedures within and across problem domains without the need for extensive change to the basic components thus handling classes of problems rather than addressing one type of problem [24-28]. While most applications of meta-heuristics explore a search space of problem solutions hyper-heuristics explore a search space of low-level heuristics in order to select and apply an appropriate low-level heuristic. The framework in which hyper-heuristic approaches operate is presented in Figure 1 where at each stage of the search the controller uses information about the past performance of the low-level heuristics in order to select one to be used in the next stage. The selection is often made using a choice function and this process continues until a stopping condition is satisfied and the best solution is determined based on the value of the cost function.

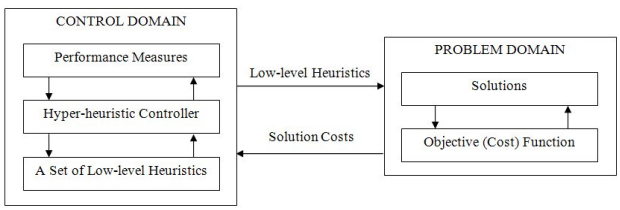



Figure 1: Hyper-heuristic framework.

The set of low level heuristics used by the controller are: pre-designed; limited in number; often involve add, drop, and swap operations; and remain the same throughout the search. They are problem specific and have limited reusability [29, 30]. Problems often include soft constraints, which may be violated, and hard constraints, which must not be violated, and these are usually assigned low and high positive weights, respectively, by the user. For any solution the value of the cost function is the sum of the weights associated with the constraints which are violated and for a feasible solution all of the hard constraints are satisfied.

The purpose of this article is to investigate three modifications to the hyper-heuristic framework The first modification (section 2.1) introduces a self learning mechanism for updating the values of choice function parameters so that the selection of low-level heuristics may be intensified or diversified appropriately. The second (section 2.2) introduces a procedure for dynamically configuring low-level heuristics in order to make a range of low-level heuristics available to the controller. In order to select effectively from this range of low-level heuristics the controller is redesigned (section 2.3) to form a hierarchy of sub-controllers each using the choice function described in section 2.1 and the dynamic configuration procedure described in section 2.2. Section 3 presents the results of experiments related to each of the three modifications including comparisons with previously published results. Section 4 discusses the results and draws conclusions. An Appendix is used to present details associated with the updating of choice function parameters.

# 2 Modifications to the hyper-heuristic framework

This section describes the three modifications to the hyper-heuristic framework in Figure 1.

## 2.1 Modifications to the choice function

The choice function proposed by Cowling et al. [30] and Soubeiga [31] is modified to allow the values of parameters to be updated automatically independently of any problem specific knowledge. The procedures work with complete rather than partial solutions and there is no need for an initial training period. The adjustment of parameters allows the search procedure to be intensified or diversified thus enhancing its applicability within and between problem domains. In cost minimisation problems the choice function selects a low-level heuristic by assessing the efficiency of the past performance of each of the low-level heuristics in decreasing the value of the cost function. Some may consistently decrease the value of the cost function and selection may be intensified on them. However, this may result in convergence to a local rather than global optimum and in such cases the choice function needs to select a low-level heuristic that diversifies the search to other parts of the solution space. Thus a suitable choice function should include factors which intensify or diversify the search appropriately. If at each stage of the search the low-level heuristics $H_1, H_2, H_3, \ldots, H_m$ are available to the choice function ( $F$ ) then a value of $F$ is computed for each low-level heuristic using,

$$F(H_j) = f_1(H_j) + f_2(H_j) + f_3(H_j), \quad \text{for } j = 1, 2, 3, \cdots, m. \tag{1}$$

The three factors in (1) represent: the past performance of the low-level heuristic ( $f_1$ ); the paired past performance of the low-level heuristic ( $f_2$ ); and the time since the low-level heuristic was last selected ( $f_3$ ). The first two factors are associated with intensifying the search while the last is associated with diversifying the search. In the Appendix section A1 each of the three factors is defined and the procedures for modifying parameters are presented in section A2. At the start of the search a solution is determined and one of the low-level heuristics is selected at random and applied to that solution. Information required in equations (A1), (A2), (A3), and (1) is updated and stored. The controller uses this information in (1) to determine the low-level heuristic with the largest value for $F$ and then using the procedures to adjust parameters this low-level heuristic or a different one is determined and used in the next iteration of the search. Subsequent iterations are conducted in the same manner until a stopping rule is satisfied and then the best solution among all of the solutions is selected as the final solution. The process is stochastic and a transition from one solution to another in the solution space is made using information about all of the previous transitions. Consequently, the process is not a Markov process and probabilistic equilibrium among the solutions is not attained [32]. Unless stated otherwise the choice function in (1) is used in all of the subsequent modifications and experiments.

## 2.2 Dynamically configured low-level heuristics

Swap-based low-level heuristics are used often and instead of generating a solution from scratch these low-level heuristics perform an exchange of attribute(s) between at least two swap candidates. For example, in a university timetabling problem an exchange may include swapping the days on which 2 classes are scheduled. Such low-level heuristics normally use problem specific knowledge in their design and applying them to different types of problems without any modification is usually infeasible. Different swap-based heuristics may be designed by choosing different configuration options at each of a set of configuration decision points. Examples of configuration options that may be selected at 4 commonly used configuration decision points are shown in Table 1.

| **Configuration Decision Points 1:** |
|---|
| The Number of Swap Candidates (λ) |
| Example Configuration Options: |

| **Configuration Decision Points 1:** The Number of Swap Candidates (λ) |
|---|
| 1. Two swap candidates?<br>2. More than two swap candidates?<br><br>*Comments: Determines the number of swap candidates involved in any trial swap process.* |
| **Configuration Decision Point 2:** Formation of λ Swap Candidate Sets |
| Example Configuration Options:<br>1. Non-violated assignments?<br>2. Violated assignments?<br><br>*Comments: Specifies the swap candidate for each of the candidate sets.* |
| **Configuration Decision Point 3:** Ordering Candidates in the λ Swap Candidate Sets |
| Example Configuration Options:<br>1. Slot number?<br>2. Ascending cost?<br><br>*Comments: Specifies the order in which the swap candidate from each candidate set enters the trial swap process.* |
| **Configuration Decision Point 4:** Acceptance Criteria |
| Example Configuration Options:<br>1. Best Solution?<br><br>*Comments: The trial swap process terminates when a solution satisfies the acceptance criteria and then the solution is returned to the controller.* |

Table 1: An example of configuration options associated with 4 configuration decision points.

From Table 1 it is seen that the number of configuration options at decision points 2 and 3 depends on the number of swap candidates (λ) chosen at decision point 1 and two or three swap candidates are commonly used. When forming swap candidate sets at decision point 2 the swap candidates may be shared among the sets formed.

The restrictions of using a fixed and limited number of problem specific low-level heuristics may be addressed by dynamically configuring swap-based low-level heuristics and using a hierarchical design for the controller. Dynamic configuration is discussed next and the design of a hierarchical controller is presented in section 2.3.

Figure 2(a) elaborates on elements of the framework in Figure 1 and represents a non-dynamic approach where the controller uses the choice function to select a low-level heuristic from a fixed set of usually no more than 10 swap-based low-level heuristics. Figure 2(b) presents the framework for an approach where the swap-based low-level heuristics are dynamically configured by the controller which selects configuration options at decision points as illustrated in Table 1 using a choice function of the same form as that used by the controller Figure 2(a) but with low-level heuristics replaced by configuration options. Dynamically configured low-level heuristics are generated and applied to the current solution and performance measures for configurations of these low-level heuristics are accumulated.

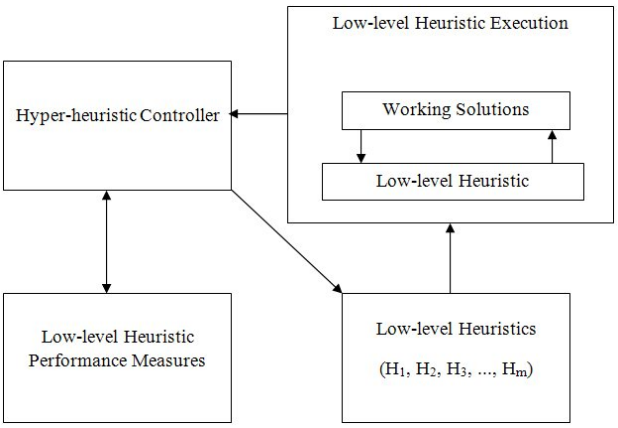



Figure 2 (a): Non-dynamic approach.

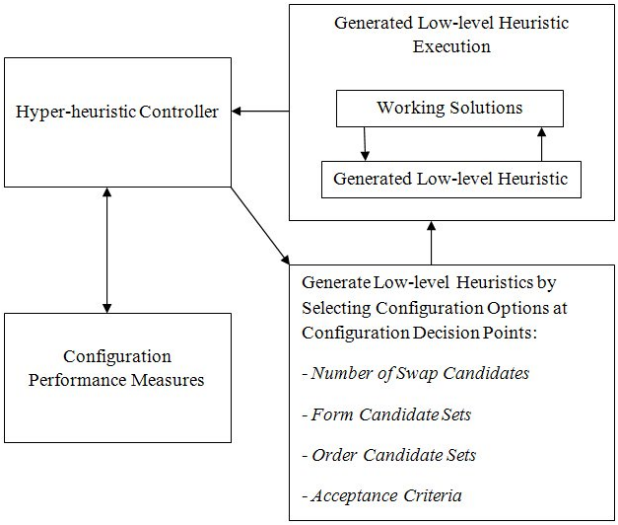



Figure 2 (b): Dynamically configured approach.

The use of a single choice function in the dynamic approach limits the total number of configuration options that the controller can work with effectively. Consequently, in order to improve the effectiveness of the dynamic approach the design of the controller needs to be reconsidered.

## 2.3 A hierarchical controller design

A new hierarchical design which operates in the controller component in Figure 2(b) is shown in Figure 3.

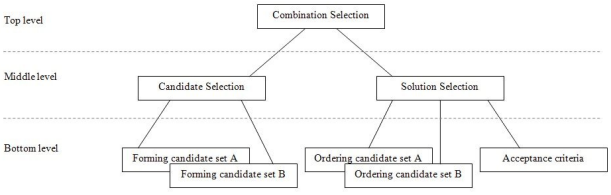



Figure 3: Hierarchical controller design.

Each component in Figure 3 is regarded as a sub-hyper-heuristic controller each of which uses a choice function where the low-level heuristic $H_j$ now represents a configuration option or a combination of configuration options depending on the level at which the sub-controller is operating. Information about the performance of configuration options or combinations of them is used to choose configuration options at different configuration decision points in the same manner as low-level heuristics were selected in the non-dynamic situation. There are 3 levels of sub-controllers in the hierarchy. The number of sub-controllers at the bottom level depends on the number of swap candidate sets formed (λ in Table 1). Figure 3 shows the case where λ = 2 and five sub-controllers are used: two for forming swap candidate sets; two for ordering swap candidate sets; and one for acceptance criteria. The middle level sub-controller chooses combinations of configuration options based on their performance in trails using the sets of configuration options generated at the bottom level. These configuration options are combined to form a low-level heuristic at the top level which is used in the next stage of the search.

# 3 Experiments

Published data sets and results for two different sets of problems are used in the experiments: international university timetabling competition problems (www.idsia.ch/Files/ttcomp2002/); and transportation services timetabling problems [33]. In order to allow comparisons experiments are designed to conform to the conditions associated with the published experimental results.

## 3.1 Experiments 1: The choice function

Two methods are investigated for generating an initial solution for a university timetabling problem: a random approach, which assigns random events (classes) to random slots (day, time, and room); and a greedy algorithm, which assigns an event to its best slot. On average across 5 experimental runs there are 1000 hard constraint violations in a randomly generated initial solution but only 200 for a greedily assigned solution. Consequently, greedy assignment is proposed for generating the initial solution used with a choice function.

Both of these methods are examined further by considering the average percentage of improvement in the cost of the initial solution if the search is allowed to continue for 1 minute and the results are shown in Figure 4. Table 2 shows the average number of hard and soft constraint violations for both methods at the end of 5 and 7 minutes.

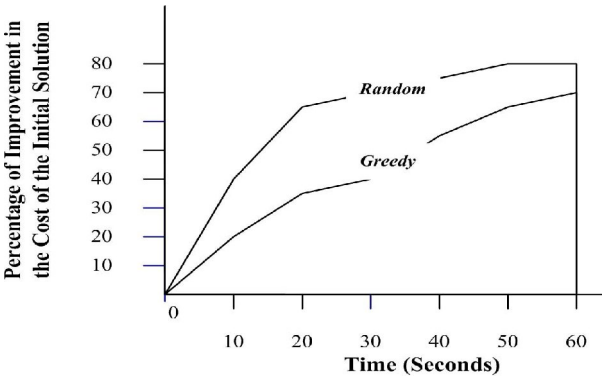



Figure 4: The percentage of improvement in the cost of the initial solution during the first minute.

| Time Limit: 5 Minutes | | |
|---|---|---|
| **Methods** | **Number of HCV** | **Number of SCV** |
| Random | 8.7 | 1152.9 |
| Greedy | 0 | 921.6 |
| **Time Limit: 7 Minutes** | | |
| **Methods** | **Number of HCV** | **Number of SCV** |
| Random | 0 | 767.8 |
| Greedy | 0 | 598.4 |

Table 2: The number of hard and soft constraint violations (HCV and SCV respectively) after 5 and 7 minutes.

From Figure 4 and Table 2 it is seen that beyond the initial solution the greedy assignment method continues to produce better results than the random method. In particular, the patterns in Figure 4 demonstrate the more general result that as the number of constraint violations decreases it becomes more difficult to reduce the number of constraint violations.

The results in Table 3 show the average costs of solutions across 10 experiment runs on each of 17 university timetabling problems using the choice function with and without automatic parameter modification. The values for the parameters $\alpha$, $\beta$, and $\delta$ in equations (A1), (A2), and (A3) are set randomly at 0.7, 0.5 and 0.1, respectively, at the start of the search and the same set of 7 low-level heuristics is used for all of the problems.

| Problems | Without Parameter Modification | With Parameter Modification |
|---|---|---|
| 1 | 118.2 | 89.5 |
| 2 | 104.1 | 77 |
| 3 | 117.5 | 80.8 |
| 4 | 234 | 175.9 |
| 5 | 199.2 | 139.5 |
| 6 | 255 | 188.5 |
| 7 | 120.8 | 84.4 |

| Problems | Without Parameter Modification | With Parameter Modification |
|---|---|---|
| 8 | 103.6 | 71.9 |
| 9 | 124.3 | 89 |
| 10 | 122.2 | 85.2 |
| 11 | 155.7 | 107.6 |
| 12 | 185.1 | 127.3 |
| 13 | 96.8 | 75.1 |
| 14 | 255.9 | 186 |
| 15 | 91 | 64.2 |
| 16 | 195.1 | 188.8 |
| 17 | 141.7 | 99.3 |

Table 3: The cost of solutions with and without parameter modification in the choice function.

From Table 3 it is seen that modification of the parameters reduces the average cost of the solutions for every problem. Although not shown here the same outcome occurs when the initial values of the parameters vary. From these experiments it is evident that automatic parameter modification is a useful enhancement to the choice function.

The next experiments examine the effect of varying the number of low-level heuristics available to the controller. In order to ensure that the results are not affected by the quality of the low-level heuristics used in each experiment all low-level heuristics are idle except for one which performs a simple swap on the solution. The idle heuristics may be selected by the controller and vary in terms of the time they take to execute but they have no effect on the solutions. Two problems are used from the university timetabling competition (U1, U2) and the transportation services timetabling (T1, T2) data sets. The number of low-level heuristics varies from 5 to 40 and in each case results are averaged across 5 experimental runs. The entries in Table 4 represent the percentage of calls received by the non-idle low-level heuristic above the percentage expected if it is called at random. For example, in problem U1 with 20 low-level heuristics the non-idle low-level heuristic received on average 27 percent of all of the calls which is 22 percent above the 5 percent expected if 20 low-level heuristics are called at random.

| Problems | Number of Low-level Heuristics | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
| Problem U1 | 56 | 52 | 30 | 22 | 12 | 11 | 10 | 10 |
| Problem U2 | 51 | 49 | 34 | 25 | 20 | 15 | 11 | 11 |
| Problem T1 | 41 | 44 | 39 | 33 | 25 | 20 | 15 | 14 |
| Problem T2 | 38 | 42 | 38 | 31 | 27 | 23 | 19 | 18 |

Table 4: The effects of increasing the number of low-level heuristics.

For each problem in Table 4 it is seen that as the number of low-level heuristics increases the idle low-level heuristics, which contribute nothing to the quality of the solution, are being called increasingly and the selection of low-level heuristics becomes almost random when there are a large number of low-level heuristics.

Figure 5 illustrates the effect over time on the cost of the solution of increasing the number of low-level heuristics.

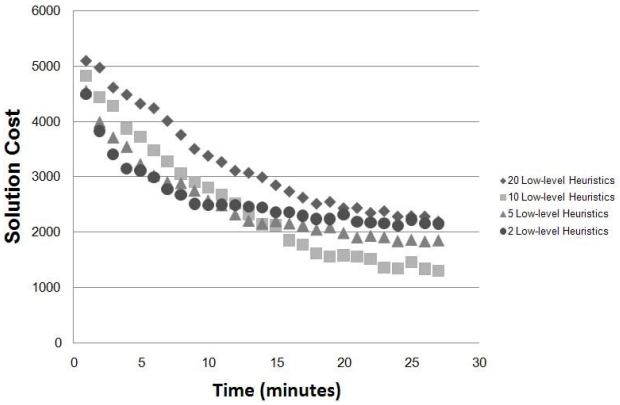



Figure 5: The number of low-level heuristics and the solution cost.

In Figure 5 it is seen that as the number of low-level heuristics increases it takes longer to establish good performance measures for them. This is evidenced by the flatter curve for 20 low-level heuristics compared to the curves for 5 or 10. For only 2 low-level heuristics performance measures are established earlier than in the other cases but there is much less opportunity to diversify the search and this makes it more difficult to escape from a local optimum. Based on the results in Table 3 and Figure 5 it is appropriate to recommended that the number of low-level heuristics should not be more than 10 or less than 5.

Figure 6 compares the average percentage of improvement across 5 experimental runs in the cost of the initial solution using the choice function approach, greedy selection, and random selection where for the greedy selection method low-level heuristics are selected and applied until no improvement is obtained and then a new low-level heuristic is selected. A university timetabling competition problem data set is used.

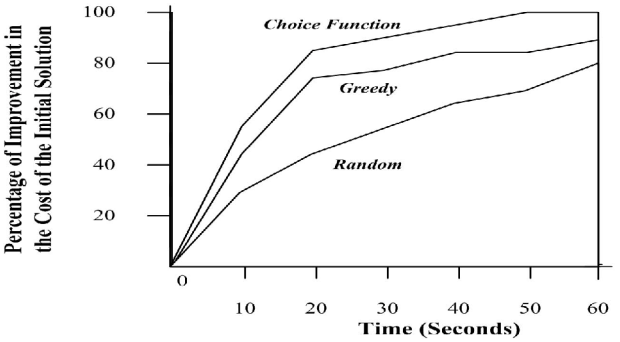



Figure 6: Improvement in the cost of the initial solution.

From Figure 6 it is seen that the choice function method consistently improves the initial solution more than either of the other methods. It is unlikely that the random approach will obtain similar quality solutions as the choice function even if more time is allowed. Random selection has a smoother improvement curve than greedy selection which has sharp improvement which flattens out quickly. The choice function has an even sharper

improvement. On average it reduces almost 90 percent of the cost of the initial solution within the first 20 seconds and reduces it by almost 100 percent after 50 seconds.

## 3.2 Experiments 2: Dynamic configuration of low-level heuristics

Using problems from the university timetabling competition these experiments examine different configurations in terms of: their ability to improve, worsen, or not change solution costs; their effect on different constraints; and their performance on different problems. Table 5 lists the configuration options used for forming candidate sets and acceptance criteria where the swap candidates in both sets are ordered based on their cost.

| Index | Forming Candidate Configuration Options |
|---|---|
| 0 | All slots |
| 1 | Occupied slots |
| 2 | Empty slots |
| 3 | Feasible slots |
| 4 | Infeasible slots |
| 5 | Constraint violated slots (specific constraint) |
| 6 | Non violated slots |
| **Index** | **Acceptance Criteria** |
| 0 | First cost improvement |
| 1 | First feasibility improvement |
| 2 | Minimum cost |
| 3 | Maximum feasibility |
| 4 | Minimum cost improvement |
| 5 | Maximum feasibility improvement |

Table 5: Configuration options for forming candidate sets and acceptance criteria.

In Table 5 the forming candidate configuration options specify which slots in the solution will be selected and used as the swap candidates. The slots can be divided into occupied and empty slots. Empty slots do not have events assigned to them and are therefore always feasible. Occupied slots can be feasible or infeasible. Infeasible slots are those assignments that violate hard constraints. These occupied slots can be further divided based on the number of violations for specific types of constraints and can also be assignments that have no constraint violations. At the end of each trial swap the acceptance criteria are checked. The acceptance criteria used improve either the cost or feasibility of the solution. The acceptance criteria can be specified to accept the first solution that satisfies one or both of these objectives. The minimum cost and maximum feasibility acceptance criteria include the selection of solutions that decrease the value of the cost function. Using the 7 configuration options and the ordering of swap candidates based on cost for each of the 2 swap candidate sets and the 6 acceptance criteria options a total of 294 (i.e. 7x1x7x1x6) low-level heuristics are generated and more are generated if there is more than 1 option for ordering swap candidates in the 2 candidate sets.

The configurations derived from Table 5 are categorised according to their performance. If the largest percentage of all of the calls made on a configuration produce an improved solution then the configuration is categorised as ‘improving’ and similarly configurations may be categorised as ‘unchanging’ or ‘worsening’. Table 6 shows the results for the first problem in the university timetabling competition when random selection is used to dynamically configure the configuration options. The top 3 configurations in each of the 3 categories are shown. The meaning of the entries in Table 6 is explained using the example of the best configuration in the ‘improving’ category (i.e. 2-5-2 (90.95%)). From Table 5 this means that the configuration is generated by choosing ‘Empty slots’ as the candidates for the first candidate set, ‘Constraint violated slots’ as the candidates for the second candidate set, ‘Minimum cost’ as the acceptance criteria, and on average 90.95 percent of the times when it is called this configuration improves the solution.

| Unchanging | Worsening | Improving |
|---|---|---|
| 2-3-3 (93.13%) | 5-0-2 (62.12%) | 2-5-2 (90.95%) |
| 2-6-3 (91.27%) | 5-0-3 (60.83%) | 2-0-5 (90.57%) |
| 3-6-3 (89.95%) | 5-6-3 (59.77%) | 2-5-5 (90.47%) |

Table 6: Top 3 configurations for performance categories.

From Table 6 it is seen that some configurations do not lead to improvement in the cost but they may be used for diversifying the search. For example, using configurations that are associated with worsening costs would lead to diversification and this is often desirable. Configurations with high chances of improving costs are appropriate when search intensification is desired especially near the end of a search.

Table 7 presents a different view of the configurations where the columns represent 5 different types of constraints: ‘Room’ (an event must be assigned to a room that has all of the resources needed); ‘Student’ (a student cannot attend more than one event at any one time); ‘One Per Day’ (a student attends only one event per day); ‘More Than Two’ (a student attends more than 2 classes consecutively); and ‘Late’ (a student attends an event at the last period of the day). The configurations are represented in the same manner as in Table 6 but the percentage now indicates the average amount of improvement they produced in the cost of the initial solution each time they were used. The best 3 configurations are shown for each constraint.

| Types of Constraints | | | | |
|---|---|---|---|---|
| **Room** | **Student** | **One Per Day** | **More Than Two** | **Late** |
| 2-4-2 (0.96%) | 0-4-2 (24.36%) | 4-2-2 (4.09%) | 2-5-4 (1.50%) | 2-1-2 (0.36%) |
| 2-5-2 | 5-4-2 | 5-6-2 | 2-5-2 | 2-5-4 |

| Types of Constraints | | | | |
|---|---|---|---|---|
| **Room** | **Student** | **One Per Day** | **More Than Two** | **Late** |
| (0.96%) | (18.61%) | (3.45%) | (1.43%) | (0.28%) |
| 0-4-2 (0.92%) | 1-4-2 (17.31%) | 4-3-2 (3.05%) | 2-1-2 (1.28%) | 2-5-2 (0.25%) |

Table 7: The best configurations for different types of constraints.

From Table 7 it is seen that configuration 2-5-2 benefits 3 constraints while configurations 2-5-4, 2-1-2, and 0-4-2 all benefit 2 constraints. When 'Empty slot' is used to form one of the candidate sets the 'Minimum cost' acceptance criteria is commonly used. The 'Student' constraint violations are removed at a high rate each time, while 'Late' constraint violations are removed at a much lower rate. If the 'Room' and 'Student' constraints are hard constraints and the other three are soft constraints then configuration 0-4-2 appears to perform well on those hard constraints.

The first 3 problems in the international timetabling competition are used to obtain the results in Table 8. The entries in the table have the same meaning as those in Table 7.

| **Problem 1** | **Problem 2** | **Problem 3** |
|---|---|---|
| 0-4-4 (0.99%) | 0-4-2 (1.53%) | 2-1-3 (0.80%) |
| 1-4-2 (0.94%) | 1-4-2 (1.32%) | 2-5-3 (0.74%) |
| 2-4-5 (0.94%) | 5-4-2 (1.21%) | 2-4-5 (0.73%) |

Table 8: Best configurations for 3 timetabling problems.

From Table 8 it is seen that the best configuration varies from one problem to another but using the 'Infeasible slots' configuration option for forming the second candidate set is beneficial across the 3 problems and combining it with the 'Empty slots' option for the first candidate set is beneficial for problems 1 and 3. The configuration 1-4-2 works well in problems 1 and 2 but with different average improvements and the 'Minimum cost' acceptance criteria is dominant for problem 2 and useful in problem 1 but not problem 3. Even when a random configuration is used consistent performance measures are observed when some configuration options are combined and these performance measures may be used to influence the configuration by the controller in much the same way that a human heuristic designer applies their past experience in selecting suitable low-level heuristics for a problem. If configuration options are good when combined then it is possible that there is a positive relationship between the options and making modifications to multiple configuration points simultaneously may assist the controller in making configurations.

## 3.3 Experiments 3: Using a hierarchical design for the controller

These experiments compare the non-dynamic approach in Figure 2(a), which uses a single choice function and a fixed set of low-level heuristics, with the dynamic approach in Figure 2(b), which incorporates the dynamic configuration of low-level heuristics and a hierarchical design for the controller as described in Figure 3. Data sets from the university timetabling competition and the transportation services timetabling problems are used in the experiments.

The non-dynamic approach uses the following 8 low-level heuristics: $\boldsymbol{H_1}$: Swap the highest cost feasible assignment with every other assignment (in ascending order based on their cost), select the best quality solution; $\boldsymbol{H_2}$: Same as $H_1$ but select the first improving quality solution; $\boldsymbol{H_3}$: Same as $H_1$ but the candidate assignments are ordered randomly; $\boldsymbol{H_4}$: Same as $H_3$ but select the first improving quality solution; $\boldsymbol{H_5}$: Swap the highest cost infeasible assignment with every other assignment (in ascending order based on their cost), select the best quality solution; $\boldsymbol{H_6}$: Same as $H_5$ but select the first improving quality solution; $\boldsymbol{H_7}$: Same as $H_5$ but the candidate assignments are ordered randomly; and $\boldsymbol{H_8}$: Same as $H_7$ but select the first improving quality solution.

For a fair comparison, the configuration options for the dynamic approach are limited to those that will generate low-level heuristics equivalent to the non-dynamic set. The configuration options for the 4 configuration points are: ***The Number of Swap Candidates (λ)***: 2; ***Forming λ Swap Candidate Sets***: Highest cost feasible assignment, Highest cost infeasible assignment; ***Ordering λ Swap Candidate Sets***: Ascending cost based, Random; and ***Acceptance Criteria***: Best quality, First improving quality. To ensure the same number of low-level heuristics as in the non-dynamic set, the ***Forming*** options selects a swap candidate for the first candidate set and the second candidate set contains all other assignments. Because there is only one assignment in the first candidate set the ***Ordering*** options are only used to order the candidates in the second candidate set.

Table 9 compares the 4 best results from the university timetabling competition (www.idsia.ch/Files/ttcomp2002/results.htm) with the results obtained using the non-dynamic and dynamic approaches where the solution costs are the averages from 10 experimental runs. It is noted that the results for the competition were obtained using algorithms specifically designed for these problems while the dynamic and non-dynamic approaches use generic configuration options and low-level heuristics, respectively.

| **Problem Data Set** | **Approach** | | | | | |
|---|---|---|---|---|---|---|
| | **Problem Specific Algorithms** | | | | **Hyper-heuristics** | |
| | **1** | **2** | **3** | **4** | **Non-Dynamic** | **Dynamic** |
| 1 | 45 | 61 | 85 | 63 | **80.1** | **79.5*** |
| 2 | 25 | 39 | 42 | 46 | 73 | 73.2 |
| 3 | 65 | 77 | 84 | 96 | **77.8** | **77.6*** |
| 4 | 115 | 160 | 119 | 166 | 174.3 | 175.7 |
| 5 | 102 | 161 | 77 | 203 | 289.9 | 292 |
| 6 | 13 | 42 | 6 | 92 | 131.2 | 133.5 |

| Problem Data Set | Approach | | | | | |
|---|---|---|---|---|---|---|
| | Problem Specific Algorithms | | | | Hyper-heuristics | |
| | 1 | 2 | 3 | 4 | Non-Dynamic | Dynamic |
| 7 | 44 | 52 | 12 | 118 | 180.2 | 170.9* |
| 8 | 29 | 54 | 32 | 66 | 82.1 | 82.2 |
| 9 | 17 | 50 | 184 | 51 | **68.9** | **69.6** |
| 10 | 61 | 72 | 90 | 81 | **83.3** | **83.3*** |
| 11 | 44 | 53 | 73 | 65 | 79.9 | 81.2 |
| 12 | 107 | 110 | 79 | 119 | 120.2 | **118.1*** |
| 13 | 78 | 109 | 91 | 160 | **101.2** | **103.5** |
| 14 | 52 | 93 | 36 | 197 | 255.7 | 253.4* |
| 15 | 24 | 62 | 27 | 114 | 119 | 123.6 |
| 16 | 22 | 34 | 300 | 38 | **64.2** | **64.8** |
| 17 | 86 | 114 | 79 | 212 | **169.9** | **170.5** |
| 18 | 31 | 38 | 39 | 40 | 61.3 | 61.3* |
| 19 | 44 | 128 | 86 | 185 | 186.1 | 186.2 |
| 20 | 7 | 26 | 0 | 17 | 93.7 | 94.7 |

Table 9: University timetabling solution costs.

In Table 9 the highlighted values represent the 40 percent of cases where one or both of the hyper-heuristic approaches achieved a lower cost than at least one of the best 4 competition results and this is encouraging considering the problem specific nature of the algorithms used in the competition. The results for the hyper-heuristic approaches are very similar but the dynamic approach achieved the same or better results to the non-dynamic approach in 35 percent of cases (marked *).

Table 10 compares the same hyper-heuristic approaches with the problem specific algorithm BOOST [32] for transportation services timetabling problems. The results are the average solution costs from 10 experimental runs.

| Problem Data Set | Approach | | |
|---|---|---|---|
| | Problem Specific Algorithm | Hyper-Heuristics | |
| | BOOST [32] | Non-Dynamic | Dynamic |
| 1 | 492 | **492** | **492*** |
| 2 | 1376 | **1376** | **1376*** |
| 3 | 1678 | **1678** | **1678*** |
| 4 | 1641 | 1761 | 1756* |
| 5 | 1396 | **1396** | **1396*** |
| 6 | 1389 | 1421 | 1434 |
| 7 | 1465 | 1606 | 1604* |
| 8 | 1858 | 2045 | 2044* |
| 9 | 3409 | **3409** | 3411 |
| 10 | 3502 | **3502** | 3533 |
| 11 | 14919 | 15598 | 15632 |
| 12 | 6028 | 6268 | 6272 |
| 13 | 21963 | 23987 | 24132 |
| 14 | 12510 | 14498 | 14498* |

Table 10: Transportation services timetabling solution costs.

In Table 10 the highlighted values represent the 43 percent of cases where one or both of the hyper-heuristic approaches achieved the same cost as BOOST which is specifically designed for the transportation problems while the hyper-heuristic approaches are using generic configurations and low-level heuristics. The results for the hyper-heuristic approaches are very similar but the dynamic approach achieved the same or better results compared to the non-dynamic approach in 57 percent of cases (marked *).

From the results in Tables 9 and 10 it is seen that the dynamic approach has performed well across 2 different types of problems using a generic set of configuration options. It was not expected that the hyper-heuristic approaches would achieve better results than algorithms specifically designed for these problems but their performance is acceptable and compares favourably with the specific algorithms. In addition, for the dynamic approach increasing the number of configuration options increases the possible number of configurations. Therefore, a longer time is required for the controller to establish reliable performance measures and it is expected that the dynamic approach may obtain equally good solutions in all cases to the non-dynamic approach given a longer search time.

The sequence of the trips in a transportation services timetabling problem determines the feasibility of the solution where no trip precedes an earlier one. A candidate selection configuration option may be added where instead of forming the second candidate set by selecting every other swap candidate these candidates must be the slots on different buses from the first candidate set. This limits the number of candidates in the second candidate set and minimises the number of swap trials needed especially when the sequence of all assignments is time feasible. The 5 problems (10 – 14) in Table 10 with the highest cost are used in the next set of experiments which examine the effect of making this simple modification to the dynamic approach based on information specific to the timetabling problem. Table 11 shows the average cost of solutions from 10 experimental runs using the dynamic approach with and without this modification and the corresponding costs for BOOST as shown in Table 10.

| Approach | Problem Data Set (as in Table 10) | | | | |
|---|---|---|---|---|---|
| | 10 | 11 | 12 | 13 | 14 |
| Modified Dynamic | **3502** | **15568** | **6066** | 24132 | **14467** |
| Dynamic (as in Table 10) | 3533 | 15632 | 6272 | 24132 | 14498 |
| BOOST (as in Table 10) | 3502 | 1419 | 6028 | 21963 | 12510 |

Table 11: Transportation services timetabling solution costs with modified dynamic approach.

From the highlighted costs in Table 11 it is seen that the modification has improved the solution using the dynamic approach in 4 of the 5 problems. For problem 10 the modified dynamic approach has an equal cost to BOOST and for problem 13 the cost has not changed. The modification has improved the performance of the dynamic approach for these transportation services

problems but, although the results are not shown, it was not as beneficial for the university timetabling problems. However, it does demonstrate that often with the dynamic approach it is easy to insert problem specific knowledge into the configuration options with beneficial results.

# 4 Conclusion

The framework within which hyper-heuristics operate has been investigated and three modifications have been developed and tested using experiments and comparisons with published results.

The first modification introduced a self learning mechanism into the choice function to modify the values of parameters in the function as the search progresses in order to allow intensification and diversification of the search. Experimental evidence showed that the modification improved the performance of the choice function which performed better than either a greedy or random method for selecting low-level heuristics. Other experiments showed that a greedy algorithm is an appropriate means of developing an initial solution and no more than 10 or less than 5 low-level heuristics should be used in the non-dynamic approach.

The second and third modifications represent two steps toward addressing the inflexibility associated with a non-dynamic approach where there is a fixed and limited number of pre-designed problem specific low-level heuristics available to a controller using a single choice function. The second modification introduced procedures for dynamically configuring low-level heuristics and the third modification redesigned the controller using a hierarchy of sub-controllers working together at different levels to generate and combine configurations. The combination of these two modifications resulted in a dynamic approach.

Experiments examined the procedure for dynamically configuring low-level heuristics in terms of: their effect on solution costs; their effect on different constraints; and their performance on different problems. The procedure was shown to be feasible but it was observed that a large number of configurations were generated and that it may be possible to combine those with desirable characteristics. However, with dynamic configuration the effectiveness of a controller using a single choice function was questionable and the controller was redesigned to form a hierarchy of sub-controllers. Experiments compared the performance of the new dynamic hyper-heuristic approach, the non-dynamic hyper-heuristic approach, and published results for algorithms that were specifically designed for the particular problems. The problems represented two different timetabling tasks and the dynamic and non-dynamic approaches used generic configuration options and low-level heuristics, respectively. It was not expected that either of the hyper-heuristic approaches would achieve better results than the problem specific algorithms but for 40 percent of the university problems and 43 percent of the transportation problems the hyper-heuristic approaches achieved a lower cost than problem specific algorithms. The results for the non-dynamic and dynamic approaches were very similar but the dynamic approach achieved the same or better results on 57 percent and 35 percent of the transportation and university problems, respectively. The dynamic approach performed well across these two different types of problems using a generic set of configuration options. For the dynamic approach increasing the number of configuration options increases the number of configurations and the controller takes longer to establish reliable performance measures so it is possible that the dynamic approach may perform even better compared to the non-dynamic approach given a longer search time. For a subset of transportation problems it was demonstrated that a simple modification to configuration options using problem specific knowledge produced an improvement in the solutions generated by the dynamic approach.

Hyper-heuristic approaches are relatively new and the findings for the modifications investigated in this study are promising. In particular, the new dynamic approach developed here is encouraging but further studies are needed to: verify its applicability in other problem domains; develop a more intelligent controller able to identify the best configuration options for particular problems; and further investigate methods suggested by Rattadilok et al. [34] to allow the search to be carried out simultaneously on multiple processors.

# References

[1] V. Maniezzo, S. Vob, P. Hansen, (Editors) (2009) Special Issue on Mathematical Contributions to Metaheuristics, *Journal of Heuristics*, 3, pp.197-312.

[2] R. Qu, E.K. Burke, B. McCollum, (2009) Adaptive automated construction of hybrid heuristics for exam timetabling and graph colouring problems, *European Journal of Operational Research*, 198(2), pp.392-404.

[3] R. Qu, E.K. Burke, B. McCollum, L.G.T. Merlot, S.Y. Lee, (2009) A Survey of Search Methodologies and Automated System Development for Examination Timetabling, *Journal of Scheduling*, 12(1), pp.55–89.

[4] Z. Rios, (Editor) (2009) Special Issue on Heuristic Research: Advances and Applications, *Journal of Heuristics*, 2, pp.105-196.

[5] E. Alba, E. Talbi, A.J. Nebro, (Editors) (2008) Special Issue on Advances in Metaheuristics for Multiobjective Optimization, *Journal of Heuristics*, 5, pp.311- 412.

[6] G.I. Zobolas, C.D. Tarantilis, G. Ioannou, (2009) Minimizing makespan in permutation flow shop scheduling problems using a hybrid metaheuristic algorithm, *Computers and Operations Research*, 36(4), pp. 1249-1267.

[7] B. McCollum, (2007) A Perspective on Bridging the Gap between Research and Practice in University Timetabling, in E.K. Burke, H. Rudova, (Editors) Practice and Theory of Automated

Timetabling VI, *Lecture Notes in Computer Science*, 3867, pp.3-23.
[8] C. Blum, A. Roli, (2003) Metaheuristics in combinatorial optimisation: overview and conceptual comparison, *ACM Comput. Sruv.*, 35, pp.268-308.
[9] F.W. Glover, G.A. Kochenberger, (2003) *Handbook of Metaheuristics*, International Series on Operational Research and Management Science, Vol.57, Springer.
[10] A.A. Najafi, F. Azimi, (2009) A priority rules-based heuristic for resource investment project scheduling problem with discounted cash flows and tardiness penalties, *Mathematical Problems in Engineering*, Vol. 2009, article ID 106425.
[11] C. Weng, X. Lu, (2005). Heuristic scheduling for bag-of-tasks applications in combination with QOS in the computational grid source, *Future Generation Computer Systems*, 21(2), pp.271-280.
[12] A. Elazouni, (2009) Heuristic method for multi-project finance-based scheduling, *Construction Management and Economics*, 27(2), pp.199-211.
[13] J.P.O. Fan, G.K. Winley, (2008) A Heuristic Search Algorithm for Flow-shop Scheduling, *Informatica*, 32, pp.453-464.
[14] H.E. Mausser, M.J. Magazine, (1996) Comparison of neural and heuristic methods for a timetabling problem, *European Journal of Operational Research*, 93(2), pp.271-287.
[15] Y. Lee, C. Chen, (2009) A heuristic for the train pathing and timetabling problem, *Transportation Research Part B: Methodological*, 43(8-9), pp.837-851.
[16] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, (1983) Optimization by Simulated Annealing, *Science*, 220(4598), pp.671-680.
[17] F. Glover, (1989) Tabu Search - Part I, *ORSA Journal on Computing*, 1(3), pp.190-206.
[18] D.E. Goldberg, (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston, MA: Kluwer Academic Publishers.
[19] M. Dorigo, T. Stützle, (2004) *Ant Colony Optimization*. MIT Press.
[20] J. Kennedy, R.C. Eberhart, Y. Shi, (2001) *Swarm Intelligence*. Morgan Kaufmann Publishers.
[21] B. Selman, H. Kautz, B. Cohen, (1993) Local Search Strategies for Satisability Testing, *Proceedings of 2nd DIMACS Challenge Workshop on Cliques, Color-ing, and Satisability*, Rutgers University, pp.290-295.
[22] K.V. Price, M.R. Storn, J.A. Lampinen, (1998) *Differential Evolution: A Practical Approach to Global Optimization*. Springer.
[23] T.G. Crainic, M. Gendreau, L-M. Rousseau, (Editors) (2010) Special Issue on Recent Advances in Metaheuristics, *Journal of Heuristics*,16, pp.235-535.
[24] R. Bai, J. Blazewicz, E.K. Burke, G. Kendall, B. McCollum, (2007) A simulated annealing hyper-heuristic methodology for flexible decision support, *Technical Report*, School of Computer Science, University of Nottingham (available at: www.asap.cs.nott.ac.uk/publications/pdf/Bai_et_al_2007-8.pdf).
[25] E.K. Burke, B. McCollum, A. Meisels, S. Petrovic, R. Qu, (2007) A Graph-Based Hyper Heuristic for Educational Timetabling Problems, *European Journal of Operational Research*, 176(1), pp.177-192.
[26] E.K. Burke, S. Petrovic, R. Qu, (2006) Case-based heuristic selection for timetabling problems, *Journal of Scheduling*, 9(2), pp.115-132.
[27] E.K. Burke, G. Kendall, E. Soubeiga, (2003) A tabu search hyperheuristic for timetabling and rostering, *Journal of Heuristics*, 9(6), pp.451-470.
[28] P. Cowling, G. Kendall, E. Soubeiga, (2000) A Hyperheuristic Approach to Scheduling a Sales Summit, in E.K. Burke, W. Erben, (Editors) Proceedings of the Third International Conference on the Practice and Theory of Automated Timetabling, *Lecture Notes in Computer Science*, 2079, pp.176 – 190.
[29] P. Cowling, G. Kendall, E. Soubeiga, (2002) Hyperheuristics: A Robust Optimisation Method Applied to Nurse Scheduling, in Parallel Problem Solving from Nature VI (PPSN 2002), *Lecture Notes in Computer Science*, pp.851–860.
[30] P. Cowling, G. Kendall, E. Soubeiga, (2001) Hyperheuristic: A Tool for Rapid Prototyping in Scheduling and Optimisation, *Proceedings of the Second European Conference on Evolutionary Computing for Combinatorial Optimisation* (EvoCop 2002), pp.1–10.
[31] E. Soubeiga, (2003) *Development and Application of Hyperheuristics to Personnel Scheduling*, Ph.D. thesis, University of Nottingham School of Computer Science.
[32] S. Karlin, (1972) *A First Course in Stochastic Processes*, Academic Press, New York.
[33] R.S.K. Kwan, M.A. Rahim, (1999) Object Oriented Bus Vehicle Scheduling – the BOOST System, in N.H.M. Wilson, (Editor) *Computer-Aided Transit Scheduling of Public Transport*, Springer, pp.177-179.
[34] P. Rattadilok, A. Gaw, R.S.K. Kwan, (2005) Distributed Choice Function Hyper-heuristics for Timetabling and Scheduling, in E. Burke, M. Trick, (Editors) Practice and Theory of Automated Timetabling V, *Lecture Notes in Computer Science*, 3616, pp.51-67.

# Appendix

## A1 Factors in the choice function

***Factor*** $f_1$***:*** A measure of the past performance of the low-level heuristic $H_j$ is calculated using,

$$f_1(H_j) = \sum_{n=1}^{l} \alpha^n \left( \frac{I_n(H_j)}{T_n(H_j)} \right). \quad \text{(A1)}$$

$I_n(H_j)$ is the change in the cost function the $n^{th}$ last time $H_j$ was used, $l$ refers to the first time that $H_j$ was selected, and if $I_n(H_j) > 0$ then the value of the cost function was decreased. $T_n(H_j)$ is the amount of CPU time in milliseconds from the time the low-level heuristic $H_j$ was used the $n^{th}$ last time until the time when it returned a solution to the controller. The parameter $\alpha$ is normalised to have a value in the interval (0, 1) and it assigns a decreasing geometric sequence of weights to the past performance measures of $H_j$. The initial value of $\alpha$ is determined randomly and if necessary it is automatically modified during the search as described below.

***Factor*** $f_2$***:*** The performance of a low-level heuristic may be affected by the low-level heuristic that was used immediately before it. Suppose that $H_k$ was used at the last iteration and the use of $H_j$ next is being considered. Then the measure of the past performance of the pair $(H_k, H_j)$ is calculated using,

$$f_2(H_k, H_j) = \sum_{n=1}^{l} \beta^n \left( \frac{I_n(H_k, H_j)}{T_n(H_k, H_j)} \right). \quad \text{(A2)}$$

$I_n(H_k, H_j)$ is the change in the cost function the $n^{th}$ last time the pair $(H_k, H_j)$ was used, $l$ refers to the first time in the search that $H_j$ was used immediately after $H_k$, and if $I_n(H_k, H_j) > 0$ then the value of the cost function decreased. $T_n(H_k, H_j)$ is the amount of CPU time in milliseconds from the time the pair $(H_k, H_j)$ was used the $n^{th}$ last time until the time when a solution was returned to the controller. The parameter $\beta$ is normalised to have a value in the interval (0, 1) and it assigns a decreasing geometric sequence of weights to the past performance measures of the pair $(H_k, H_j)$. The initial value of $\beta$ is determined randomly and if necessary it is automatically modified during the search as described below.

***Factor*** $f_3$***:*** The two factors $f_1$ and $f_2$ intensify the search on low-level heuristics which have performed well in the past. The third factor $f_3$ diversifies the search by considering low-level heuristics that may not have been used for some time and this is relevant in situations where the search is stuck at a local optimum. The value of $f_3$ is calculated for each low-level heuristic $H_j$ using,

$$f_3(H_j) = \delta.\tau(H_j). \quad \text{(A3)}$$

$\tau(H_j)$ is the amount of CPU time in milliseconds since the low-level heuristic $H_j$ was last used and each time $H_j$ is used $\tau(H_j)$ is reset to zero. The initial value of $\delta$ is selected randomly in the interval (0, 1) and if necessary it is automatically modified during the search as described below.

## A2 Modification of parameters in the choice function

Suppose that there are $m$ low-level heuristics $H_1, H_2, H_3, \ldots, H_m$, $H_k$ has just been used, and the choice function suggests the use of $H_j$ at the next iteration. Before using $H_j$ determine which of the factors $f_1(H_j)$, $f_2(H_j)$, and $f_3(H_j)$ has the largest value $L$.

**1.** If $L = f_1(H_j)$ (or $f_2(H_j)$) then use $H_j$ in the next iteration and modify the value of $\alpha$ to $\alpha(1+\varepsilon)$ (or $\beta$ to $\beta(1+\varepsilon)$) where $\varepsilon = \frac{I_1(H_j)}{mc_0}$ $\left( \text{or } \frac{I_1(H_k, H_j)}{mc_0} \right)$ and $c_0$ is the value of the cost function for the low-level heuristic used at the start of the search. Thus the value of $\alpha$ (or $\beta$) increases as confidence grows in the forecasts provided by the choice function and decreases when a low-level heuristic cannot be found that has decreased the value of the cost function the last time it was used.

If $I_1(H_j) = 0$ (or $I_1(H_k, H_j) = 0$) and this has not been occurring regularly then no change in the value of the cost function is preferable to a decrease and the value of $\alpha$ (or $\beta$) needs to be decreased by a small amount where $\varepsilon = \frac{-T_1(H_j)}{m^2 n_j}$ $\left( \text{or } \frac{-T_1(H_k, H_j)}{m^2 n_j} \right)$ and $n_j$ is the number of times $H_j$ has been used in the search. Here $\varepsilon$ is proportional to the time that might be wasted by using $H_j$ and it is a small value if $n_j$ is large which means $H_j$ (or the pair $(H_k, H_j)$) has often performed well in the past. If $I_1(H_j) = 0$ (or $I_1(H_k, H_j) = 0$) and this has been occurring regularly (as defined by the user) then the value of $\delta$ is modified as in part 4 of the modification procedures below.

**2.** If $L = f_3(H_j)$ then determine the trial low-level heuristic $H_i$ which maximises the value of $f_1(H_h) + f_2(H_h)$ for *h* = 1, 2, 3, ⋯, *m*. Use $H_i$ as a trial and if $F(H_i) < F(H_k)$ then decrease the value of $\delta$ to $\delta(1-q)$ and accept that $H_i$ is the low-level heuristic to use in the next iteration. This means that diversification of the search using $H_j$ has been suggested prematurely.

Before the trial use of $H_i$ is conducted which $F(H_j) > F(H_i) > f_1(H_j) + f_2(H_j) + f_3(H_i)$ means that $f_3(H_j) > f_3(H_i)$. If the use of $H_i$ decreases the value of the cost function then it is preferred to $H_j$ and the value of $\delta$ needs to be decreased in order to lessen the effect of the factor $f_3$ in the choice function. If $H_j$ has been suggested prematurely then it is desirable to use $H_i$ and have $F(H_i) > F(H_j)$ which means that if the value of $\delta$ changes to $\delta(1-q)$ then $F(H_i) - qf_3(H_i) > F(H_j) - qf_3(H_j)$ and so $q > \frac{F(H_j) - F(H_i)}{f_3(H_j) - f_3(H_i)} > 0$. Thus an appropriate value for *q* is $\frac{F(H_j) - F(H_i)}{f_3(H_j) - f_3(H_i)} + \gamma$, where $\gamma$ is an arbitrarily selected small positive number.

Otherwise, use $H_j$ as suggested by the choice function and do not change the value of $\delta$. This means that diversification of the search using $H_j$ is appropriate.

**3.** If the values of $f_1(H_j)$, $f_2(H_j)$, and $f_3(H_j)$ are the same then use $H_j$ as suggested by the choice function and do not change the values of $\alpha, \beta,$ and $\delta$.

**4.** Regardless of the value of *L* if the suggested low-level heuristic $H_j$ has been selected and used many times in recent iterations and continually fails to decrease the value of the cost function then increase the value of $\delta$ to $\delta + p$ in order to diversify the search using $H_n$ which maximises the value of $\tau(H_h)$ for *h* = 1, 2, 3, … , *m*. $H_n$ is the low-level heuristic which was last used the longest time ago and $F(H_j) > F(H_n)$. However, it is desirable to diversify the search so that $F(H_n) + p\tau(H_n) > F(H_j) + p\tau(H_j)$ and so $p > \frac{F(H_j) - F(H_n)}{\tau(H_n) - \tau(H_j)} > 0$. Thus and an appropriate value for *p* is $\frac{F(H_j) - F(H_n)}{\tau(H_n) - \tau(H_j)} + \nu$, where $\nu$ is an arbitrarily selected small positive number.

# Expanding Mental Outlook with the Help of Concept Maps

Burdescu Dumitru Dan, Mihăescu Marian Cristian, Ionaşcu Marian Costel and Buligiu Ionut
University of Craiova, Romania
E-mail: burdescu@software.ucv.ro

Logofătu Bogdan
University of Bucharest, Romania
E-mail: logofatu@credis.ro

**Keywords:** mental outlook, concept maps, e-Learning, clustering

**Received:** October 25, 2009

*Expanding mental outlook for learners is one of the important open problems in e-Learning. Online should have the property of expanding and enhancing the mental outlook of learners in general and also, in particular, concerning the studied discipline. This paper presents an approach to this issue. The tool used for expanding the mental outlook is represented by concept maps. Concept maps are used for representing relationships among concepts that define a certain area (e.g. discipline in e-Learning). The concept map is build such that it represents a concrete and broad representation of the domain. As an example, this paper presents a concept map built for Data structures and Algorithms course and more exactly for Binary Search Trees chapter.*

*Povzetek: Na primeru binarnih iskalnih dreves je predstavljen nov koncept e-učenja.*

## 1 Introduction

Concept maps are a result of Novak and Gowin's [3] research into human learning and knowledge construction. Novak [1] proposed that the primary elements of knowledge are concepts and relationships between concepts are propositions. Novak [2] defined concepts as "perceived regularities in events or objects, or records of events or objects, designated by a label". Propositions consist of two or more concept labels connected by a linking relationship that forms a semantic unit. Concept maps are a graphical two-dimensional display of concepts (usually represented within boxes or circles), connected by directed arcs encoding brief relationships (linking phrases) between pairs of concepts forming propositions.

This paper uses concept maps for presenting the very higher general structure of a studied discipline. The concept map is used by learners as a top level reference material that may be consulted. This structured high level overview of the discipline is aimed to expand the mental outlook for learners in general by exercising this ability on a particular discipline.

## 2 Concept maps

Concept mapping may be used as a tool for understanding, collaborating, validating, and integrating curriculum content that is designed to develop specific competencies. Concept mapping, a tool originally developed to facilitate student learning by organizing key and supporting concepts into visual frameworks, can also facilitate communication among faculty and administrators about curricular structures, complex cognitive frameworks, and competency-based learning outcomes. To validate the relationships among the competencies articulated by specialized accrediting agencies, certification boards, and professional associations, faculty may find the concept mapping tool beneficial in illustrating relationships among, approaches to, and compliance with competencies [4].

Concept maps are also effective in identifying both valid and invalid ideas held by students, and this will be discussed further in another section. They can be as effective as more time-consuming clinical interviews for identifying the relevant knowledge a learner possesses before or after instruction [7].

Recent decades have seen an increasing awareness that the adoption of refined procedures of evaluation contributes to the enhancement of the teaching/learning process. In the past, the teacher's evaluation of the pupil was expressed in the form of a final mark given on the basis of a scale of values determined both by the culture of the institution and by the subjective opinion of the examiner. This practice was rationalized by the idea that the principal function of school was selection - i.e. only the most fully equipped (outstanding) pupils were worthy of continuing their studies and going on to occupy the most important positions in society.

Ausubel [1] made the very important distinction between rote learning and meaningful learning. Meaningful learning requires three conditions: 1. The material to be learned must be conceptually clear and presented with language and examples relatable to the

learner's prior knowledge. Concept maps can be helpful to meet this condition, both by identifying large general concepts held by the leaner prior to instruction of more specific concepts, and by assisting in the sequencing of learning tasks though progressively more explicit knowledge that can be anchored into developing conceptual frameworks; 2. The learner must possess relevant prior knowledge. This condition can be met after age 3 for virtually any domain of subject matter, but it is necessary to be careful and explicit in building concept frameworks if one hopes to present detailed specific knowledge in any field in subsequent lessons. We see, therefore, that conditions (1) and (2) are interrelated and both are important; 3. The learner must choose to learn meaningfully. The one condition over which the teacher or mentor has only indirect control is the motivation of students to choose to learn by attempting to incorporate new meanings into their prior knowledge, rather than simply memorizing concept definitions or propositional statements or computational procedures. The indirect control over this choice is primarily in instructional strategies used and the evaluation strategies used. Instructional strategies that emphasize relating new knowledge to the learner's existing knowledge foster meaningful learning. Evaluation strategies that encourage learners to relate ideas they possess with new ideas also encourage meaningful learning. Typical objective tests seldom require more than rote learning [9].

According to this approach, the responsibility for failure at school was to be attributed exclusively to the innate (and, therefore, unalterable) intellectual capacities of the pupil. The learning/ teaching process was, then, looked upon in a simplistic, linear way: the teacher transmits (and is the repository of) knowledge, while the learner is required to comply with the teacher and store the ideas being imparted. [5].Usage of concept maps may be very useful for students when starting to learn about a subject. The concept map may bring valuable general overlook of the subject for the whole period of study. It may be advisable that at the very first meeting of students with the subject to include a concept map of the subject.

# 3 Data structures and algorithms / Binary search tree concept map

The experimental concept map was used on Tesys e-Learning platform [6]. On this platform there was set an Algorithms and Data Structures discipline. The tests were performed for five chapters: Simply/Double Linked Lists, Binary Search Trees, Height Balanced Trees, B Trees and Graphs.

The concepts are presented in table 1.

| ID | Concept | ID | Concept |
|---|---|---|---|
| C1 | BST | C9 | Right child |
| *C2* | Dynamic Structure | C10 | No child |
| *C3* | Node(s) | C11 | Root |
| *C4* | Traversed | C12 | Leaf |
| *C5* | Key | C13 | Preorder |
| *C6* | Parent | C14 | Inorder |
| C7 | Child | C15 | Postorder |
| *C8* | Left child | C16 | Ascending order |

Table 1: List of concepts

The concept map for Binary Search Trees is presented in figure 1. It contains 16 concepts, 11 linking phrases and 16 propositions.

The list of propositions with two concepts and one linking phrase is presented in table 2. The list of propositions with three concepts and two linking phrases is presented in table 3.

Once the concept map has been built the general graph of the each chapter may be created. In this graph, each proposition will become an edge that links the first concept and the last concept. The domain knowledge expert will assign a weight for each edge. While the students answers questions the number of correct and wrong answers will determine the knowledge weight of that edge.

There is one proposition with five concepts and four linking phrases: "BST" may be "Traversed" in "Preorder" determines "Key" in "Ascending Order". The concepts are bolded and put between quotation marks, while linking phrases are italic and underlined.

| Id | Concept | Linking phrase | Concept |
|---|---|---|---|
| P1 | BST | is | Dynamic Structure |
| P2 | BST | is made of | Node(s) |
| P3 | Node | has | key |
| P4 | Node | is | Parent |
| P5 | Node | is | Child |
| P6 | Parent | is greater than | Left child |
| P7 | Parent | is smaller than | Right child |
| P8 | Node | may have | Left child |
| P9 | Node | may have | Right child |
| P10 | Node | may have | No child |

Table 2: List of propositions with two concepts and one linking phrase.

| Id | C | LP | C | LP | C |
|---|---|---|---|---|---|
| P11 | Node | without | parent | is | root |
| P12 | BST | may be | traversed | in | preorder |
| P13 | BST | may be | traversed | in | inorder |
| P14 | BST | may be | traversed | in | postorder |

Table 3: List of propositions with three concepts and two linking phrases.

Once the Conncept Map has been set up the professor has to set up the quiz questions for the chapter. For each edge in the map it will correspond a certain number of quiz questions. There is no specification regarding the number of quiz questions but a minimum (e.g. five) number is still required. The activity performed by learner is monitored and stored. The Calculus engine will reconstruct an Annotated Concept Map which will present to the learner the current status of his knowledge level at Concept level. In this way, the learner will have an exact outlook of his knowledge level regarding that chapter. This information may be used by learners and course managers to obtain important information regarding the reached knowledge level.

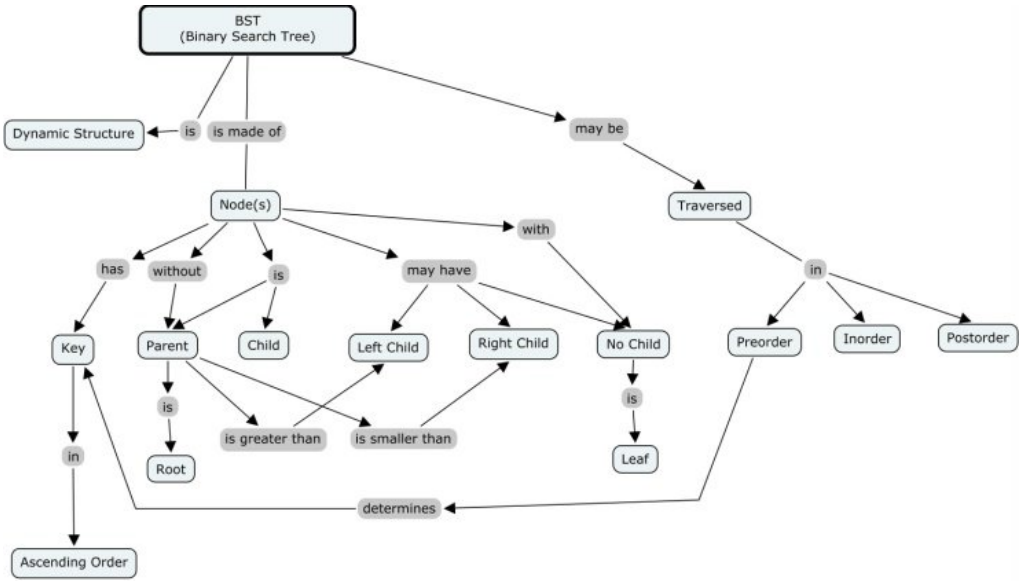



Figure 1: Binary Search Tree Concept Map

Defining a metric for computing the accumulated knowledge is accomplished. If

**W** is the weight of the edge,
**CA** is the number of correct answers,
**WA** is the number of wrong answers,
**N** is the number of questions

Than

**KW** is the knowledge weight of the edge

$$KW = \frac{CA - WA}{N} \frac{1}{W} * 100$$

Under these circumstances the knowledge weight may also be negative. At any time there may be estimated the overall knowledge level of the learner as the ratio between overall knowledge weight and overall weight.

| Proposition | Weight | Number of questions |
|---|---|---|
| P1 | 10 | 8 |
| P2 | 4 | 7 |
| P3 | 7 | 6 |
| P4 | 3 | 5 |
| P5 | 2 | 7 |

Table 4: Sample setup of BST chapter.

| Proposition (weight) | No. of questions | CA | WA | KW (%) |
|---|---|---|---|---|
| P1 (10) | 8 | 3 | 2 | 1.25 |
| P2 (4) | 7 | 4 | 2 | 7.14 |
| P3 (7) | 6 | 1 | 3 | -4.76 |
| P4 (3) | 5 | 3 | 1 | 13.3 |

Table 5: Sample values for learner's associated graph.

| Chapter | Weight |
|---|---|
| Simply/Double Linked Lists | 15 |
| Binary Search Trees | 15 |
| Height Balanced Trees | 25 |
| B Trees | 25 |
| Graphs | 20 |

Table 6: Sample weights assigned to chapters.

Table 4 presents a sample of the setup of the Binary SearchTrees chapter. Table 5 presents a sample of the of the values of the Learner's Associated Graph corresponding to BST chapter.

The values from table five are marked in an Annotated Concept Map that is finally presented to the learner. The Annotated Concept Map is the final outcome of the Decision Support System and is supposed to guide the learner regarding the necessary future efforts.

Table 6 presents the weights of chapters as they were assigned by the domain expert.

# 4 General infrastructure

For running such a process the e-Learning infrastructure must have some characteristics. The process is designed to run at chapter level. This means a discipline needs to be partitioned into chapters. The chapter has to have assigned a concept map which may consist of about 20 concepts. Each concept has assigned a document and a set of quiz questions. Each concept and each quiz has a weight, depending of its importance in the hierarchy. It is supposed that the platform run for some time such that there is a history of performed activities for a large number of learners. Figure 2 presents a general e-Learning infrastructure for a discipline.

Once a course manager has been assigned a discipline, he has to set up its chapters by specifying their names and their associated concept maps. For each concept managers have the possibility of setting up one document and one pool of questions. When the discipline

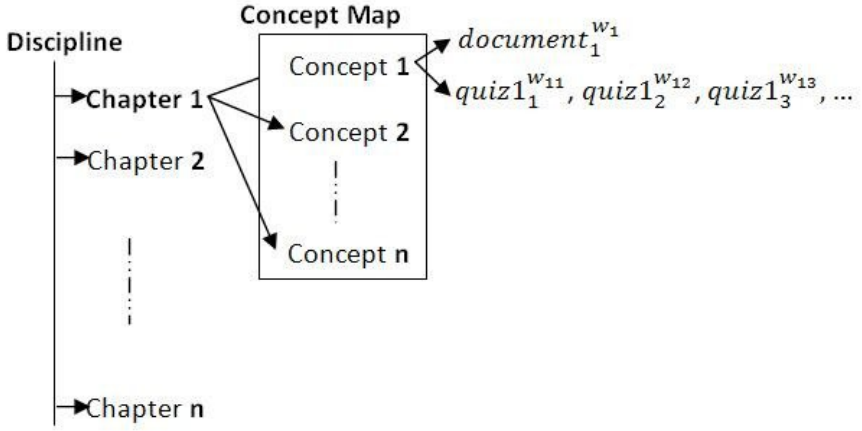



Figure 2: General structure of a discipline

is fully set, the learning process may start for learners. Any opening of the document and any test quiz that is taken by a learner is registered. The business logic of document retrieval tool wills using these data in order to determine the moment when it is able to determine the document (or the documents) that are considered to need more attention from the learner. The course manager specifies the number of questions that will be randomly extracted for creating a test or an exam.

Let us suppose that for a chapter the professor created 50 test quizzes and he has set to 5 the number of quizzes that are randomly withdrawn for testing and 15 the number of quizzes that are randomly withdrawn for final exam. It means that when a student takes a test from this chapter 5 questions from the pool of test question are randomly withdrawn. When the student takes the final examination at the discipline from which the chapter is part, 15 questions are randomly withdrawn. This manner of creating tests and exams is intended to be flexible enough for the professor.

# 5 Experimental results

The setup and experimental results were obtained on Tesys e-Learning platform [6]. On this platform there was set Algorithms and Data Structures discipline. The tests were performed for the chapter named Binary Search Trees. Figure 1 presents the concept map for Binary Search Trees chapter.

For each concept there is associated a set of quizzes. As learners start answering quizzes the table data regarding performed activities is continuously updated. Table 1 presents the structure of recorded activities. Table 2 presents a snapshot for the activities table. The total number of students of 400 and the total number of problems is 200. The time during which results have been taken is six months.

In order to obtain relevant results we pruned noisy data. We considered that students for which the number of taken tests or the time spent for testing is close to zero are not interesting for our study and degrade performance and that is why all such records were deleted. After this step there remained only 268 instances.

| Event type | Details |
|---|---|
| login | start time and duration |
| view document | start time, duration, concept |
| self test | start time, duration, quizzes |
| messages | start time, duration, receiver categories |

Table 7: Structure of recorded activities

| User id | Event type | Details |
|---|---|---|
| 10 | login | 10.10.2009 11:20, 25 min |
| 11 | view document | 10.10.2009 11:21, 10 min, concept11 |
| 12 | self test | 10.10.2009 11:31, 3 min, concept11 |
| 13 | messages | 10.10.2009 11:34, 5 min, prof3, prof4 |

Table 8: Sample data from activities table

After data has been recorded, there was run a clustering process on learners. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. Basically, for our platform we create clusters of learners based on their activity. There are many clustering methods in the literature: partitioning methods, hierarchical methods, density-based methods, grid-based methods or model-based methods. From all of these we chose to have a closer look on partitioning methods.

Given a database of n objects and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion, often called similarity function, such as distance, so that objects within a cluster are "similar", whereas the objects of different clusters are "dissimilar" in terms of database attributes. So, the first step is to define a list of attributes that may be representative for modelling and characterizing student's activity.

Among the attributes there may be:

- the number of logins,

- the number of taken tests,
- the average grade for taken tests

The classic k-means algorithm is a very simple method of creating clusters. Firstly, it is specified how many clusters are being thought: this is the parameter k. Then k points are chosen at random as cluster centres. Instances are assigned to their closest cluster centre according to the ordinary Euclidean function. Next the centroid, or the mean, of all instances in each cluster is calculated – this is the "means" part. These centroids are taken to be the new centre values for their respective clusters. Finally, the whole process is repeated with the new cluster centres. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at each point the cluster centres have stabilized and will remain the same thereafter.

From a different perspective for a cluster there may be computed the following parameters:

- $\mu = \frac{x_1 + x_2 + ... + x_n}{n}$ , the means
- $\sigma = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_n - \mu)^2}{n-1}$ ,the standard deviation
- p, the probability

The sum of all probabilities for all clusters is 1. If we know which of the distributions each instance came from, finding the parameters is easy. On the other hand, if the parameters are known finding the probabilities that a given instance comes from each distribution is easy. Given an instance x, the probability that it belongs to cluster A is:

$$\Pr[A \mid x] = \frac{\Pr[x \mid A] - \Pr[A]}{\Pr[x]} = \frac{f(x; \mu_A, \sigma_A) p_A}{\Pr[x]}$$

where $f(x; \mu_A, \sigma_A)$ is the normal distribution function for cluster A, that is:

$$f(x; \mu_A, \sigma_A) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The EM algorithm takes into consideration that we know neither of these things: not the distribution that each training instance came from, nor the parameters μ, σ or the probability. So, we adopt the procedure used for the k-means clustering algorithm and iterate. Start with initial guess for the five parameters, use them to calculate the cluster probabilities for each instance, use these probabilities to reestimate the parameters, and repeat. This is called the EM algorithm for "expectation-maximization". The first step, the calculation of cluster probabilities (which are the "expected" class values) is "expectation"; the second, calculation of the distribution parameters is "maximization" of the likelihood of the distributions given the data.

A slight adjustment must be made to the parameter estimation equations to account for the fact that it is only cluster probabilities, not the clusters themselves, that are known for each instance. These probabilities just act like weights. If wi is the probability that instance i belongs to cluster A, the mean and standard deviation are:

$$\mu_A = \frac{w_1 x_1 + w_2 x_2 + ... + w_n x_n}{w_1 + w_2 + ... + w_n}$$

$$\sigma_A^2 = \frac{w_1 (x_1 - \mu)^2 + w_2 (x_2 - \mu)^2 + ... + w_n (x_n - \mu)^2}{w_1 + w_2 + ... + w_n}$$

where $x_i$ are all the instances, not just those belonging to cluster A. Technically speaking, this is a "maximum likelihood" estimator of the variance.

Now we shall discuss about how iterating process is terminated. The k-means algorithm stops when the classes of instances don't change from one iteration to the next – a "fixed point" has been reached. In the EM algorithm things are not quite so easy: the algorithm converges toward a fixed point but never actually gets there. But we can see how close it is by calculating the overall likelihood that the data came from this dataset, given the values for the parameters (mean, standard deviation and probability). This overall likelihood is obtained by multiplying the probabilities of the individual instances *i*:

$$\prod_i \left( p_A \Pr[x_i \mid A] + p_B \Pr[x_i \mid B] \right)$$

Where the probabilities given the clusters A and B are determined from the normal distribution function $f(x; \mu, \sigma)$. This overall likelihood is a measure of the "goodness" of clustering and increases at each iteration of the EM algorithm. Again, there is a technical difficulty with equating the probability of a particular value of x with $f(x; \mu, \sigma)$, and in this case the effect does not disappear because no probability normalization operation is applied. The upshot is that the likelihood expression above is not a probability and does not necessarily lie between zero and one: nevertheless, its magnitude still reflects the quality of the clustering. In practical implementations its logarithm is calculated instead: this is done by summing the logs of individual components, avoiding all the multiplications. But the overall conclusion still holds: iterate until the increase in log-likelihood becomes negligible. For example, a practical implementation might iterate until the difference between successive values of log-likelihood is less than 10-10 for ten successive iterations. Typically, the log-likelihood will increase very sharply over the first few iterations and then converge rather quickly to a point that is virtually stationary.

Although the EM algorithm is guaranteed to converge to a maximum, this is a local maximum and may not necessarily be the same as the global maximum. For a better chance of obtaining the global maximum, the whole procedure should be repeated several times, with different initial guess for the parameter values. The overall log-likelihood figure can be used to compare the different final configuration obtained: just choose the largest of the local maxima.

The EM algorithm is implemented in Weka package [10] and needs the input data to be in a custom format

called *arff*. Under these circumstances we have developed an offline Java application that queries the platform's database and crates the input data file called *activity.arff*. This process is automated and is driven by a *properties* file in which there is specified what data will lay in *activity.arff* file.

The most important step in this procedure is the attribute selection and the granularity of their nominal values. The number of attributes and their meaning has a crucial importance for the whole process since irrelevant attributes may degrade classification performance in sense of relevance. On the other hand, the more attributes we have the more time the algorithm will take to produce a result. Domain knowledge and of course common sense are crucial assets for obtaining relevant results.

Running the EM algorithm created four clusters. The procedure clustered 40 instances (15%) in cluster 0, 83 instances (31%) in cluster 1, 112 instances (42%) in cluster 2 and 33 instances (12%) in cluster 3. The final step is to check how well the model fits the data by computing the likelihood of a set of test data given the model. Weka measures goodness-of-fit by the logarithm of the likelihood, or log-likelihood: and the larger this quantity, the better the model fits the data. Instead of using a single test set, it is also possible to compute a cross validation estimate of the log-likelihood. For our instances the value of the log-likelihood is -3.75 which represents a promising result in the sense that instances (in our case students) may be classified in three disjoint clusters based on their activity.

# 6 Conclusions and future work

Tesys e-Learning platform has been designed such that on-line testing activities may be performed as they were set up by course managers.

A Concept Map for a Binary Search Trees chapter as well as for each chapter of Algorithms and Data Structures course has been created. The Concept maps have been the staring point in creating the sets of quiz questions. Each quiz question refers to a certain proposition from the concept map.

After the setup has been put in place, the learners started using the platform. At request, from the general concept map there was derived the learner's associated concept map. This concept map provides important information regarding the level of knowledge reached by learner.

The business logic computes the knowledge of the student regarding the chapter as a knowledge weight and regarding the discipline as a percentage of covered concepts. This weight is computed as a function of proposition's weight, number of questions assigned to that proposition, the number of correct answered questions and number of wrong answered questions.

We have created a procedure of data analysis which may provide interesting conclusions regarding the classification of students from an e-learning platform.

The platform was developed, it is currently running and has built in capabilities of monitoring students testing activities. An off-line application was developed for creating the input data files that are analysed. Data analysis is done using EM clustering algorithm implemented by Weka system. One of the main goals is clustering of students. This may lead to important information regarding the learners for which expanding mental outlook performed better.

Student's clustering may have a predictive value in the sense that from the performed testing activities a student has made there may be pulled conclusions about his learning proficiency. On the other hand, platform's characterization may have as result an estimation of the capability of an e-learning system to grade and order students according to their accumulated knowledge. This analysis is critical for having as conclusion that a system can expand mental outlook.

This whole mechanism represents the functionality of a decision support system that runs along the Tesys e-Learning platform.

Whenever needed, the learner may study his own associated concept map at discipline level or at chapter level. This functionality has as main benefit expanding the mental outlook in general and for studied discipline in particular.

As future works, a computational framework may be created which estimates the proficiency of the learners after using the concept map as structuring facility. This kind of analysis may give important information concerning the improvement of the used concept map.

# References

[1] Novak, J. D. (1977). *A Theory of Education*. Cornell University Press, Ithaca, NY.

[2] Novak, J. D. (1998). *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Lawrence Erlbaum Associates, Mahwah, NJ.

[3] Novak, J. D. and Gowin, D. B. (1984). *Learning How to Learn*. Cambridge University Press, New York.

[4] McDaniel,E., Roth, B., and Miller, M. (2007) Concept Mapping as a Tool for Curriculum Design. *Issues in Informing Science and Information Technology*, Vol. 4.

[5] Vecchia, L., Pedroni, M. (2007). Concept Maps as a Learning Assessment Tool. *Issues in Informing Science and Information Technology*, Vol. 4.

[6] Burdescu, D.D., Mihăescu, M.C., 2006. Tesys: e-Learning Application Built on a Web Platform. In *Proceedings of International Joint Conference on e-Business and Telecommunications*. INSTICC Press, Setubal, Portugal,, pp. 315-318

[7] Edwards, J. And Fraser, K. (1983). Concept maps as reflections of conceptual understanding. *Research in Science Education*, 13, pp. 19-26.

[8] Ausubel, D. P., Novak, J. D., and Hanesian, H., (1978). *Educational Psychology: A Cognitive View (2nd ed.)*. Holt, Rinehart and Winston, New York.

[9] Holden, C. (1992). Study flunks science and math tests. *Science Education*, 26, 541.

[10] WEKA, www.cs.waikato.ac.nz/ml/weka

# CONTENTS OF *Informatica* Volume 34 (2010) pp. 1–544

## Papers


ALBERTI, G. & , J. KLEIBER. 2010. Grammar of ReALIS and the Implementation of its Dynamic Interpretation. Informatica 34:103–110.

ÁVILA-ARGÜELLES, R. & , H. CALVO, A. GELBUKH, S. GODOY-CALDERÓN. 2010. Assigning Library of Congress Classification Codes to Books Based Only on their Titles. Informatica 34:77–84.

BÉCHET, N. & , J. CHAUCHÉ, V. PRINCE, M. ROCHE. 2010. Corpus and Web: Two Allies in Building and Automatically Expanding Conceptual Classes. Informatica 34:279–286.

BOYTCHEVA, S. & , I. NIKOLOVA, E. PASKALEVA, G. ANGELOVA, D. TCHARAKTCHIEV, N. DIMITROVA. 2010. Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records. Informatica 34:269–278.

CARCHIOLO, V. & , A. LONGHEU, M. MALGERI, G. MANGIONI. 2010. Context-based Global Expertise in Recommendation Systems. Informatica 34:409–417.

CHIFARI, A. & , G. CHIAZZESE, L. SETA, G. MERLO, S. OTTAVIANO, M. ALLEGRA. 2010. A Reflection on Some Critical Aspects of Online Reading Comprehension. Informatica 34:491–495.

CHU, Y.-M. & , L.-L. HSU, J.-T. YANG. 2010. A Study of Analysing IT Digital Coping Strategies. Informatica 34:169–179.

DAN, B.D. & , M.M. CRISTIAN, I.M. COSTEL, B. IONUT, L. BOGDAN. 2010. Expanding Mental Outlook with the Help of Concept Maps. Informatica 34:535–540.

DAUDARAVICIUS, V. & . 2010. Automatic Identification of Lexical Units. Informatica 34:85–91.

DEBIAO, H. & , C. JIANHUA, H. JIN. 2010. Cryptanalysis of a Simple Three-party Key Exchange Protocol. Informatica 34:337–339.

EKBAL, A. & , S. BANDYOPADHYAY. 2010. Named Entity Recognition Using Appropriate Unlabeled Data, Post-processing and Voting. Informatica 34:55–76.

FOMICHOV, V.A. & . 2010. Theory of K-representations as a Comprehensive Formal Framework for Developing a Multilingual Semantic Web. Informatica 34:287–296.

GARCÍA-HERNÁNDEZ, R.A. & , J.FCO. MARTÍNEZ-TRINIDAD, J.A. CARRASCO-OCHOA. 2010. Finding Maximal Sequential Patterns in Text Document Collections and Single Documents. Informatica 34:93–102.

GHIDUK, A.S. & , M.R. GIRGIS. 2010. Using Genetic Algorithms and Dominance Concepts for Generating Reduced Test Data. Informatica 34:377–386.

HE, R. & , B. QIN, T. LIU, S. LI. 2010. Cascaded Regression Analysis Based Temporal Multi-document Summarization. Informatica 34:119–124.

KICHKAYLO, T. & , T. RYUTOV, M.D. OROSZ, R. NECHES. 2010. Planning to Discover and Counteract Attacks. Informatica 34:151–160.

KIM, K. & , I. YIE, D. NYANG. 2010. On the Security of Two Group Signature Schemes with Forward Security. Informatica 34:229–234.

KRAVARI, K. & , E. KONTOPOULOS, N. BASSILIADES. 2010. Trusted Reasoning Services for Semantic Web Agents. Informatica 34:429–440.

KRYŠTOF, J. & . 2010. An LPGM method: Platform Independent Modeling and Development of Graphical User Interface. Informatica 34:353–367.

LEDENEVA, Y. & , G. SIDOROV. 2010. Recent Advances in Computational Linguistics. Informatica 34:3–18.

LINTEAN, M.C. & , V. RUS. 2010. Paraphrase Identification Using Weighted Dependencies and Word Semantics. Informatica 34:19–28.

LLORET, E. & , M. PALOMAR. 2010. Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. Informatica 34:29–35.

MA, H. & , R. NOACK, K.-D. SCHEWE, B. THALHEIM. 2010. Using Meta-Structures in Database Design. Informatica 34:387–403.

MARINČIČ, D. & . 2010. Parsing with Intraclausal Coordination and Clause Detection. Informatica 34:243–244.

MEGHANATHAN, N. & . 2010. A Simulation Study on the Impact of Mobility Models on the Network Connectivity, Hop Count and Lifetime of Routes for Ad hoc Networks. Informatica 34:199–213.

NGOMO, A.-C. N. & . 2010. Low-Bias Extraction of Domain-Specific Concepts. Informatica 34:119–124.

NOVALIJA, I. & , D. MLADENIČ. 2010. Ontology Extension Towards Analysis of Business News. Informatica 34:517–522.

OBITKO, M. & , P. VRBA, V. MAŘÍK, M. RADAKOVIČ, P. KADERA. 2010. Applications of Semantics in Agent-Based

Manufacturing Systems. Informatica 34:315–330.

PARALIČ, J. & , F. BABIČ, J. WAGNER, P. BEDNÁR, M. PARALIČ. 2010. KP-Lab System for the Support of Collaborative Learning and Working Practices, Based on Trialogical Learning. Informatica 34:341–351.

PAVLIN, G. & , M. KAMERMANS, M. SCAFES. 2010. Dynamic Process Integration Framework: Toward Efficient Information Processing in Complex Distributed Systems. Informatica 34:477–490.

PETCU, D. & , S. PANICA, M. NEAGUL, M. FRÎNCU, D. ZAHARIE, R. CIORBA, A. DINIŞ. 2010. Earth Observation Data Processing in Distributed Systems. Informatica 34:463–476.

PETRIČ, I. & . 2010. Text Mining for Discovering Implicit Relationships in Biomedical Literature. Informatica 34:241–242.

PIPER, I. & , D. KEEP, T. GREEN, I. ZHANG. 2010. Application of Microsimulation to the Modelling of Epidemics and Terrorist Attacks. Informatica 34:141–150.

POPESCU, E. & , C. BADICA, L. MORARET. 2010. Accommodating Learning Styles in an Adaptive Educational System. Informatica 34:451–462.

POŽENEL, M. & , V. MAHNIČ, M. KUKAR. 2010. Separating Interleaved HTTP Sessions Using a Stochastic Model. Informatica 34:191–197.

RAHIM, A. & , F. BIN MUHAYA, Z.S. KHAN, M.A. ANSARI, M. SHER. 2010. Enhanced Relevance-Based Approach for Network Control. Informatica 34:215–218.

RATTADILOK, P. & . 2010. An Investigation and Extension of a Hyper-heuristic Framework. Informatica 34:523–534.

RUPNIK, J. & , M. GRČAR, T. ERJAVEC. 2010. Improving Morphosyntactic Tagging of Slovene Language through Metatagging. Informatica 34:161–168.

SETH, R. & , R. KAPOOR, H. AL-QAHERI, S. SANYAL. 2010. Piecemeal Journey to 'HALCYON' World of Pervasive Computing: From Past Progress to Future Challenges. Informatica 34:181–190.

SHAO, Z. & . 2010. Multisignature Scheme Based on Discrete Logarithms in the Plain Public Key Model. Informatica 34:509–515.

ŠTAJNER, T. & , D. RUSU, L. DALI, B. FORTUNA, D. MLADENIČ, M. GROBELNIK. 2010. A Service Oriented Framework for Natural Language Text Enrichment. Informatica 34:307–313.

STANESCU, L. & , G. MIHAI, D. BURDESCU, M. BREZOVAN, C.S. SPAHIU. 2010. A Software System for Viewing and Querying Automatically Generated Topic Maps in the E-learning Domain. Informatica 34:441–450.

SVETEL , I. & , M. PEJANOVIĆ. 2010. The Role of the Semantic Web for Knowledge Management in the Construction Industry. Informatica 34:331–336.

TAYLOR, M.E. & , C. KIEKINTVELD, C. WESTERN, M. TAMBE. 2010. A Framework for Evaluating Deployed Security Systems: Is There a Chink in your ARMOR?. Informatica 34:129–139.

TÉLLEZ-VALERO, A. & , M. MONTES-Y-GÓMEZ, L. VILLASEÑOR-PINEDA, A. PEÑAS-PADILLA. 2010. Towards Multi-Stream Question Answering Using Answer Validation. Informatica 34:119–124.

VEMULAPALLI, S. & , X. LUO, J.F. PITRELLI, I. ZITOUNI. 2010. Using Bagging and Boosting Techniques for Improving Coreference Resolution. Informatica 34:111–118.

WANG, H. & , X. JIANG, L.-T. CHIA, A.-H. TAN. 2010. Wikipedia2Onto — Building Concept Ontology Automatically, Experimenting with Web Image Retrieval. Informatica 34:297–306.

WENG, Y. & , C. HU, X. ZHANG. 2010. BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis. Informatica 34:419–428.

XIA, Z. & , S. LU, J. LI, J. TANG. 2010. Enhancing DDoS Flood Attack Detection via Intelligent Fuzzy Logic. Informatica 34:497–507.

YASUDA, H. & . 2010. A Risk Management System to Oppose Cyber Bullying in High School: Warning System with Leaflets and Emergency Staffs. Informatica 34:235–239.

YOU, L. & , J. ZENG. 2010. Fast Scalar Multiplications on Hyperelliptic Curve Cryptosystems. Informatica 34:219–228.

ZHANG, X. & , Z. TANG, J. YU, M. GUO. 2010. A Fast Convex Hull Algorithm for Binary Image. Informatica 34:369–376.

## Editorials

LEDENEVA, Y. & , G. SIDOROV. 2010. Editor's Introduction to the Special Issue on System Modeling and Transformation Principles. Informatica 34:1–1.

KIEKINTVELD, C. & , J. MARECKI, P. PARUCHURI, K. SYCARA. 2010. Editor's Introduction to the Special Issue on Quantitative Risk Analysis Techniques for Security Applications. Informatica 34:127–127.

FOMICHOV, V.A. & . 2010. Editor's Introduction to the Special Issue on Semantic Informational Technologies. Informatica 34:267–267.

BADICA, C. & . 2010. Editor's Introduction to the Special Issue on E-Service Intelligence. Informatica 34:407–407.

# 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 S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean 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 LaTeX 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: .....

Title and Profession (optional): .....

.....

Home Address and Telephone (optional): .....

.....

Office Address and Telephone (optional): .....

.....

E-mail Address (optional): .....

Signature and Date: .....

## Informatica WWW:

## http://www.informatica.si/

**Referees from 2008 on:**

Ajith Abraham, Siby Abraham, Renato Accornero, Hameed Al-Qaheri, Gonzalo Alvarez, Wolfram Amme, Nicolas Anciaux, Rajan Arora, Costin Badica, Zoltán Balogh, Andrea Baruzzo, Norman Beaulieu, Paolo Bellavista, Zbigniew Bonikowski, Marco Botta, Pavel Brazdil, Andrej Brodnik, Ivan Bruha, Wray Buntine, Yunlong Cai, Juan Carlos Cano, Tianyu Cao, Norman Carver, Marc Cavazza, Jianwen Chen, LM Cheng, Chou, Cheng-Fu, Girija Chetty, G. Chiola, Yu-Chiun Chiou, Ivan Chorbev, Shauvik Roy Choudhary, Lawrence Chung, Jean-Noël Colin, Jinsong Cui, Alfredo Cuzzocrea, Gunetti Daniele, Grégoire Danoy, Manoranjan Dash, Paul Debevec, Fathi Debili, Carl James Debono, Joze Dedic, Abdelkader Dekdouk, Bart Demoen, Sareewan Dendamrongvit, Tingquan Deng, Gaël Dias, Ivica Dimitrovski, Jana Dittmann, Simon Dobrišek, Quansheng Dou, Jeroen Doumen, Dejan Drajic, Jozo Dujmovic, Umut Riza Ertürk, Ling Feng, YiXiong Feng, Andres Flores, Vladimir A. Fomichov, Stefano Forli, Massimo Franceschet, Alberto Freitas, Chong Fu, Gabriel Fung, Andrea Gambarara, Matjaž Gams, Juan Garbajosa, David S. Goodsell, Jaydeep Gore, Zhi-Hong Guan, Donatella Gubiani, Bidyut Gupta, Marjan Gusev, Zhu Haiping, Juha Hyvärinen, Dino Ienco, Natarajan Jaisankar, Imad Jawhar, Yue Jia, Ivan Jureta, Džani Juričić, Zdravko Kačič, Boštjan Kaluža, Dimitris Kanellopoulos, Rishi Kapoor, Daniel S. Katz, Mustafa Khattak, Ivan Kitanovski, Tomaž Klobučar, Ján Kollár,, Peter Korošec, Agnes Koschmider, Miroslav Kubat, Chi-Sung Laih, Niels Landwehr, Andreas Lang, Yung-Chuan Lee, John Leggett, Aleš Leonardis, Guohui Li, Guo-Zheng Li, Jen Li, Xiang Li, Xue Li, Yinsheng Li, Yuanping Li, Lejian Liao, Huan Liu, Xin Liu, Hongen Lu, Mitja Luštrek, Inga V. Lyustig, Matt Mahoney, Dirk Marwede, Andrew McPherson, Zuqiang Meng, France Mihelič, Nasro Min-Allah, Vojislav Misic, Mihai L. Mocanu, Jesper Mosegaard, Marta Mrak, Yi Mu, Josef Mula, Phivos Mylonas, Marco Di Natale, Pavol Navrat, Nadia Nedjah, R. Nejabati, Wilfred Ng, Zhicheng Ni, Fred Niederman, Omar Nouali, Franc Novak, Petteri Nurmi, Barbara Oliboni, Matjaž Pančur, Gregor Papa, Marcin Paprzycki, Marek Paralič, Byung-Kwon Park, Gert Schmeltz Pedersen, Torben Bach Pedersen, Zhiyong Peng, Ruggero G. Pensa, Dana Petcu, Macario Polo, Victor Pomponiu, Božidar Potočnik, S. R. M. Prasanna, HaiFeng Qian, Lin Qiao, Jean-Jacques Quisquater, Jean Ramaekers, Jan Ramon, Wilfried Reimche, Juan Antonio Rodriguez-Aguilar, Pankaj Rohatgi, Wilhelm Rossak, Sattar B. Sadkhan, Khalid Saeed, Motoshi Saeki, Evangelos Sakkopoulos, M. H. Samadzadeh, MariaLuisa Sapino, Piervito Scaglioso, Walter Schempp, Barabara Koroušić Seljak, Mehrdad Senobari, Subramaniam Shamala, Heung-Yeung Shum, Tian Song, Andrea Soppera, Alessandro Sorniotti, Liana Stanescu, Martin Steinebach, Xinghua Sun, Marko Robnik,vSikonja, Jurij,vSilc, Carolyn Talcott, Camillo J. Taylor, Drago Torkar, Christos Tranoris, Denis Trček, Katarina Trojacanec, Mike Tschierschke, Filip De Turck, Aleš Ude, Alessia Visconti, Petar Vračar, Valentino Vranić, Chih-Hung Wang, Huaqing Wang, Hao Wang, YunHong Wang, Sigrid Wenzel, Woldemar Wolynski, Allan Wong, Stefan Wrobel, Konrad Wrona, Bin Wu, Xindong Wu, Li Xiang, Yan Xiang, Di Xiao, Fei Xie, Yuandong Yang, Chen Yong-Sheng, Jane Jia You, Ge Yu, Mansour Zand, Dong Zheng, Jinhua Zheng, Albrecht Zimmermann, Blaz Zupan, Meng Zuqiang

# *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 2010 (Volume 34) 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: ACM Digital Library, 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.*

**Volume 34 Number 4 December 2010** **ISSN 0350-5596**

# *Informatica*

## An International Journal of Computing and Informatics


| | | |
|---|---|---|
| Editor's Introduction to the Special Issue on E-Service Intelligence | C. Badica | **407** |
| Context-based Global Expertise in Recommendation Systems | V. Carchiolo, A. Longheu, M. Malgeri, G. Mangioni | **409** |
| BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis | Y. Weng, C. Hu, X. Zhang | **419** |
| Trusted Reasoning Services for Semantic Web Agents | K. Kravari, E. Kontopoulos, N. Bassiliades | **429** |
| A Software System for Viewing and Querying Automatically Generated Topic Maps in the E-learning Domain | L. Stanescu, G. Mihai, D. Burdescu, M. Brezovan, C.S. Spahiu | **441** |
| Accommodating Learning Styles in an Adaptive Educational System | E. Popescu, C. Badica, L. Moraret | **451** |
| Earth Observation Data Processing in Distributed Systems | D. Petcu, S. Panica, M. Neagul, M. Frîncu, D. Zaharie, R. Ciorba, A. Diniş | **463** |
| Dynamic Process Integration Framework: Toward Efficient Information Processing in Complex Distributed Systems | G. Pavlin, M. Kamermans, M. Scafes | **477** |

End of Special Issue / Start of normal papers

| | | |
|---|---|---|
| A Reflection on Some Critical Aspects of Online Reading Comprehension | A. Chifari, G. Chiazzese, L. Seta, G. Merlo, S. Ottaviano, M. Allegra | **491** |
| Enhancing DDoS Flood Attack Detection via Intelligent Fuzzy Logic | Z. Xia, S. Lu, J. Li, J. Tang | **497** |
| Multisignature Scheme Based on Discrete Logarithms in the Plain Public Key Model | Z. Shao | **509** |
| Ontology Extension Towards Analysis of Business News | I. Novalija, D. Mladenič | **517** |
| An Investigation and Extension of a Hyper-heuristic Framework | P. Rattadilok | **523** |
| Expanding Mental Outlook with the Help of Concept Maps | B.D. Dan, M.M. Cristian, I.M. Costel, B. Ionut, L. Bogdan | **535** |

Informatica **34** (2010) Number 4, pp. 407–544