Volume 35 Number 4 December 2011 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Advances in Semantic Information Retrieval

Guest Editor:

Vitaly Klyuev

Maxim Mozgovoy

EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive 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

Advances in Semantic Information Retrieval

Recent advances in semantic technologies have resulted in methods and tools that allow creating and managing domain knowledge. They influence the ways and forms of document representation in computer memory; they define approaches to analyze documents, and techniques to mine and retrieve knowledge. Searching for video, voice and speech materials raises new challenging problems for information retrieval systems.

We are pleased to introduce this special issue of Informatica journal that includes four revised and extended papers, presented at the 1st International Workshop on Advances in Semantic Information Retrieval (Szczecin, Poland, September 18-21, 2011).

The first paper entitled A Query Expansion Technique using the EWC Semantic Relatedness Measure by V. Klyuev and Y. Haralambous proposes the usage of a new relatedness measure in the query expansion task. One of the goals of query expansion is to reformulate the user query in order to reduce the number of non-relevant documents retrieved by search systems. EWC extends the gaining popularity explicit semantic analysis (ESA). EWC stands for ESA, plus Wordnet, plus Collocations. Whereas ESA takes into account only encyclopedic knowledge from Wikipedia, EWC considers encyclopedic, ontological, and collocational knowledge about terms. The authors use this advantage of EWC over ESA to find more precise terms to expand the user queries. They tested proposed techniques on the NTCIR data collection which is similar to that of TREC. The authors discuss the details of proposed technique, investigate the nature of improvements in the retrieval performance, and outline the directions for the future experiments.

The next paper by A. Patyk-Łońska, M. Czachor, and D. Aerts entitled Distributed Representations Based on Geometric Algebra: the Continuous Model introduces a new model of distributed representations of complex structures (sentences in natural languages) based on geometric algebra. They compare it with two other models: Holographic Reduced Representation and Binary Spatter Codes (BSC). Results of their evaluations show that the best models for storing and recognizing multiple similar statements in natural languages are the new model and BSC with recognition percentage above 90%.

The third paper entitled Experiments on Preserving Pieces of Information in a Given Order in Holographic Reduced Representations and the Continuous Geometric Algebra Model by A. Patyk-Łońska addresses the need to develop a new scheme for encoding and decoding complex structures that is based entirely on geometric symbols. The author studies the properties of Geometric Analogues of Holographic Reduced Representations as the ability to store pieces of information in a given order by means of trajectory association. The work describes the results of three types of the experiments: finding correct item or correct place of an item in a sequence and finding the alignment of items in a sequence without the precise knowledge of trajectory vectors.

The fourth paper entitled Grammar Checking with Dependency Parsing: a Possible Extension for LanguageTool by M. Mozgovoy examines the use of dependency-based syntactic parsing in the problem of grammar checking. The author proposes a possible extension for a well-known open source grammar checker LanguageTool. This extension will allow the users to compose new grammar rules, based on word-word dependency links. The paper demonstrates real situations, where such capabilities are helpful. The author also proposes rule syntax, similar to existing conventions of LanguageTool, and discusses implementation and testing issues.

The first year of the ASIR workshop was successful. We received 12 submissions, and 7 of them were accepted and presented on-site. We believe that the workshop will facilitate discussion of new research results in this area, and will serve as a meeting place for researchers from all over the world. Our aim is to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics and information retrieval. The workshop is now established as an event within Federated conference on computer science and information systems (FedCSIS), annually organized by the System Research Institute of the Polish Academy of Sciences and the Polish Information Processing Society, and sponsored by the IEEE.

In its turn, ASIR is supported by the University of Aizu (Japan), known as Japan's first university, solely dedicated to computer science engineering. The University of Aizu is a major center of international education and the home of several conferences, sponsored by the ACM and the IEEE.

We would wish to acknowledge selfless efforts of our committee members and FedCSIS conference organizers, who ensured high quality of publications and flawless

arrangement of the forum. We would like to specially mention professors Marcin Paprzycki, Maria Ganzha, and Halina Kwasnicka, responsible for FedCSIS. We had a great support from our international team of reviewers, consisting of: Wladyslaw Homenda, Maciej Piasecki (Warsaw University of Technology, Poland); Antoni Ligeza (AGH University of Science and Technology, Poland); Nikolay Mirenkov, Alexander Vazhenin, Ryuichi Oka (University of Aizu, Japan); Marek Reformat (University of Alberta, Canada); Qun Jin (Waseda University, Japan); Eloisa Vargiu (University of Cagliari, Italy); Tuomo Kakkonen (University of Eastern Finland); Roman Shtykh (Rakuten Inc., Japan); Slawomir Zadrozny (Systems Research Institute of the Polish Academy of Sciences, Poland); Vladimir Oleshchuk (University of Agder, Norway); Kamen Kanev (Shizuoka University, Japan); Cristian Lai (CRS4, Italy); Troels Andreasen (Roskilde University, Denmark); Anna Fensel (Vienna FTW, Austria); Evgeny Pyshkin (Saint-Petersburg State Polytechnical University, Russia); Shih-Hung Wu (Chaoyang University of Technology, Taiwan); Vladimir Dobrynin (Saint-Petersburg State University, Russia); Simone Ludwig (North Dakota State University, USA).

We also thank Professor Matjaz Gams (managing editor of Informatica), who supported the publication of this special issue.

This year, we are organizing ASIR'2012 within FedCSIS in Wroclaw, Poland. We will continue to maintain high standards of quality and organization, set by the first workshop. We welcome all the researchers, interested in semantics and information retrieval, to join our event.

ASIR Chairs.
Vitaly Klyuev, Associate Professor
Maxim Mozgovoy, Assistant Professor
The University of Aizu
Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima,
965-8580 JAPAN
{vkluev, mozgovoy}@u-aizu.ac.jp

A Query Expansion Technique Using the EWC Semantic Relatedness Measure

Vitaly Klyuev
University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, 965-8580 Japan
E-mail: vkluev@u-aizu.ac.jp

Yannis Haralambous
Institut Télécom – Télécom Bretagne, Dép. Informatique,
UMR CNRS 3192 Lab-STICC Technopôle Brest Iroise, CS 83818, 29238 Brest Cedex 3, France
E-mail: yannis.haralambous@telecom-bretagne.eu

Keywords: relatedness measure, Wikipedia, WordNet, search engine, query expansion

Received: October 28, 2011

This paper analyses the efficiency of the EWC semantic relatedness measure in an ad-hoc retrieval task. This measure combines the Wikipedia-based Explicit Semantic Analysis (ESA) measure, the WordNet path measure and the mixed collocation index. EWC considers encyclopaedic, ontological, and collocational knowledge about terms. This advantage of EWC is a key factor to find precise terms for automatic query expansion. In the experiments, the open source search engine Terrier is utilised as a tool to index and retrieve data. The proposed technique is tested on the NTCIR data collection. The experiments demonstrated superiority of EWC over ESA.

Povzetek: Članek obravnava razširjanje poizvedb z uporabo semantične povezave med besedami.

1 Introduction

A bag-of-words representation of documents by information retrieval systems results in queries expressed utilising the language of keywords. Users face a vocabulary problem: A keyword language is not adequate to describe the information needs. Statistical analysis of styles of user behaviour showed that the queries submitted to search engines are short (2 to 3 terms, on average) and ambiguous, and users rarely look beyond the first 10 to 20 links retrieved [20].

To improve user queries, search engines provide many tools.

Manuals and instructions are among them. They help a little, although ordinary users do not like reading.

A query suggestion feature is common for general purpose search engines. Its disadvantage is in the way to generate recommended terms. They are based on the first term typed by the user, which is always the first one in all expanded queries [22].

The relevance feedback feature is another instrument to help users. There are at least two steps in the interaction with a search engine: The first one is the submission of the original query, and the second one is the user reaction on the results retrieved in order to provide the system with the user opinion (marking some documents as relevant). This feature is not popular among users because the mechanisms of changing queries are not clear and users cannot control the process [9].

In addition, query suggestion and relevance feedback are used to modify the queries in order to make them more accurate to express the user information needs.

Query expansion is a well-known and popular technique to reformulate the user query in order to reduce the number of non-relevant pages retrieved by information retrieval systems. Another goal of query expansion is to provide the user with additional relevant documents. Automatic query expansion is an important area of information retrieval: Many scientists are involved in designing new methods, techniques, and approaches.

This paper presents authors' technique to automatically expand the user queries. This technique is based on the EWC semantic relatedness measure [21]. This measure takes into account encyclopaedic, ontological, and collocational knowledge about terms. The environment for the experiments includes Terrier as a search engine and NTCIR-1 CLIR data collection for the Japanese-English cross-lingual retrieval task.

The rest of the paper is organised as follows. The next section reviews the approaches to automatic query expansion. Section 3 describes the nature of the measure used. Section 4 provides the necessary details related to this technique to expand the queries. The tools and data utilised in the experiments are presented in Section 5.

[*] This paper is based on V. Klyuev and Y. Haralambous, *Query Expansion: Term Selection using the EWC Semantic Relatedness Measure* published in the proceedings of the 1st International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2011 conference).

The results of the experiments are discussed in Section 6. Comments on the future directions of the investigations are presented there. Concluding remarks can be found in Section 7.

2 Related Work

A comprehensive review of the classical approaches to expand queries can be found in [9]. They propose different ways to obtain semantically (topically) related terms, techniques to evaluate importance of the terms found, mechanisms to define the number of terms to add (expand) the user query, and strategies to evaluate the quality of obtained results.

Generally, the semantics of the terms are clear when they are in the sentences because the meanings of the words are fixed and only one is usually selected from the set of all possible variants. However, in the queries, the terms are separate instances, and their semantics are unclear. This is called the polysemy problem. On the other hand, the same things can be described by using different terms. This is the nature of the synonymy problem. The query should be rich enough to include the possible candidates for expressing user information needs.

The general goal of query expansion is to find a solution for these two aforementioned problems. The classical solution for the synonymy problem is to apply thesauri as instruments to obtain the candidates for expansion. WordNet is widely used for this purpose [19]. Modern techniques suggest Wikipedia as a valuable source to find synonyms [12].

Many techniques are used to solve the polysemy problem. Approaches described in [13] and [16] are based on the analysis of the query log files of search engines and clicked URLs. Authors of this study [18] utilised WordNet for a deep analysis of the queries submitted to the information retrieval system in order to find the concepts and then obtain the candidate terms for expansion. The involvement of users is the feature of the approach discussed in [17]. They should select the correct ontology for each query submitted to expand the query.

The authors of this study [14] also pointed out that the information exploited by different approaches differs, and combining the different query expansion approaches is more efficient than the use of any of them separately. They investigated techniques to rank the terms extracted from the retrieved documents. One is based on the measures of occurrence of the candidate and query terms in the retrieved documents. The other one utilises the differences between the probability distribution of terms in the collection and in the top ranked documents retrieved by the system. A similar idea is discussed in [15].

The authors of [10] combined the concept-based retrieval, based on explicit semantic analysis (ESA), with keyword-based retrieval. At the first step, they use keyword-based retrieval to obtain the candidates for query expansion. Then, they tune queries applying ESA. After that, they perform the final retrieval in the space of concepts.

It is difficult to compare the aforementioned approaches, because different data sets were used to evaluate them. In many cases, it is not clear wherever the test queries cover a wide range of data set topics. The performance evaluation is done automatically for some approaches, whereas for others, the authors involve the users to judge the quality of retrieval.

3 Measure Description

In study [21], the new measure of words relatedness is introduced. It combines the ESA measure μ_{ESA} [10], the ontological WordNet path measure μ_{WNP}, and the collocation index C_ξ. This measure is called EWC (ESA plus WordNet, plus collocations) and is defined as follows:

$$\mu_{EWC}(w_1, w_2) = \mu_{ESA}(w_1, w_2) * \alpha$$
$$\alpha = (1 + \lambda_\sigma(\mu_{WNP}(w_1, w_2))) * \gamma$$
$$\gamma = (1 + \lambda'_\sigma(C_\xi(w_1, w_2)))$$

where λ_σ weights the WordNet path measure (WNP) with respect to ESA, and λ'_σ weights the mixed collocation index with respect to ESA. This index is defined as follows:

$$C_\xi = \frac{2 * f(w_1, w_2)}{f(w_1) + f(w_2)} + \xi \frac{2 * f(w_2, w_1)}{f(w_1) + f(w_2)}$$

where $f(w_1, w_2)$, $f(w_2, w_1)$ are the frequency of the collocations of $w_1 w_2$ and $w_2 w_1$ in the corpus, and $f(w_i)$ is the frequency of word w_i. The values for constants λ_σ, λ'_σ, and ξ are set to 5.16, 48.7, and 0.55, respectively on the basis of empirical tests.

Study [21] demonstrated the superiority of this measure over ESA on the WS-353 test set.

The current implementation of EWC does not take into account Wikipedia articles with titles consisting of multiple terms (they are dimensions in the Wikipedia space). As a result, the proposed technique cannot distinct multiple term items from collocations and give them the highest score.

4 Technique to Expand Queries

Assume that Z is a pool of term-candidates for query expansion. The formulas below present the method to select terms to expand queries. N is a number of original query terms, and j is an index of them. Values for the

WordNet component and collocation component should be above zero in order to choose related terms. Thresholds t_2 for EWC values and t_1 are parameters adjusted in the experiments. For every word $w_i \in Z$, the weight is calculated. Word w_i is selected for expansion if its weight is equal to 1.

$$weight(w_i) = \begin{cases} 1, if \sum_{j=0}^{N} \frac{score(w_1, w_2)}{N} > t_1 \\ 0, otherwise \end{cases}$$

$$score(w_1, w_2) = \begin{cases} 1, if \mu_{WNP} > 0; C_{\xi} > 0; \mu_{EWC} > t_2 \\ 0, otherwise \end{cases}$$

This approach can be interpreted as follows: A term is selected from the list of term-candidates, if the similarity score between this term and the majority of original query terms is higher than a given threshold t_1. The term-candidate should have non-zero values for μ_{WNP} and C_{ς} components.

5 Tools and Data Sets Used

The open source search engine Terrier [1] was used as a tool to index and retrieve data. It provides the different retrieval approaches. TF-IDF and Okapi's B25 schemas [6, 9] are among them.

As a data set for experiments, the NTCIR CLIR data collection [2] was used. It consists of 187,000 articles in English. These articles are summaries of papers presented at scientific conferences hosted by Japanese academic societies. The collection covers a variety of topics such as chemistry, electrical engineering, computer science, linguistics, and library science. The size of the collection is approximately 275.5 MB. A total of 83 topics are in Japanese. A structure of the dataset and topics is similar to that of TREC [3].

A straightforward approach was applied to translate queries into English: Google's translation service [4] generated queries in English. This method was selected because on-line dictionaries do not work well with terms in katakana and specific terminology [7]. Katakana is one of four sets of characters used in Japanese writing. It is primarily applied for the transcription of foreign language words into Japanese.

A Porter Stemmer algorithm was applied to the documents and queries, and a standard stop word list provided by Terrier was also utilised. Only the title fields were considered as a source of the queries. They are relatively short: each query consists of a few keywords. The authors of the study reported in [5] experimented with Terrier applying the same conditions to the TREC data.

To measure the term similarities, an experimental tool described in [21] was utilised.

6 Results of Experiments

The authors implemented the proposed technique to expand queries as follows.

To obtain the candidates for query expansion, a query expansion functionality offered by Terrier was adopted. It extracts the most informative terms (in this case 10) of the top-ranked documents (in this case 3) by using a particular DFR (divergence from randomness) term weighting model [8].

Table 1 provides the list of original queries (topics 1, 12, and 24), candidate terms for expansion (arranged by the decreasing score calculated by Terrier), and the final sets of terms used to expand queries (they are in bold).

Table 1: Original and Expanded Queries for Topics: 1, 12 and 24.

Topic	Original query	Terms for expansion
1	Robot	**Robot** person **human** multi comput sice design will confer paper
12	Mining methods	Mine method rule data databas associ discoveri **larg** tadashi solv amount
24	Machine translation system	Machin translat **system** exampl **base** **masahiro** **method** nation convert **problem**

Table 2: Thresholds Tuning: Topics 1 to 30.

t1	t2	EWC: Average Precision R-Precision			ESA/ BM25
		InL2	TF-IDF	BM25	
0.5	0.1	0.2940 0.3216	0.3031 0.3314	0.3072 0.3324	**0.3101** **0.3347**
0.65	0.09	0.2940 0.3300	0.2936 **0.3332**	0.2955 0.3278	**0.2961** 0.3276
0.67	0.07	**0.2973** **0.2977**	0.2954 0.2963	0.2960 0.2916	0.2959 0.2916
0.67	0.08	0.3101 0.3277	0.3151 0.3268	0.3140 0.3265	**0.3172** **0.3315**
0.67	0.09	0.3073 0.3256	0.3105 **0.3373**	**0.3106** 0.3352	0.3080 0.3373
0.67	0.1	0.3030 0.3295	**0.3103** **0.3350**	0.3099 0.3349	0.3099 0.3318
0.67	0.11	0.3049 0.3239	**0.3121** **0.3309**	0.3110 0.3292	0.3102 0.3302
0.67	0.12	0.3049 0.3239	**0.3121** **0.3309**	0.3110 0.3292	0.3092 0.3284
0.67	0.13	0.3049 0.3125	**0.3114** 0.3245	0.3099 0.3248	0.3111 **0.3282**
0.67	0.15	0.3033 0.3143	**0.3115** 0.3267	0.3111 **0.3282**	0.3110 0.3282
0.69	0.09	0.3073 0.3256	0.3105 **0.3373**	**0.3106** 0.3352	0.3080 0.3373
0.75	0.1	0.3030 0.3295	0.3103 0.3330	**0.3099** **0.3349**	0.3099 0.3318
System	system	0.2980 0.2995	0.3034 0.3166	0.3017 0.3163	

Candidate terms for expansion are presented in stemmed form after applying the aforementioned Porter algorithm. One to five terms were selected by this method. As one can see from this table, this technique does not usually select the top-ranked terms as candidates for expansion from the Terrier engine point of view.

As mentioned in Section 5, a total of 83 topics are available to retrieve documents from the collection. The original goal of topics 0001 to 0030 is to tune the parameters of the retrieval system. Relevance judgments for some of them are known in advance. Topics 0031 to 0083 were used in official runs at the NTCIR 1 Workshop. Organisers found that the number of relevant documents for 13 topics of the 53 contained less than five relevant documents per topic in cross-lingual retrieval. Hence, they discarded these topics from evaluation [25]. The full set was used in these experiments because the goal is to compare the performance of different methods implemented in the same environment.

Table 3: Evaluation Results for Topics 31 to 83.

Recall level	Precision		
	EWC	ESA	System
at 0.00	0.9594	0.9702	0.9676
at 0.10	0.7911	0.7852	0.7972
at 0.20	0.7007	0.6904	0.5859
at 0.30	0.5288	0.5245	0.5001
at 0.40	0.3642	0.3586	0.3599
at 0.50	0.2878	0.2856	0.3011
at 0.60	0.176	0.1767	0.181
at 0.70	0.1597	0.1571	0.1394
at 0.80	0.1061	0.101	0.0862
at 0.90	0.0557	0.0533	0.0354
at 1.00	0.0404	0.0405	0.0242

In the evaluations, the partially relevant documents were considered as irrelevant. To archive this, the corresponding file with the answers provided by NTCIR Workshop organisers was applied when evaluating the retrieval results.

Table 2 summarises the results of retrieval to tune thresholds t_1 and t_2. The test queries were generated from topics 0001 to 0030. It is important to note that when the queries are expanded with all the terms proposed by Terrier, the retrieval results drop to zero. The retrieval utilising the TF-IDF schema produced better results for the original queries (without expansion) compared to the BM25 and InL2 models [1]. The line *system* shows this result. The first number in the cells is the value of average precision, and the second one is the value of R-precision. The performance of retrieval with expanded queries utilising the ESA and EWC approaches for the threshold values (t_1 equals 0.67 and t_2 ranges from 0.08 to 0.15) is better compared to the variant without expansion. For the EWC measure, the maximum of the retrieval performance is reached when the values of thresholds t_1 and t_2 are set to 0.67 and 0.12. For ESA, the optimal threshold values are 0.67 and 0.08. The performance of ESA is higher than EWC.

Table 3 shows precision/recall evaluation results across 53 queries.

Table 4 summarises the results of retrieval for topics 31 to 83. The threshold values were set to the optimal parameters (see Table 2). Six runs were executed. The EWC measure demonstrated better performance (see values in bold) over ESA in both cases (average

Table 4: Retrieval Results: Topics 31 to 83.

t1	t2	EWC: Average Precision R-Precision		ESA: Average Precision R-Precision	
		BM25	TF-IDF	BM25	TF-IDF
0.67	0.08	**0.2363** **0.2416**		0.2349 0.2390	
0.67	0.12		0.2200 0.2317		0.2161 0.2284
System	system	0.2225 0.2350	**0.2218** **0.2359**		

Table 5: Differences in the Expanded Queries for Topics 31 to 83.

Topic	Original Queries	Expansion Terms	
		EWC	ESA
32	Network Collaboration	design technolog web world tool focu wide	design technolog world tool focu wide
37	buffer control	buffer control memori input	memori input
38	TCP / IP Throughput Performance Communications	network	
46	Reset period algorithm	system	
56	Information Lifecycle artifacts knowledge sharing	knowledg design product life	knowledg
62	Lifelong learning and volunteer	learn educ	educ
69	Computer-aided teaching	aid teach educ instruct person system	aid educ person system
75	Simulation exercise	system work model	system work
81	Sex differences in brain	differ mean	differ
82	Antimalarial drugs	antimalari drug	antimalari

precision and R-Precision). The line *system* shows the retrieval results without expansion for TF-IDF and BM25 schemes

Table 5 presents the differences in the expanded queries for topics 31 to 83 for EWC and ESA metrics. Expansion terms are presented in stemmed form in this table. One can see that some expansion terms are taken from original queries. This is the case for topics 37, 56, 62, 69, 81, and 82. Such a selection increases the importance of the respective search terms at the searching process.

Among 53 topics, there are only the differences in topics 32, 37, 38, 46, 56, 62, 69, 75, 81, and 82. For the remaining topics, both approaches (EWC and ESA) generated the same queries. In other words, only these 10 queries contributed in improvements of the performance of the search.

Original queries for topics 38 and 46 were not expanded when the ESA approach was applied.

As we mentioned in Section 3, the implementation of EWC discards Wikipedia articles with titles consisting of multiple terms (they are dimensions in the Wikipedia space). Multiple word terms cannot be recognized and scored accurately. These terms are widely used in scientific terminology. Recognition of such terms seems to be the most promising direction to enhance EWC.

To summarize, one can conclude that the EWC measure provides little benefit over ESA, as the results of the retrieval are better. Ontological knowledge combined with collocational knowledge helps in the selection of expansion terms.

7 Conclusion

This study tested the semantic relatedness measure when selecting the terms to expand queries. Key components of this measure are the ESA measure, the WordNet path measure, and the mixed collocation index. Results produced by the Terrier search engine were a base line in the experiments.

Term candidates for the expansion were also generated by Terrier. The proposed techniques were applied to the ad-hoc retrieval task. As a data set, the NTCIR-1 CLIR Test collection was used. The initial English queries were obtained automatically applying Google translate. The queries were expanded by applying the Wikipedia-based Explicit Semantic Analysis measure, and the DFR mechanism, and the semantic relatedness measure. The retrieval results showed superiority of the last one over ESA and DFR.

A promising new direction to enhance the EWC measure is to take into account Wikipedia articles with titles consisting of multiple terms in order to get knowledge about scientific terminology and general purpose terminology. The expected outcome of this is the more precise term selection to expand the user queries submitted to search engines.

References

[1] Terrier (2011). [Online document], http://terrier.net

[2] NTCIR-1 CLIR data collection (1999). [Online document], http://research.nii.ac.jp/ntcir/data/data-en.html

[3] TREC (2011). [Online document], http://trec.nist.gov/

[4] Google Translate (2011). [Online document], http://translate.google.com/

[5] Ben He and Iadh Ounis (2009). Studying Query Expansion Effectiveness. In Proc. *The 31st European Conference on Information Retrieval (ECIR09)*. Toulouse, France.

[6] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne (1995). Okapi at TREC-4. In Proc. *TREC 4*.

[7] Aitao Chen, Fredric C. Gey, Kazuaki Kishida, Hailing Jiang and Qun Liang (1999). Comparing Multiple Methods for Japanese and Japanese-English Text Retrieval, NTCIR Workshop 1. In Proc. *The First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition,* Tokyo.

[8] G. Amati and C.J. Van Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

[9] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze (2008). *Introduction to Information Retrieval*, Cambridge University Press.

[10] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich (2011). Concept-Based Information Retrieval using Explicit Semantic Analysis, *ACM Transactions on Information Systems*, 29(2).

[11] Philipp Sorg and Philipp Cimiano (2009). An Experimental Comparison of Explicit Semantic Analysis Implementations for Cross-Language Retrieval. In Proc. *The International Conference on Applications of Natural Language to Information Systems (NLDB)*, Saarbrücken.

[12] Yinghao Li, Wing Pong Robert Luk, Kei Shiu Edward Ho, and Fu Lai Korris Chung (2007). Improving weak ad-hoc queries using Wikipedia as external corpus. In Proc. *The 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, pp 797 - 798.

[13] Hamada M.Zahera, Gamal F. El Hady, and Waiel.F Abd El-Wahed (2010). Query Recommendation for Improving Search Engine Results. In Proc. *The World Congress on Engineering and Computer Science 2010* Vol I, WCECS 2010, San Francisco, USA.

[14] Jose R. Perez-Aguera1 and Lourdes Araujo (2008). Comparing and Combining Methods for Automatic Query Expansion. *Advances in Natural Language Processing and Applications Research in Computing Science 33*, pp. 177-188.

[15] Ming-hung Hsu, Ming-feng Tsai, and Hsin-hsi Chen (2008). Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach. In Proc. *Asia Information Retrieval Symposium*, pp. 213-224.

[16] Hang Cui1, Ji-Rong Wen, Jian-Yun Nie3, and Wei-Ying Ma (2002). Probabilistic query expansion using query logs. In Proc. *The 11th international conference on World Wide Web*, ACM New York, NY, USA.

[17] J. Malecka and V. Rozinajova (2006). An Approach to Semantic Query Expansion. In Proc. *Tools for Acquisition, Organization and Presenting of Information and Knowledge, Research Project Workshop*, Bystra dolina, Tatry, pp. 148-153.

[18] Jiuling Zhang, Beixing Deng, and Xing Li (2009). Concept Based Query Expansion Using WordNet. In Proc. *The 2009 International e-Conference on Advanced Science and Technology*, pp. 52-55.

[19] M. Ellen Voorhees (1994). Query expansion using lexical-semantic relations. In Proc. *The 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*.

[20] D. Gayo-Avello Brenes (2009). Stratifies analysis of AOL query log. *Information Sciences*, 179, pp. 1844-1858.

[21] Yannis Haralambous and Vitaly Klyuev (2011). A Semantic Relatedness Measure Based on Combined Encyclopedic, Ontological and Collocational Knowledge. In the Proc. *IJCNLP*.

[22] Vitaly Klyuev, Ai Yokoyama (2010). Web Query Expansion: A Strategy Utilizing Japanese WordNet. *Journal of Convergence*, V. 1, Number 1.

[23] Space ALC (2011). [Online document], http://www.alc.co.jp/

[24] Mecab (2011). [Online document], http://mecab.sourceforge.net/

[25] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato and Soichiro Hidaka (1999). Overview of IR tasks. In Proc. *The First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Tokyo.

Distributed Representations Based on Geometric Algebra: The Continuous Model

Agnieszka Patyk-Łońska, Marek Czachor
Gdańsk University of Technology
ul. Narutowicza 11/12, Gdańsk 80-233, Poland
E-mail: {patyk, mczachor}@pg.gda.pl,
http://www.pg.gda.pl/ patyk
http://www.mif.pg.gda.pl/kft/czachor.html

Diederik Aerts
Centrum Leo Apostel (CLEA), Vrije Universiteit Brussel
Krijgskundestraat 33, 1160 Brussels, Belgium
E-mail: diraerts@vub.ac.be
http://www.vub.ac.be/CLEA/aerts/

Keywords: distributed representation of data, geometric algebra, HRR, BSC, scaling

Received: October 23, 2011

Authors revise the concept of a distributed representation of data as well as two previously developed models: Holographic Reduced Representation (HRR) and Binary Spatter Codes (BSC). A Geometric Analogue (GA_c — "c" stands for continuous as opposed to its discrete version) of HRR is introduced – it employs role-filler binding based on geometric products. Atomic objects are real-valued vectors in n-dimensional Euclidean space while complex data structures belong to a hierarchy of multivectors. The paper reports on a test aimed at comparison of GA_c with HRR and BSC. The test is analogous to the one proposed by Tony Plate in the mid 90s. We repeat Plate's test on GA_c and compare the results with the original HRR and BSC — we concentrate on comparison of recognition percentage for the three models for comparable data size, rather than on the time taken to achieve high percentage. Results show that the best models for storing and recognizing multiple similar structures are GA_c and BSC with recognition percentage highly above 90. The paper ends with remarks on perspective applications of geometric algebra to quantum algorithms.

Povzetek: Članek se ukvarja s porazdeljeno predstavitvijo podatkov, ki uporablja geometrijsko algrebro.

1 Introduction

Distributed representations of data are very different from traditional structures (e.g. trees, lists) and complex structures bare little resemblance to their components, therefore great care must be taken when composing or decomposing a complex structure. The most widely used definition of a distributed representation is due to Hinton *et al.* [13]. In a *distributed representation of data* each concept is represented over a number of units and each unit participates in the representation of some number of concepts. The size of a distributed representation is usually fixed and the units have either binary or continuous-space values. In most distributed representations only the overall pattern of activated units has a meaning.

Let us consider an example of storing the following information: "Fido bit Pat". The action in this statement is $bite$ and the features (i.e. $roles$) of this action are an agent and an object, denoted $bite_{agt}$ and $bite_{obj}$, while their $fillers$ are $Fido$ and Pat respectively. If we consider storing the way that the action is performed, we can add a third feature ($role$), e.g. $bite_{way}$. If we store $Fido$, Pat, $bite_{agt}$ and $bite_{obj}$ as vectors, we are able to encode "Fido bit Pat" as

$$bite_{agt} * Fido + bite_{obj} * Pat.$$

The operation of $binding$, denoted by "$*$", takes two vectors and produces another vector, often called a $chunk$ of a sentence. It would be ideal for the resulting vector not to be similar to the original vectors but to have the same dimensions as the original vectors. $Superposition$, denoted by "+", is an operation that takes any number of vectors and creates another one that is similar to the original vectors. Usually, the superimposed vectors are already the result of the binding operation.

Irrespectively of the mathematical model, the above operations are defined in a way that allows to build complex

This paper is based on A. Patyk-Łońska, M. Czachor and D. Aerts *Some tests on geometric analogues of Holographic Reduced Representations and Binary Spatter Codes* published in the proceedings of the 1st International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2011 conference).

statements, such as “John saw Fido bit Pat":

$$John * see_{agt} + (bite_{agt} * Fido + bite_{obj} * Pat) * see_{obj}.$$

In order to decode information, we have to use the operation of $unbinding$ — it is the inverse (an exact inverse or a pseudo-inverse) of binding enabling us to extract an information from a complex statement, provided that we have one of the bound vectors or a very similar vector as a cue. Marking the unbinding operation by “♯" we obtain the following answer to “Who bit Pat?":

$$(bite_{agt} * Fido + bite_{obj} * Pat) \sharp\, bite_{agt} = Fido'.$$

We cannot definitely say that the resulting vector $Fido'$ will be an exact copy of $Fido$, as even an optimal scheme will generate a considerable amount of noise. Since we cannot expect that a noisy decoded information will be identical to what was encoded, we have to rely heavily on various similarity measures — they vary mostly by time taken by computation and the accuracy.

Clean-up memory is an auto-associative collection of all atomic objects and complex statements produced by the system. Given a noisy extracted vector such structure must be able to recall the most similar item stored or indicate, that no matching object had been found.

Independently of the scheme considered, any representation should possess the following qualities

- composition and decomposition — rules of composition and decomposition must be applicable to all elements of the domain, irrespectively of the degree of complication of a given element. Further, decomposition should support structure-sensitive processing.

- fixed size — structures of different degree of complication should take up the same amount of space in order to facilitate generalization. In the GA_c model this feature has been given up. Still, structures of different complexity will be of the same type.

- similarity — the representation scheme should provide a quick way to compute similarity between analogous structures (e.g. $Fido\ bit\ Pat\ Smith$ and $Fido\ bit\ John$).

- noise reduction — decomposed statements should resemble their original counterpart.

- productivity — the model should be able to construct complex nested structures using a set of only few rules.

As far as previously developed models are concerned, Holographic Reduced Representations (HRR), Binary Spatter Codes (BSC), and Associative-Projective Neural Networks (APNN) are distributed representations of cognitive structures where binding of role–filler codevectors maintains predetermined data size. In HRR [23] binding is performed by means of circular convolution

$$(x \circledast y)_j = \sum_{k=0}^{n-1} x_k y_{j-k \bmod n}.$$

of real n-tuples or, in ‘frequency domain’, by componentwise multiplication of (complex) n-tuples,

$$(x_1, \ldots, x_n) \circledast (y_1, \ldots, y_n) = (x_1 y_1, \ldots, x_n y_n).$$

Bound n-tuples are superposed by addition, and unbinding is performed by an approximate inverse. A dual formalism, where real data are bound by componentwise multiplication, was discussed by Gayler [9]. In BSC [14, 15] one works with binary n-tuples, bound by componentwise addition mod 2,

$$\begin{aligned} (x_1, \ldots, x_n) \oplus (y_1, \ldots, y_n) &= \\ &= (x_1 \oplus y_1, \ldots, x_n \oplus y_n), \\ x_j \oplus y_j &= x_j + y_j \bmod 2, \end{aligned} \tag{1}$$

and superposed by pointwise majority-rule addition; unbinding is performed by the same operation as binding. APNN, introduced and further developed by Kussul [16] and his collaborators [17], employ binding and superposition realized by a context-dependent thinning and bitwise disjunction, respectively. As opposed to HRR and BSC, APNN do not require an unbinding procedure to retrieve component codevectors from their bindings. A detailed comparison of HRR, BSC and APNN can be found in [24].

2 Geometric Algebra

One often reads that the above models represent data by *vectors*, which is not exactly true. Given two vectors one does not know how to perform, say, their convolution or componentwise multiplication since the result depends on basis that defines the components. Basis must be fixed in advance since otherwise all the above operations become ambiguous. It follows that neither of the above reduced representations can be given a true and meaningful geometric interpretation. Geometric analogues of HRR [5] can be constructed if one defines binding by the geometric product, a notion introduced in 19th century works of Grassmann [11] and Clifford [8].

The fact that a geometric analogue of HRR is intrinsically geometric may be important for various conceptual reasons — for example, the rules of geometric algebra may be regarded as a mathematical formalization of the process of *understanding* geometry. The use of geometric algebra distributed representations has been inspired by a well-known fact, that most people think in pictures, i.e. two- and three-dimensional shapes, not by using sequences of ones and zeroes. Mere strings of bits are not meaningful to (most) humans, no matter how technically advanced they are.

In order to grasp the main ideas behind a geometric analogue of HRR let us consider an orthonormal basis $b_1, \ldots, b_n$ in some n-dimensional Euclidean space. Now consider two vectors $x = \sum_{k=1}^{n} x_k b_k$ and $y = \sum_{k=1}^{n} y_k b_k$. The *scalar*

$$x \cdot y = y \cdot x$$

is known as the *inner product*. The *bivector*

$$x \wedge y = -y \wedge x$$

is the *outer product* and may be regarded as an oriented plane segment (alternative interpretations are also possible, cf. [7]). $\mathbf{1}$ is the identity of the algebra. The geometric product of x and y then reads

$$xy = \underbrace{\sum_{k=1}^{n} x_k y_k \, \mathbf{1}}_{x \cdot y} + \underbrace{\sum_{k<l} (x_k y_l - y_k x_l) b_k b_l}_{x \wedge y} .$$

Grassmann and Clifford introduced geometric product by means of the basis-independent formula involving the *multivector*

$$xy = x \cdot y + x \wedge y \tag{2}$$

which implies the so-called Clifford algebra

$$b_k b_l + b_l b_k = 2\delta_{kl}\mathbf{1}.$$

when restricted to an orthonormal basis. Inner and outer product can be defined directly from xy:

$$\begin{aligned} x \cdot y &= \frac{1}{2}(xy + yx), \\ x \wedge y &= \frac{1}{2}(xy - yx). \end{aligned}$$

The most ingenious element of (2) is that it adds two apparently different objects, a scalar and a plane element, an operation analogous to addition of real and imaginary parts of a complex number. Geometric product for vectors x, y, z can be axiomatically defined by the following rules:

$$\begin{aligned} (xy)z &= x(yz), \\ x(y+z) &= xy + xz, \\ (x+y)z &= xz + yz, \\ xx = x^2 &= |x|^2, \end{aligned}$$

where $|x|$ is a positive scalar called the magnitude of x. The rules imply that $x \cdot y$ must be a scalar since

$$xy + yx = |x+y|^2 - |x|^2 - |y|^2.$$

Geometric algebra allows us to speak of inverses of vectors: $x^{-1} = x/|x|^2$. x is invertible (i.e. possesses an inverse) if its magnitude is nonzero. Geometric product of an arbitrary number of invertible vectors is also invertible. The possibility of inverting all nonzero-magnitude vectors is perhaps the most important difference between geometric and convolution algebras.

Geometric products of *different* basis vectors

$$b_{k_1 \ldots k_j} = b_{k_1} \ldots b_{k_j},$$

$k_1 < \cdots < k_j$, are called basis blades (or just blades). In n-dimensional Euclidean space there are 2^n different blades. This can be seen as follows. Let $\{x_1, \ldots, x_n\}$ be a sequence of bits. Blades in an n-dimensional space can be written as

$$c_{x_1 \ldots x_n} = b_1^{x_1} \ldots b_n^{x_n}$$

where $b_k^0 = \mathbf{1}$, which shows that blades are in a one-to-one relation with n-bit numbers. A general multivector is a linear combination of blades,

$$\psi = \sum_{x_1 \ldots x_n = 0}^{1} \psi_{x_1 \ldots x_n} c_{x_1 \ldots x_n}, \tag{3}$$

with real or complex coefficients $\psi_{x_1 \ldots x_n}$. Clifford algebra implies that

$$\begin{aligned} &c_{x_1 \ldots x_n} c_{y_1 \ldots y_n} = \\ &\quad = (-1)^{\sum_{k<l} y_k x_l} c_{(x_1 \ldots x_n) \oplus (y_1 \ldots y_n)}, \end{aligned} \tag{4}$$

where $\oplus$ is given by (1). Multiplication of two basis blades is thus, up to a sign, in a one-to-one relation with exclusive alternative of two binary n-tuples. Accordingly, (4) is a projective representation of the group of binary n-tuples with addition modulo 2.

The GA_c model is based on binding defined by geometric product (4) of blades while superposition is just addition of blades (3). The discrete GA_d [19] is a version of the GA_c model obtained if $\psi_{x_1 \ldots x_n}$ in (3) equal ± 1. The first recognition tests of GA_d, as compared to HRR and BSC, were described in [19]. In the present paper we go further and compare HRR and BSC with GA_c, a version employing "projected products" [5] and arbitrary real $\psi_{x_1 \ldots x_n}$. We also repeat Plate's scaling test ([22], [23] – Appendix I) and compare test results for GA_c, HRR and BSC models.

Throughout this paper we shall use the following notation: "$*$" denotes binding roles and fillers by means of the geometric product and "$+$" denotes the superposition of sentence chunks. Additionally, "$\circledast$" will denote binding performed by circular convolution used in the HRR model and a^* denotes the involution of a HRR vector a. A "$+$"' in the superscript of x^+ denotes the operation of reversing a blade or a multivector x: $(b_{k_1 \ldots k_j})^+ = b_{k_j} \ldots b_{k_1}$. Asking a question (i.e. decoding) will be denoted with "$\sharp$". The *size* of a (multi)vector means the number of memory cells it occupies in computer's memory, while the *magnitude* of a (multi)vector $V = \{v_1, \ldots, v_n\}$ is its Euclidean norm $\sqrt{\sum_{i=1}^{n} v_i^2}$.

For our purposes it is important that geometric calculus allows us to define in a very systematic fashion a hierarchy of associative, non-commutative, and invertible operations that can be performed on 2^n-tuples. The resulting

superpositions are less noisy than the ones based on convolutions, say. Such operations are in general unknown to a wider audience, which explains popularity of tensor and convolution algebras. Geometric product preserves dimensionality at the level 2^n-dimensional *multivectors*, where n is the number of bits indexing basis vectors. Moreover, all nonzero vectors are invertible with respect to geometric product, a property absent for convolutions and important for unbinding and recognition. A detailed analysis of links between GA_c, HRR and BSC can be found in [5]. In particular, it is shown that both GA_c model and BSC are based on two different representations (in group theoretical sense) of the additive group of binary n-tuples with addition modulo 2. Actually, the latter observation was the starting point for studying geometric algebra forms of reduced representations [6].

3 The GA_c Model

Multivector (3) associated with n-dimensional Euclidean space can be represented by the 2^n-tuple $(\psi_{0_1\ldots 0_n}, \ldots, \psi_{1_1\ldots 1_n})$. Geometric product of two such 2^n-tuples is again a 2^n-tuple. In this sense geometric product is analogous to bindings employed in HRR or BSC, but we can still proceed in several inequivalent ways. For example, since a product of two basis blades is again a basis blade multiplied by ± 1, we can require that $\psi_{x_1\ldots x_n} = \pm 1$. Such a discrete version of GA HRR was tested vs. HRR and BSC in [19], and will be denoted here by GA_d (discrete GA HRR).

The continuous GA_c model differs greatly from GA_d. First of all, we do not begin with a general 2^n-dimensional multivector. Atomic objects are real-valued vectors in n-dimensional Euclidean space, in practice represented by n-tuples of components taken in some basis. A hierarchy of multivectors is reserved for complex statements, formed by binding and superposition of atomic objects. An n-dimensional vector, when seen from the multivector perspective, is a highly sparse 2^n-tuple: Only n out of 2^n components can be nonzero.

The procedure we employ was suggested in [5]. The space of 2^n-tuples is split into subspaces corresponding to scalars (0-vectors), vectors (1-vectors), bivectors (2-vectors), and so on. At the bottom of the hierarchy lay vectors $V \in \mathbb{R}^n$, having rank 1 and being denoted as $\overset{1}{V}$. An object of rank 2 is created by multiplying two elements of rank 1 with the help of the geometric product. Let $\overset{1}{V} = \{\alpha_1, \alpha_2, \alpha_3\}$ and $\overset{1}{W} = \{\beta_1, \beta_2, \beta_3\}$ be vectors in $\mathbb{R}^3$. A multivector $\overset{2}{X}$ of rank 2 in $\mathbb{R}^3$ comprises the following elements (cf. [18])

$$\overset{2}{X} = \overset{1}{V}\overset{1}{W} = \begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix} = \begin{bmatrix}\alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_3\beta_3\\ \alpha_1\beta_2 - \alpha_2\beta_1\\ \alpha_1\beta_3 - \alpha_3\beta_1\\ \alpha_2\beta_3 - \alpha_3\beta_2\end{bmatrix},$$

the first entry in the array on the right being a scalar and the remaining three entries being 2-blades. For arbitrary vectors in $\mathbb{R}^n$ we would have obtained one scalar (or, more conviniently: $\binom{n}{0}$ scalars) and $\binom{n}{2}$ 2-blades.

Let $\overset{2}{X} = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ and $\overset{1}{V} = \{\alpha_1, \alpha_2, \alpha_3\}$ be two multivectors in $\mathbb{R}^3$. A multivector $\overset{3}{Z}$ of rank 3 in $\mathbb{R}^3$ may be created in two ways: as a result of multiplying either $\overset{1}{V}$ by $\overset{2}{X}$ or $\overset{2}{X}$ by $\overset{1}{V}$. Let us concentrate on the first case

$$\overset{3}{Z} = \overset{1}{V}\overset{2}{X} = \begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix}\begin{bmatrix}\gamma_1\\ \gamma_2\\ \gamma_3\\ \gamma_4\end{bmatrix} = \begin{bmatrix}\alpha_1\gamma_1 - \alpha_2\gamma_2 - \alpha_3\gamma_3\\ \alpha_1\gamma_2 + \alpha_2\gamma_1 - \alpha_3\gamma_4\\ \alpha_1\gamma_3 + \alpha_2\gamma_4 + \alpha_3\gamma_1\\ \alpha_1\gamma_4 - \alpha_2\gamma_3 + \alpha_3\gamma_2\end{bmatrix}.$$

Here, the first three entries in the resulting matrix are 1-blades, while the last entry is a 3-blade. For arbitrary multivectors of rank 1 and 2 in $\mathbb{R}^n$ we would have obtained $\binom{n}{1}$ vectors and $\binom{n}{3}$ trivectors. We cannot generate multivectors of rank higher than 3 in $\mathbb{R}^3$, but it is easy to check that in spaces $\mathbb{R}^{n>3}$ a multivector of rank 4 would have $\binom{n}{0}$ scalars, $\binom{n}{2}$ bivectors and $\binom{n}{4}$ 4-blades. The number of k-blades in a multivector of rank r is described by Table 1. It becomes clear that a multivector of rank r over $\mathbb{R}^n$ is actually a vector over a $\sum_{i=0}^{\lfloor \frac{r}{2} \rfloor}\binom{n}{2i + r \bmod 2}$-dimensional space.

As an example let us consider the following roles and fillers being normalized vectors drawn randomly from $\mathbb{R}^n$ with Gaussian distribution $N(0, \frac{1}{n})$

$$\begin{array}{rclcrcl} Pat &=& \{a_1, \ldots, a_n\}, & & name &=& \{x_1, \ldots, x_n\},\\ male &=& \{b_1, \ldots, b_n\}, & & sex &=& \{y_1, \ldots, y_n\},\\ 66 &=& \{c_1, \ldots, c_n\}, & & age &=& \{z_1, \ldots, z_n\}. \end{array}$$

$PSmith$, who is a 66 year old male named Pat, is created by first multiplying roles and fillers with the help of the geometric product

$$\begin{aligned} PSmith &= \\ &= name * Pat + sex * male + age * 66\\ &= name \cdot Pat + name \wedge Pat + sex \cdot male +\\ &\quad sex \wedge male + age \cdot 66 + age \wedge 66\\ &= \begin{bmatrix} \sum_{i=1}^{n}(a_i x_i + b_i y_i + c_i z_i)\\ a_1x_2 - a_2x_1 + b_1y_2 - b_2y_1 + c_1z_2 - c_2z_1\\ a_1x_3 - a_3x_1 + b_1y_3 - b_3y_1 + c_1z_3 - c_3z_1\\ \vdots\\ a_{n-1}x_n - a_nx_{n-1} + b_{n-2}y_n\\ -b_ny_{n-1} + c_{n-1}z_n - c_nz_{n-1}\end{bmatrix}\\ &= [d_0, d_{12}, d_{13}, \ldots, d_{(n-1)n}]^T\\ &= d_0 + d_{12}e_{12} + d_{13}e_{13} + \cdots + d_{(n-1)n}e_{(n-1)n}, \end{aligned}$$

where $e_1, \ldots, e_n$ are orthonormal basis blades. In order to be decoded as much correctly as possible, $PSmith$

Table 1: Numbers of k-blades in multivectors of various ranks in $\mathbb{R}^n$

rank	scalars	vectors	bivectors	trivectors	4-blades	...	data size
1	0	$\binom{n}{1}$	0	0	0	...	$O(\binom{n}{1})$
2	$\binom{n}{0}$	0	$\binom{n}{2}$	0	0	...	$O(\binom{n}{0}+\binom{n}{2})$
3	0	$\binom{n}{1}$	0	$\binom{n}{3}$	0	...	$O(\binom{n}{1}+\binom{n}{3})$
4	$\binom{n}{0}$	0	$\binom{n}{2}$	0	$\binom{n}{4}$	...	$O(\binom{n}{0}+\binom{n}{2}+\binom{n}{4})$
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
$2r$	$\binom{n}{0}$	0	$\binom{n}{2}$	0	$\binom{n}{4}$	...	$O(\sum_{i=0}^{r}\binom{n}{2i})$
$2r+1$	0	$\binom{n}{1}$	0	$\binom{n}{3}$	0	...	$O(\sum_{i=0}^{r}\binom{n}{2i+1})$

should have the same magnitude as vectors representing atomic objects, therefore it needs to be normalized. Finally, $PSmith$ takes the form of

$$PSmith \;=\; [\hat{d}_0, \hat{d}_{12}, \hat{d}_{13}, \ldots, \hat{d}_{(n-1)n}]^T,$$

where $\hat{d}_i = \frac{d_i}{\sqrt{\sum_{j=0,12}^{(n-1)n} d_j^2}}$.

$PSmith$ is now a multivector of rank 2. The decoding operation

$$\begin{aligned} & name^+ PSmith \\ &= name^+(name \cdot Pat + name \wedge Pat \\ &\quad + sex \cdot male + sex \wedge male + age \cdot 66 \\ &\quad + age \wedge 66) \end{aligned}$$

will produce a multivector of rank 3 consisting of vectors and trivectors. However, the original Pat did not contain any trivector components — they all belong to the noise part and the only interesting blades in $name^+PSmith$ are vectors. The expected answer is a vector, therefore there is no point in calculating the whole multivector $name^+PSmith$ and only then comparing it with items stored in the clean-up memory. To be efficient, one should generate only the vector-part while computing $name^+PSmith$ and skip the noisy trivectors.

Let $\langle\cdot\rangle_k$ denote the projection of a multivector on k-blades. To decode $PSmith$'s $name$ we need to compute

$$\begin{aligned} & \langle name^+ PSmith\rangle_1 \\ &= name^+ namePat + \langle\, name^+(name \wedge Pat \\ &\quad + sex \cdot male + sex \wedge male + age \cdot 66 \\ &\quad + age \wedge 66)\,\rangle_1 \\ &= Pat + \text{noise} = Pat'. \end{aligned}$$

The resulting Pat' will still be noisy, but to a lesser degree than it would have been if the trivectors were present.

Formally, we are using a map $*^1_{1,2}$ that transforms a multivector of rank 1 (i.e. an n-tuple) and a multivector of rank 2 (i.e. a $(1+\frac{(n-1)n}{2})$-tuple) into a multivector of rank 1 without computing the unnecessary blades. Let X be a multivector of rank 2

$$X = \langle X\rangle_0 + \langle X\rangle_2 = x_0 + \sum_{l<m} x_{lm} e_l e_m,$$

where $x_{lm} = -x_{ml}$. If $A = (A_1, \ldots, A_n)$ is a decoding vector (actually, an inverse of a role vector), then

$$\begin{aligned} A *^1_{1,2} X &= x_0 A + \sum_{l,m} A_l x_{lm} e_m \\ &= \sum_k \big(x A_k + \sum_l A_l x_{lk}\big) e_k \\ &= \sum_k Y_k e_k = Y, \end{aligned}$$

with $Y = (Y_1, \ldots, Y_n)$ being an n-tuple, i.e. a multivector of rank 1. More explicitly,

$$Y_k = (A *^1_{1,2} X)_k = x_0 A_k + \sum_{l=1}^{k-1} A_l x_{lk} - \sum_{l=k+1}^{n} A_l x_{kl}.$$

The map $*^1_{1,2}$ is an example of a *projected product*, introduced in [5], reconstructing the vector part of AX without computing the unnecessary parts. The projected product is basis independent, as opposed to circular convolutions. In general, $*^m_{l,k}$ transforms the geometric product of two multivectors $\overset{l}{A}$ and $\overset{k}{B}$ into a multivector $\overset{m}{C}$.

We now need to compare Pat' with other items stored in the clean-up memory using the dot product, and since Pat' is a vector, we need to compare only the vector part. That means, if the clean-up memory contained a multivector $\overset{2t+1}{M}$ of an odd rank, we would also need to compute $Pat' \cdot \langle \overset{2t+1}{M} \rangle_1$ while searching for the right answer.

This method of decoding suggests that items stored in the clean-up memory should hold information about their ranks, which is dangerously close to employing fixed data slots present in localist architectures. However, a rank of a clean-up memory item can be "guessed" from its size. In a distributed model we also should not "know" for sure how many parts the projected product should reject, but it can certainly reject parts spanned by blades of highest grades. Unfortunately, since the geometric product is non-commutative, questions concerning roles and fillers need to be asked on different sides of a sentence, forcing atomic objects to hold information on whether they are roles or fillers and thus, forcing them to be partly hand-generated. We can either ask question always on the same side of a sentence and be satisfied with less precise answers or always ask

about only the roles or only the fillers. It becomes clear, that recognition based on the hierarchy of multivectors and the projected product is best applicable to tasks in which questions need to be asked only on one side of the sentence or in which sentences have predetermined structure.

Before providing formulas for encoding and decoding a complex statement we need to introduce additional notation for the projected product and the projection. We have already introduced the projected product $*^m_{l,k}$ transforming the geometric product of two multivectors of ranks l and k into a multivector of rank m. This will not always be the case for complex statements, since we can produce a multivector that will not be of any given rank. Let $*^m_{l,\{\alpha_1,\alpha_2,\ldots,\alpha_k\}}$ denote the projected product transforming the geometric product of a multivector $\overset{l}{A}$ and a multivector B containing α_1-blades, α_2-blades,... and α_k-blades into a multivector $\overset{m}{C}$. In this way, the projected product $*^1_{1,2}$ may be written down as $*^1_{1,\{0,2\}}$. By analogy, let $\langle\cdot\rangle_{\{\alpha_1,\alpha_2,\ldots,\alpha_k\}}$ denote the projection of a multivector on components spanned by α_1-blades, α_2-blades,... and α_m-blades.

Let Ψ denote the normalized multivector encoding the sentence "*Fido bit PSmith*", i.e.

$$\Psi = \underbrace{bite_{agt} * Fido}_{\text{rank 2}} + \underbrace{bite_{obj} * \underbrace{PSmith}_{\text{rank 2}}}_{\text{rank 3}}.$$

Multivector Ψ will contain scalars, vectors, bivectors and trivectors and can be written down as the following vector of dimension $\sum_{i=0}^{3}\binom{n}{i}$

$$\begin{aligned}\Psi &= \underbrace{\alpha}_{\text{a scalar}} + \underbrace{\sum_{i=1}^{n}\beta_i e_i}_{\text{vectors}} + \underbrace{\sum_{1=i<j}^{n}\gamma_{ij}e_{ij}}_{\text{bivectors}} \\ &\quad + \underbrace{\sum_{1=i<j<k}^{n}\delta_{ijk}e_{ijk}}_{\text{trivectors}}.\end{aligned}$$

4 More Examples of Encoding and Decoding Sentences

The following examples illustrate various ways of asking questions in the GA$_c$ architecture.

"*Who was bitten?*"
The answer to that question will be a multivector of rank 2

$$\begin{aligned}\Psi \sharp bite_{obj} &= \langle bite^+_{obj}\Psi\rangle_{\{0,2\}} \\ &= bite^+_{obj} *^2_{1,\{0,1,2,3\}} \Psi \\ &= PSmith' \approx PSmith.\end{aligned}$$

Let $bite_{obj} = \{y_1,\ldots,y_n\}$, $PSmith'$ will then have the form

$$\begin{aligned}PSmith' &= (y_1e_1+\cdots+y_ne_n) *^2_{1,\{0,1,2,3\}} \\ &\quad (\sum_{i=1}^{n}\beta_i e_i + \sum_{1=i<j<k}^{n}\delta_{ijk}e_{ijk}) \\ &= \underbrace{\sum_{k=1}^{n} y_k\beta_k}_{\text{a scalar}} + \underbrace{\sum_{1=i<j}^{n}\theta_{ij}e_{ij}}_{\text{bivectors}},\end{aligned}$$

where

$$\theta_{ij} = y_i\beta_j - y_j\beta_i + \sum_{\substack{t=1\\ t\notin\{i,j\}}}^{n} y_t\delta_{ijt}$$

with $\delta_{ijt} = \delta_{tij} = -\delta_{itj}$. As previously, $PSmith'$ should be compared with appropriate items from the clean-up memory to produce the most probable answer.

"*What happened to PSmith?*"
Asking about roles poses a problem of inverting a (multi)vector. Since not all multivectors are invertible, we have to be satisfied with reverses [5] of multivectors. We will need another type of a projected product: let $*^m_{\{\alpha_1,\alpha_2,\ldots,\alpha_l\},k}$ denote the projected product transforming the geometric product of a multivector B containing α_1-blades, α_2-blades,... and α_l-blades and a multivector $\overset{k}{A}$ into a multivector $\overset{m}{C}$. The answer to our question will be a vector

$$\begin{aligned}\Psi \sharp PSmith &= \langle \Psi PSmith^+\rangle_1 \\ &= \Psi *^1_{\{0,1,2,3\},2} PSmith^+ \\ &= bite'_{obj} \approx bite_{obj}.\end{aligned}$$

If we denote $PSmith$ as

$$PSmith = z_0 + z_{12}e_{12} + \cdots + z_{(n-1)n}e_{(n-1)n}$$

then

$$\begin{aligned}bite'_{obj} &= (\sum_{i=1}^{n}\beta_i e_i + \sum_{1=i<j<k}^{n}\delta_{ijk}e_{ijk}) \\ &\quad *^1_{\{0,1,2,3\},2}(z_0 - \sum_{1=i<j}^{n} z_{ij}e_{ij}) \\ &= \zeta_1e_1+\cdots+\zeta_ne_n,\end{aligned}$$

where

$$\zeta_k = \beta_k z_0 - \sum_{i=1}^{k-1}\beta_i z_{ik} + \sum_{i=k+1}^{n}\beta_i z_{ki} - \sum_{\substack{1=i<j\\ i,j\neq k}}^{n}\delta_{ijk}z_{ij},$$

with $\delta_{ijk} = \delta_{kij} = -\delta_{ikj}$.

"What did Fido do?"

The last question in this example will produce an answer having the form of a vector

$$\begin{aligned}\Psi \sharp Fido &= \langle \Psi Fido^{+} \rangle_1 \\ &= \Psi *^{1}_{\{0,1,2,3\},1} Fido^{+} \\ &= bite'_{agt} \approx bite_{agt}.\end{aligned}$$

If $Fido = \{v_1, \ldots, v_n\}$, then

$$\begin{aligned}bite'_{agt} &= (\alpha + \gamma_{12}e_{12} + \cdots + \gamma_{(n-1)n}e_{(n-1)n}) \\ &\quad *^{1}_{\{0,1,2,3\},1}(v_1e_1 + \cdots + v_ne_n) \\ &= \vartheta_1 e_1 + \ldots \vartheta_n e_n,\end{aligned}$$

where

$$\vartheta_k = \alpha v_k - \sum_{i=1}^{k-1} \gamma_{ik} v_i + \sum_{i=k+1}^{n} \gamma_{ki} v_i.$$

Those tedious calculations imply that the GA_c model is best applicable to sentences having a similar or identical complexity structure, otherwise it may be hard to automatize the process of asking questions and retrieving answers. Because of this limitation, this construction seems to be a promising candidate for a holographic database.

5 Overview of Plate's Scaling Test

Plate [23] describes a simulation in which approximately 5000 HRR 512-dimensional vectors were stored in the clean-up memory. The purpose of his simulation was to study efficiency of the HRR model but also to provide a counterexample to the claim that role-filler representations do not permit one component of a relation to be retrieved given the others. We will repeat Plate's test on several models and compare the results.

Let us consider the following atomic objects

$$\left.\begin{array}{l} num_x\ (x = 0, \ldots, 2500), \\ times, \\ plus, \end{array}\right\} \text{fillers},$$

$$\left.\begin{array}{l} result, \\ operand. \end{array}\right\} \text{roles}$$

At the beginning of the scaling test, relations concerning multiplication and addition are constructed. For example, "$2 \cdot 3 = 6$" is constructed as

$$\begin{aligned}times_{2,3} &= times + operand * (num_2 + num_3) \\ &\quad + result * num_6.\end{aligned}$$

Generally, relations are constructed in the following way

$$\begin{aligned}times_{x,y} &= times + operand * (num_x + num_y) \\ &\quad + result * num_{x \cdot y}, \\ plus_{x,y} &= plus + operand * (num_x + num_y) \\ &\quad + result * num_{x+y}.\end{aligned}$$

x and y range from 0 to 50 with $y \leq x$ making a total of 2501 number vectors and 2652 instances of each $times_{x,y}$ and $plus_{x,y}$. As one can notice, the same $operand$ role is used for both x and y to preserve commutativity of multiplication and addition.

Plate writes, that a relation can be "looked up" by supplying enough information to distinguish a specific relation from others. For example, to look up "$2 \cdot 3 = 6$" one needs to find the most similar relation R to any of the following fragmentary statements

$$\begin{aligned}&\text{(case 1)} && times + operand * num_2 \\ &&& + operand * num_3, \\ &\text{(case 2)} && times + operand * num_2 \\ &&& + result * num_6, \\ &\text{(case 3)} && times + operand * num_3 \\ &&& + result * num_6, \\ &\text{(case 4)} && operand * num_2 + operand * num_3 \\ &&& + result * num_6.\end{aligned}$$

Retrieving the missing piece of information in the first three cases can be done by asking any of the subquestions

$$\begin{aligned}&\text{(case 1)} && R \sharp result, \\ &\text{(case 2)} && R \sharp operand, \\ &\text{(case 3)} && R \sharp operand.\end{aligned}$$

Case 4 is somewhat more problematic — to look up a missing relation name ($times$ or $plus$) one needs to have a separate clean-up memory containing only relation names or to use an alternative encoding in which there is a role for relation names. We will alter Plate's test by using the latter method.

Plate states that he had tried one run of the system making a query for each component missing in every relation — this amounted to 10608 queries. A further 7956 queries had been made to decode the missing component except for the relation name. Plate goes on to claim, that the system made no errors.

There appear to be two misstatements in Plate's claims. Firstly, we cannot treat subquestions regarding cases 2 and 3 separately, as there are two equally probable answers to each of these subquestions, provided that relations R_2 and R_3 point correctly to $times_{x,y}$. Secondly, consider a fragmentary piece of information

$$times + operand * num_0 + result * num_0.$$

In this situation, the missing component can be any of the numbers num_x where $x \in \{0, \ldots, 50\}$ and thus, there are 51 atomic objects that are equally probable to be the right answer. This suggests that Plate regards several answers as valid ones, as long as the similarity of these answers exceeds some threshold. To work out the missing component, one then needs to check which of those potential answers is not in the original set used for retrieval.

Such a method of investigating scaling properties has more than a few advantages:

– Inaccuracies mentioned above act as a test if all atomic objects are created and treated equally. Ideally, every atomic object of the num_x form should be recognized as a correct answer to the "zero problem" for $\frac{\text{number of trials}}{51} \cdot 100\%$ of the time.

– Prime numbers greater than 100 do not appear in any of $times_{x,y}$ and $plus_{x,y}$ relations, therefore they test if the model is immune to garbage data.

– Numbers ranging from num_0 to num_{100} may be constructed in a multitude of ways by addition (num_0 by multiplication) and any given sentence chunk $result*$ num_z will appear quite often in the $plus_{x,y}$ relation. Hence, this is a great way of checking if the model deals with excessive similarity of a number of complex statements.

– Atomic objects bound with $operand$ and $result$ range in variety. On the other hand, there are just two atomic objects acting as an $operation$ — does it affect in any way the recognition of $operation$ filler? Indeed, it will be shown in Section 7 that recognition of the $operation$ chunk turns out to be quite interesting depending on the choice of the architecture.

6 Notation

For the purpose of explaining test results, let us introduce the following notation. Let $S^*_{x,y}$ and $S^+_{x,y}$ denote relations

$$\begin{aligned} S^*_{x,y} &= operation * times + operand * num_x + \\ &\quad operand * num_y + result * num_{x \cdot y}, \\ S^+_{x,y} &= operation * plus + operand * num_x + \\ &\quad operand * num_y + result * num_{x+y}, \end{aligned}$$

for $y \leq x$. We chose to use a separate role for a relation name to enable encoding the information given only operands and the result. Let $F^{op}_{i,x,y}$ denote fragmentary statements for $i \in \{1,2,3,4\}$ and $op \in \{*,+\}$

$$\begin{aligned} F^{op}_{1,x,y} &= S^{op}_{x,y} - result * num_{x\ op\ y}, \\ F^{op}_{2,x,y} &= S^{op}_{x,y} - operand * num_x, \\ F^{op}_{3,x,y} &= S^{op}_{x,y} - operand * num_y, \\ F^{op}_{4,x,y} &= S^{op}_{x,y} - operation * op. \end{aligned}$$

If v is an element of the clean-up memory, then let $N(v)$ denote the closest $neighbor$ of v, i.e. an element of the clean-up memory that is most similar to v. If v has more than one neighbor, then all subquestions during the test are asked to all of v's neighbors. In HRR, GA_d (with the Hamming measure of similarity) and GA_c it is extremely unlikely for an element of the clean-up memory to have more than one neighbor due to the continuous nature of data in these architectures. Let $Q^{op}_{i,x,y} = N(F^{op}_{i,x,y})$ for $i \in \{1,2,3,4\}$ and $op \in \{*,+\}$. During the test we asked subquestions concerning components missing in $F^{op}_{i,x,y}$ and obtained the following (sets of) answers

$$\begin{aligned} q^{op}_{1,x,y} &= N(Q^{op}_{1,x,y} \sharp result), \\ q^{op}_{2,x,y} &= N(Q^{op}_{2,x,y} \sharp operand), \\ q^{op}_{3,x,y} &= N(Q^{op}_{3,x,y} \sharp operand), \\ q^{op}_{4,x,y} &= N(Q^{op}_{4,x,y} \sharp operation). \end{aligned}$$

We assume that a missing component is identified correctly if it is the only neighbor to appropriate answer $q^{op}_{\cdot,x,y}$ or it belongs to the set of neighbors of $q^{op}_{\cdot,x,y}$.

7 Test Results

The software for all tests was developed by A. Patyk-Łońska in Java language. All tests were performed on an ordinary PC with dualcore AMD processor with 2 GB RAM.

Tables 2 through 4 compare scaling test results for

– GA_c and HRR, both using dot-product as a similarity measure.

– BSC using Hamming distance as a similarity measure.

Although BSC and HRR models need only n-dimensional vectors, this is not quite the case for and GA_c, which needs $1 + \frac{n(n-1)}{2}$ numbers to represent multivectors of rank 2 over $\mathbb{R}^n$. We present recognitions test results close to 100% and comment on vector length required for each model to achieve such percentage. The real number of memory cells used up by each model is given in brackets in the table headings.

The answers to subquestions $Q^{op}_{2,x,y} \sharp operand$ and $Q^{op}_{3,x,y} \sharp operand$ were considered to be correct if any of the two possible operands came up as the item most similar to those subquestions. In case of other questions and subquestions only exact answers were taken into consideration.

50 runs of the test were performed on each model. Unlike in Plate's test, x and y ranged from 0 to only 20. Hence, there are 401 number vectors and 462 relation vectors.

The "zero problem" is clearly visible in each tested model, as the recognition percentage of $Q^*_{3,x,y}$ barely exceeds 90%. Nevertheless, $Q^*_{3,x,y}$ almost always contains at least one of the operands from the original sentence $S^*_{x,y}$ since the recognition percentage of $q^*_{3,x,y}$ reaches 100% for sufficiently large data size. On the whole, the recognition percentage of $q^*_{2,x,y}$ and $q^*_{3,x,y}$ does not differ greatly from the recognition percentage of $q^+_{2,x,y}$ and $q^+_{3,x,y}$ in any model.

Table 2: Recognition percentage for GA_c.

Questions	$\mathbf{R}^{10}$ (46)	$\mathbf{R}^{20}$ (191)	$\mathbf{R}^{30}$ (436)	$\mathbf{R}^{40}$ (781)
$Q^*_{1,x,y}$	89.76%	99.98%	99.99%	100.0%
$q^*_{1,x,y}$	39.44%	95.28%	99.58%	99.88%
$Q^*_{2,x,y}$	91.12%	99.73%	99.98%	100.0%
$q^*_{2,x,y}$	36.24%	83.86%	97.92%	99.81%
$Q^*_{3,x,y}$	83.97%	91.15%	91.33%	91.34%
$q^*_{3,x,y}$	41.27%	84.92%	98.05%$^{\Delta}$	99.82%$^{\Delta}$
$Q^*_{4,x,y}$	98.90%	99.60%	99.63%	99.59%
$q^*_{4,x,y}$	42.01%	95.56%	99.24%	99.52%
$Q^+_{1,x,y}$	89.39%	99.99%	100.0%	100.0%
$q^+_{1,x,y}$	39.09%	95.99%	99.76%	99.95%
$Q^+_{2,x,y}$	86.96%	99.59%	99.96%	100.0%
$q^+_{2,x,y}$	35.32%	83.84%	97.97%	99.79%
$Q^+_{3,x,y}$	87.00%	99.63%	99.96%	100.0%
$q^+_{3,x,y}$	35.12%	83.84%	97.98%	99.79%
$Q^+_{4,x,y}$	99.05%	99.53%	99.51%	99.54%
$q^+_{4,x,y}$	45.84%	94.73%	99.14%	99.49%

Table 3: Recognition percentage for HRR.

Questions	$N = 200$	$N = 300$	$N = 400$	$N = 500$
$Q^*_{1,x,y}$	29.1%	27.06%	26.28%	28.51%
$q^*_{1,x,y}$	31.08%$^{\Delta}$	30.03%$^{\Delta}$	30.30%$^{\Delta}$	32.23%$^{\Delta}$
$Q^*_{2,x,y}$	54.72%	52.06%	53.10%	53.32%
$q^*_{2,x,y}$	98.99%$^{\Delta}$	99.92%$^{\Delta}$	99.98%$^{\Delta}$	100.0%$^{\Delta}$
$Q^*_{3,x,y}$	50.53%	47.93%	49.80%	51.21%
$q^*_{3,x,y}$	98.92%$^{\Delta}$	99.90%$^{\Delta}$	99.97%$^{\Delta}$	100.0%$^{\Delta}$
$Q^*_{4,x,y}$	89.23%	90.56%	90.51%	90.29%
$q^*_{4,x,y}$	90.28%$^{\Delta}$	92.69%$^{\Delta}$	92.42%$^{\Delta}$	92.31%$^{\Delta}$
$Q^+_{1,x,y}$	28.26%	29.46%	28.03%	28.81%
$q^+_{1,x,y}$	27.32%	29.37%	28.02%	28.80%
$Q^+_{2,x,y}$	53.91%	54.48%	55.26%	54.68%
$q^+_{2,x,y}$	98.72%$^{\Delta}$	99.90%$^{\Delta}$	99.99%$^{\Delta}$	99.99%$^{\Delta}$
$Q^+_{3,x,y}$	53.73%	55.23%	55.34%	54.62%
$q^+_{3,x,y}$	98.67%$^{\Delta}$	99.91%$^{\Delta}$	99.98%$^{\Delta}$	100.0%$^{\Delta}$
$Q^+_{4,x,y}$	98.70%	98.75%	98.66%	98.75%
$q^+_{4,x,y}$	97.16%	98.55%	98.64%	98.74%

Table 4: Recognition percentage for BSC.

Questions	$N = 200$	$N = 300$	$N = 400$	$N = 500$
$Q^*_{1,x,y}$	86.71%	91.65%	93.78%	94.74%
$q^*_{1,x,y}$	82.82%	90.62%	93.87%$^\Delta$	94.95%$^\Delta$
$Q^*_{2,x,y}$	94.42%	97.60%	99.03%	99.44%
$q^*_{2,x,y}$	99.68%$^\Delta$	99.97%$^\Delta$	99.98%$^\Delta$	100.0%$^\Delta$
$Q^*_{3,x,y}$	86.87%	89.43%	90.50%	90.97%
$q^*_{3,x,y}$	99.15%$^\Delta$	99.47%$^\Delta$	99.65%$^\Delta$	100.0%$^\Delta$
$Q^*_{4,x,y}$	94.39%	95.58%	95.39%	95.50%
$q^*_{4,x,y}$	90.78%	94.89%	95.22%	95.44%
$Q^+_{1,x,y}$	86.38%	91.59%	93.65%	94.71%
$q^+_{1,x,y}$	81.71%	90.28%	93.27%	94.57%
$Q^+_{2,x,y}$	94.23%	97.77%	99.19%	99.52%
$q^+_{2,x,y}$	99.36%$^\Delta$	99.94%$^\Delta$	100.0%$^\Delta$	100.0%$^\Delta$
$Q^+_{3,x,y}$	94.54%	97.39%	98.77%	99.48%
$q^+_{3,x,y}$	99.41%$^\Delta$	99.94%$^\Delta$	100.0%$^\Delta$	100.0%$^\Delta$
$Q^+_{4,x,y}$	95.40%	95.38%	95.65%	95.66%
$q^+_{4,x,y}$	91.81%	94.27%	95.02%	95.27%

Table entries marked with a "$^\Delta$" indicate that despite the wrong recognition of a fragmentary sentence, the missing component has been identified correctly. In all tested models such situations arise for sentences with one of the operands missing. For HRR, however the missing item has been "accidentally" correctly identified also in cases of missing $operation * times$ and $result * times_{x,y}$ components. Such recognition did not occur in cases of missing $operation * plus$ and $result * plus_{x,y}$ components, which is distressingly asymmetric.

HRR turned out to be the worst model during this experiment. The recognition percentage of $Q^*_{1,x,y}$ and $Q^+_{1,x,y}$ is dangerously low when compared to other Q's. Both $Q^*_{1,x,y}$ and $Q^+_{1,x,y}$ are retrieved from the clean-up memory given only two operands and the operation type. Since we have only two operation types, $Q^*_{1,x,y}$ and $Q^+_{1,x,y}$ will not differ greatly from each other. This phenomenon is also observable in BSC (but not in GA$_c$), where the recognition percentage of Q_1's is only slightly lower than that of the other Q's. Apart from that weakness, BSC performs as well as GA$_c$ for adequate data size.

8 Conclusion

Authors developed a new model of distributed representations of data based on geometric algebra. Although the data representations of sentences encoded in this model may have varying lengths (as opposed to HRR and BSC), it can be justified by the fact that it is quite logical for sentences that hold more information to have larger "volume".

Tedious calculations presented in Section 3 imply that the GA$_c$ model is best applicable to sentences having a similar or identical complexity structure, otherwise it may be hard to make the process of asking questions and retrieving answers automatic. Because of this limitation, this construction seems to be a promising candidate for a holographic database.

Although research in distributed representations has been thriving in the past decades, no one has yet developed a software tool that would employ distributed representations to implement databases with real-life contents. Of course, some attempts at scaling has been made so far, but they were rather narrowly aimed at specific tasks. Authors hope to develop such a tool in the (near) future.

9 Further Perspectives – Quantum-like Computation Based on Geometric Algebra

Quantum algorithms [17] employ tensor product binding and thus are analogous to Smolensky's tensor product representations [25]. The peculiarity of quantum computation is in its putative implementation: hardware based on the rules of micro-world automatically guarantees parallelism of processing the entire superposition of bound objects. The same property, however, makes quantum processors extremely sensitive to noise so it is by no means evident that working devices will be practically constructed.

The question is if we really have to look for micro-world implementations of quantum computation. Replacing tensor products by geometric products one obtains a one-to-one map between quantum mechanical superpositions and multivectors [2, 4], and all elementary quantum gates have geometric analogues [3]. This proves that quantum algorithms can be, in principle, implemented in systems described by geometric algebra.

Acknowledgement

This work was supported by *Grant G.0405.08* of the *Fund for Scientific Research Flanders*.

References

[1] D. Aerts and M. Czachor (2004), "Quantum aspects of semantic analysis and symbolic artificial intelligence", *J. Phys. A*, vol. 37, pp. L123-L13.

[2] D. Aerts and M. Czachor (2007), "Cartoon computation: Quantum-like algorithms without quantum mechanics", *J. Phys. A*, vol. 40, pp. F259-F266.

[3] M. Czachor (2007), "Elementary gates for cartoon computation", *J. Phys. A*, vol. 40, pp. F753-F759.

[4] D. Aerts and M. Czachor (2008), "Tensor-product versus geometric-product coding", *Physical Review A*, vol. 77, id. 012316.

[5] D. Aerts, M. Czachor, and B. De Moor (2009), "Geometric Analogue of Holographic Reduced Representation", *J. Math. Psychology*, vol. 53, pp. 389-398.

[6] D. Aerts, M. Czachor, and B. De Moor (2006), "On geometric-algebra representation of binary spatter codes". preprint arXiv:cs/0610075 [cs.AI].

[7] D. Aerts, M. Czachor, and Ł. Orłowski (2009), "Teleportation of geometric structures in 3D ", *J. Phys. A* vol. 42, 135307.

[8] W.K. Clifford (1878), "Applications of Grassmann's extensive algebra", *American Journal of Mathematics Pure and Applied*, vol. 1, 350–358.

[9] R. W. Gayler (1998), "Multiplicative binding, representation operators, and analogy", *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, K. Holoyak, D. Gentner, and B. Kokinov, eds., Sofia, Bulgaria: New Bulgarian University, p. 405.

[10] H. Grassmann (1877), "Der Ort der Hamilton'schen Quaternionen in der Ausdehnungslehre", *Mathematische Annalen*, vol. 3, 375–386.

[11] G. E. Hinton, J. L. McClelland and D. E. Rumelhart (1986), ""Parallel distributed processing: Explorations in the microstructure of cognition", vol. 1, 77Ű109, *"Distributed representations"*, The MIT Press, Cambridge, MA.

[12] P. Kanerva (1996), "Binary spatter codes of ordered k-tuples". In C. von der Malsburg et al. (Eds.), *Artificial Neural Networks ICANN Proceedings, Lecture Notes in Computer Science* vol. 1112, pp. 869-873.

[13] P. Kanerva (1997), "Fully distributed representation". *Proc. 1997 Real World Computing Symposium (RWCŠ97, Tokyo)*, pp. 358-365.

[14] E. M. Kussul (1992), *Associative Neuron-Like Structures*. Kiev: Naukova Dumka (in Russian).

[15] E.M. Kussul and T.N. Baidyk (1990), "Design of Neural-Like Network Architecture for Recognition of Object Shapes in Images", *Soviet J. Automation and Information Sciences*, vol. 23, no. 5, pp. 53-58.

[16] N.G. Marchuk, and D.S. Shirokov (2008), "Unitary spaces on Clifford algebras", *Advances in Applied Clifford Algebras*, vol 18, pp. 237-254.

[17] M.A. Nielsen and I.L. Chuang (2000), *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.

[18] A. Patyk (2010), "Geometric Algebra Model of Distributed Representations", in *Geometric Algebra Computing in Engineering and Computer Science*, E. Bayro-Corrochano and G. Scheuermann, eds. Berlin: Springer. Preprint arXiv:1003.5899v1 [cs.AI].

[19] T. Plate (1995), "Holographic Reduced Representations", *IEEE Trans. Neural Networks*, vol. 6, no. 3, pp. 623-641.

[20] T. Plate (2003), *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. CSLI Publications, Stanford.

[21] D.A. Rachkovskij (2001), "Representation and Processing of Structures with Binary Sparse Distributed Codes", *IEEE Trans. Knowledge Data Engineering*, vol. 13, no. 2, pp. 261-276.

[22] P. Smolensky (1990), "Tensor product variable binding and the representation of symbolic structures in connectionist systems". *Artificial Intelligence*, vol. 46, pp. 159-216.

Experiments on Preserving Pieces of Information in a Given Order in Holographic Reduced Representations and the Continuous Geometric Algebra Model

Agnieszka Patyk-Łońska
Gdańsk University of Technology
ul. Narutowicza 11/12, Gdańsk 80-233, Poland
patyk@pg.gda.pl, http://www.pg.gda.pl/ patyk

Keywords: distributed representations of data, geometric algebra, HRR, BSC, word order, trajectory associations, bag of words

Received: October 23, 2011

Geometric Analogues of Holographic Reduced Representations (GA_c, which is the continuous version of the previously developed discrete GA model) employ role-filler binding based on geometric products. Atomic objects are real-valued vectors in n-dimensional Euclidean space and complex statements belong to a hierarchy of multivectors. The property of GA_c and HRR studied here is the ability to store pieces of information in a given order by means of trajectory association. We describe results of three experiments: finding correct item or correct place of an item in a sequence and finding the alignment of items in a sequence without the precise knowledge of trajectory vectors.

Povzetek: Članek preučuje ohranitev informacij pri obliki holografske hrambe podatkov.

1 Introduction

The work presented here is a result of experimenting with trajectory association technique using the newly-developed distributed representation model GA_c [20].

An ideal distributed representation system needs to meet several criteria in order to successfully perform cognitive tasks. These include computational efficiency, noise tolerance, scaling and ability to represent complex structures. The most widely used definition of a distributed representation of data is due to Hinton *et al.* [13]: in a *distributed representation of data* each concept is represented over a number of units and each unit participates in the representation of some number of concepts. The size of a distributed representation is usually fixed and the units have either binary or continuous-space values. In most distributed representations only the overall pattern of activated units has a meaning.

Such patterns of activity are hard to understand and interpret, therefore they are often compared to greyscale images. Distributed representations usually take the form of one-dimensional vectors, while greyscale images are two-dimensional matrices, but the way the pixels are aligned (one-dimensional string or two-dimensional array) is of no relevance. Since the information is distributed over the elements of a vector, a great percentage of units ("pixels") can be changed without making the vector (overall "picture") unrecognizable.

Let us consider an example of storing the following information: "Fido bit Pat". The action in this statement is $bite$ and the features (i.e. $roles$) of this action are an agent and an object, denoted $bite_{agt}$ and $bite_{obj}$, while their $fillers$ are $Fido$ and Pat respectively. If we consider storing the way that the action is performed, we can add a third feature ($role$), e.g. $bite_{way}$. If we store $Fido$, Pat, $bite_{agt}$ and $bite_{obj}$ as vectors, we are able to encode "Fido bit Pat" as

$$bite_{agt} * Fido + bite_{obj} * Pat.$$

The operation of $binding$, denoted by "$*$", takes two vectors and produces another vector, often called a $chunk$ of a sentence. It would be ideal for the resulting vector not to be similar to the original vectors but to have the same dimensions as the original vectors. $Superposition$, denoted by "$+$", is an operation that takes any number of vectors and creates another one that is similar to the original vectors. Usually, the superimposed vectors are already the result of the binding operation.

For more details and examples on distributed representations of data the reader should refer to [20].

2 Preserving Pieces of Information in a Given Order

While some solutions to the problem of preserving pieces of information in a given order have proved ingenious, others are obviously flawed. Let us consider the representation of the word eye — it has three letters, one of which occurs

This paper is based on A. Patyk-Łońska *Preserivng pieces of information in a given order in HRR and GA_c* published in the proceedings of the 1st International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2011 conference).

twice. The worst possible choice of binding and superposition would be to store quantities of letters, e.g.

$$eye = twice * e + once * y,$$

since we would not be able to distinguish eye from eey or yee. Another ambiguous representation would be to remember the neighborhood of each letter

$$eye = before_y * e + between_e * y + after_y * e.$$

Unfortunately, such a method of encoding causes words eye and $eyeye$ to have the same representation

$$\begin{aligned} eyeye &= before_y * e + 2 \cdot between_e * y + \\ &\quad (before_y + after_y) * e + after_y * e \\ &= 2\big(before_y * e + between_e * y \\ &\quad + after_y * e\big) \\ &= 2\, eye. \end{aligned}$$

Real-valued vectors are normalized in most distributed representation models, therefore the factor of 2 would be most likely lost in translation. Such $contextual\ roles$ (Smolensky [25]) cause problems when dealing with certain types of palindromes. Remembering positions of letters is also not a good solution

$$eye = letter_{first} * e + letter_{second} * y + letter_{third} * e$$

as we need to redundantly repeat the first letter as the third letter, otherwise we could not distinguish eye from ey or ye. Secondly, this method of encoding will not detect similarity between eye and $yeye$.

Pike argues in [21] that matrix-based memory is multidirectional, i.e. it allows both forward and backward association — having two vectors a and b and their binding $M = ab$ we can extract both a and b by performing a reverse operation on the appropriate side of the matrix. Convolution-correlation systems, on the other hand, regard bindings $a \circledast b$ and $b \circledast a$ as identical. We will use a similar technique, asking right-hand-side and left-hand-side questions during experiments described in the following sections.

A quantum-like attempt to tackle the problem of information ordering was made in [1] — a version of semantic analysis, reformulated in terms of a Hilbert-space problem, is compared with structures known from quantum mechanics. In particular, an LSA matrix representation ([1, 10]) is rewritten by the means of quantum notation. Geometric algebra has also been used extensively in quantum mechanics ([2, 4, 3]) and so there seems to be a natural connection between LSA and GA_c, which is the ground for fututre work on the problem of preserving pieces of information in a given order.

As far as convolutions are concerned, the most interesting approach to remembering information in a given order has been described in [12]. Authors present a model that builds a holographic lexicon representing both word meaning and word order from unsupervised experience with natural language texts comprising altogether 90000 words. This model uses simple convolution and superposition to construct n-grams recording the frequency of occurrence of every possible word sequence that is encountered, a window of about seven words around the target word is usually taken into consideration. To predict a word in a completely new sentence, the model looks up the frequency with which the potential target is surrounded by words present in the new sentence. To be useful, n-gram models need to be trained on massive amounts of text and therefore require extensive storage space. We will use a completely different approach to remembering information order — trajectory association described by Plate in [23]. Originally, this technique also used convolution and correlation, but this time items stored in a sequence are actually superimposed, rather than being bound together.

3 Trajectory Associacion

In the HRR model vectors are normalized and therefore can be regarded as radii of a sphere of radius 1. If we attach a sequence of items, say A, B, C, D, E to arrowheads of five of those vectors, we obtain a certain $trajectory$ associated with sequence $ABCDE$. This is a geometric analogue to the $method\ of\ loci$ which instructs to remember a list of items by associating each term with a distinctive location along a familiar path. Let k be a randomly chosen HRR vector and let

$$k^i = k \circledast k^{i-1} = k^{i-1} \circledast k, \quad i > 1$$

be its ith $power$, with $k^1 = k$. The sequence S_{ABCDE} is then stored as

$$S_{ABCDE} = A \circledast k + B \circledast k^2 + C \circledast k^3 + D \circledast k^4 + E \circledast k^5.$$

Of course, each power of k needs to be normalized before being bound with a sequence item. Otherwise, every subsequent power of k would be larger or smaller than its predecessor. As a result, every subsequent item stored in a sequence would have a bigger or a smaller share in vector S_{ABCDE}. Obviously, this method cannot be applied to the discrete GA model or to BSC, since it is impossible to obtain more than two distinct powers of a vector with the use of XOR as a means of binding.

This technique has a few obvious advantages present in HRR but not in GA_c had we wished to use ordinary vectors as first powers — different powers of a vector k would then be multivectors of different ranks. While k^i and $k^{i\pm1}$ are very similar in HRR, in GA_c they would not even share the same blades. Further, the similarity of k^i and k^{i+m} in HRR is the same as the similarity of k^j and k^{j+m}, whereas in GA_c that similarity would depend on the parity of i and j. In the light of these shortcomings, we need to use another structure acting as a first power in order to make trajectories work in GA_c. Let t be a random normalized full

multivector over $\mathbb{R}^n$ and let us define powers of t in the following way

$$\begin{aligned} t^1 &= t, \\ t^i &= (t^{i-1})t \text{ for } i > 1. \end{aligned}$$

We will store vectors $a_1 \dots a_l$ in a sequence $S_{a_1 \dots a_l}$ using powers of the multivector t

$$S_{a_1 \dots a_l} = a_1 t + a_2 t^2 + \dots + a_l t^l.$$

To answer a question "*What is the second item in a sequence?*" in GA_c we need to use the projected product

$$\langle S_{a_1 \dots a_l} (t^2)^+ \rangle_1 \approx a_2,$$

and to find out the place of item a_i we need to compute

$$(a_i)^+ S_{a_1 \dots a_l} \approx t^i.$$

Some may argue that such encoding puts a demand on items in the clean-up memory to hold information if they are roles or fillers, which is dangerously close to employing fixed data slots present in localist architectures. Actually, elements of a sequence can be recognized by their size, relatively shorter than the size of multivector t and its powers.

We present three experiments using trajectory association and we comment on test results for HRR and GA_c models. Firstly, we studied if an item can be retrieved given a sequence and an appropriate power of t, and vice versa — if a sequence and an item can lead to the power of t associated with that item. Finally, we tested whether both HRR and GA_c models can find the alignment of items in a sequence without the precise knowledge of vector t or its powers. Since the normalization using the square root of the number of chunks proved very noisy in statements containing powers of the trajectory vector, we decided to improve the HRR model. The HRR vectors in our tests were normalized by dividing them with their magnitude.

4 Correct Place Detection

In this experiment we investigated if powers of a (multi)vector t carry enough information about the original t. During 1000 tests for (multi)vector sizes ranging from 2^4 to 2^8 we asked the following question for every sequence S (a permutation of letters $\{A, B, C, D, E\}$): "*Where is A?*"

$$S \sharp A = \begin{cases} S \circledast A^* \approx t^x & \text{in HRR}, \\ A^+ S \approx t^x & \text{in } GA_c. \end{cases}$$

This amounted to 120000 questions altogether. For the purpose of the experiment described in Section 6 we used the clean-up memory consisting of powers of t only.

Ideally, every position of the letter A should come up as the correct answer the same number of times. However, since high powers of t acquire noise, lower powers should be recognized more often as the correct answer. Indeed, Figure 1 shows that in HRR the frequencies of the powers of t align with t being recognized most often and t^5 being recognized least often. In GA_c the percentage diagrams for various powers of t lay close to each other and often intertwine, still the relationship between the powers of t is similar to the one observed in HRR.

Since lower powers of t are recognized correctly more often, higher powers of t come up more often as the incorrect answer to $S \sharp A$. Vector t^3 is the correct answer to $S_{\bullet\bullet A \bullet\bullet} \sharp A$. However, if t^3 is not recognized, the next most similar answer will be t^5 because it contains three "copies" of t^3, indicated here by brackets

$$\Big\{ t * \Big(t * [t \Big\} * t \Big) * t].$$

The second most similar item will be t^4 because it contains two "copies" of t, and so on. The item least similar to t^3 will be t. This relationship should be best observable in HRR, since the powers can be multiplied from either side. In GA_c the powers of t can be increased from one side only and the relation between them should be less visible. Figure 2 shows that high powers of t are recognized more often in cases when the proper answer is not recognized — we will use this relationship in an experiment described in Section 6.

5 Correct Item Detection

Here we tested if trajectory association allows us to ask "*What is the xth item in a sequence?*"

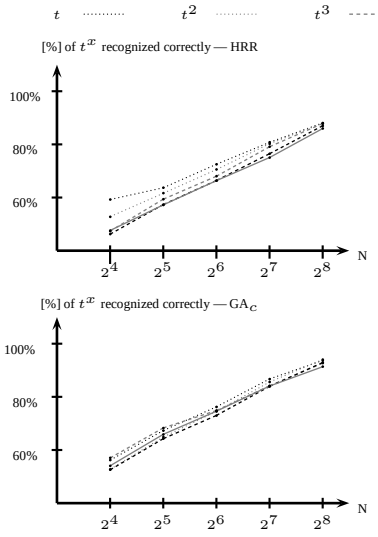
$$L_x \approx S \sharp t^x = \begin{cases} S \circledast (t^x)^* & \text{in HRR}, \\ S (t^x)^+ & \text{in } GA_c, \end{cases}$$

where $L_x \in \{A, B, C, D, E\}$ denotes the xth letter in a sequence. During 1000 tests for (multi)vector sizes ranging from 2^5 to 2^9 we asked that question for every permutation sequence of the set $\{A, B, C, D, E\}$, there were 120000 questions altogether for every (multi)vector t^x.

Again, we tested both HRR and GA_c models using a clean-up memory consisting only of expected answers, i.e. letters $\{A, B, C, D, E\}$. The results for both models (Figure 3) were similar with GA_c performing slightly better than HRR. In both models the first few letters of a sequence were more often recognized correctly than the last letters. Among the erroneously recognized letters, the last few letters of a sequence were most often offered as the most probable answer, which will come in handy in the next Section. The diagrams for GA_c lie closer together, once again indicating that trajectory association spreads information more evenly in GA_c than in HRR.

6 Item Alignment

The three previous tests were not very demanding for trajectory associations. Finally, we tested whether the HRR



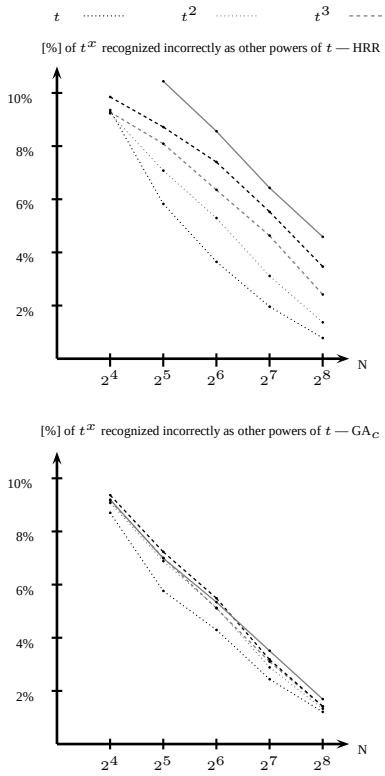
100% = $\frac{120000}{5}$ = 24000

	N=2^4	N=2^5	N=2^6	N=2^7	N=2^8
t	59.25%	63.67%	72.51%	80.77%	88.08%
t^2	52.80%	61.60%	70.52%	80.26%	87.93%
t^3	47.37%	59.37%	68.00%	79.11%	87.85%
t^4	46.22%	57.35%	66.40%	76.48%	87.01%
t^5	47.60%	57.23%	66.35%	75.05%	85.97%

100% = $\frac{120000}{5}$ = 24000

	$\mathbb{R}^4$	$\mathbb{R}^5$	$\mathbb{R}^6$	$\mathbb{R}^7$	$\mathbb{R}^8$
t	56.17%	67.31%	76.24%	86.73%	93.70%
t^2	52.72%	64.88%	74.53%	85.58%	94.05%
t^3	57.00%	68.30%	74.85%	84.13%	92.68%
t^4	52.53%	64.25%	73.00%	83.93%	92.90%
t^5	54.05%	65.89%	74.59%	84.06%	91.33%

Figure 1: Correct recognition of $S \sharp A \approx t^x$ in HRR and GA$_c$ using clean-up memory of $\{t, t^2, t^3, t^4, t^5\}$, 1000 trials.



100% = $\frac{120000}{5}$ = 24000

	N=2^4	N=2^5	N=2^6	N=2^7	N=2^8
t	9.35%	5.83%	3.65%	1.96%	0.78%
t^2	9.23%	7.08%	5.29%	3.11%	1.37%
t^3	9.28%	8.09%	6.35%	4.63%	2.42%
t^4	9.85%	8.72%	7.40%	5.53%	3.47%
t^5	11.63%	10.44%	8.56%	6.43%	4.59%

100% = $\frac{120000}{5}$ = 24000

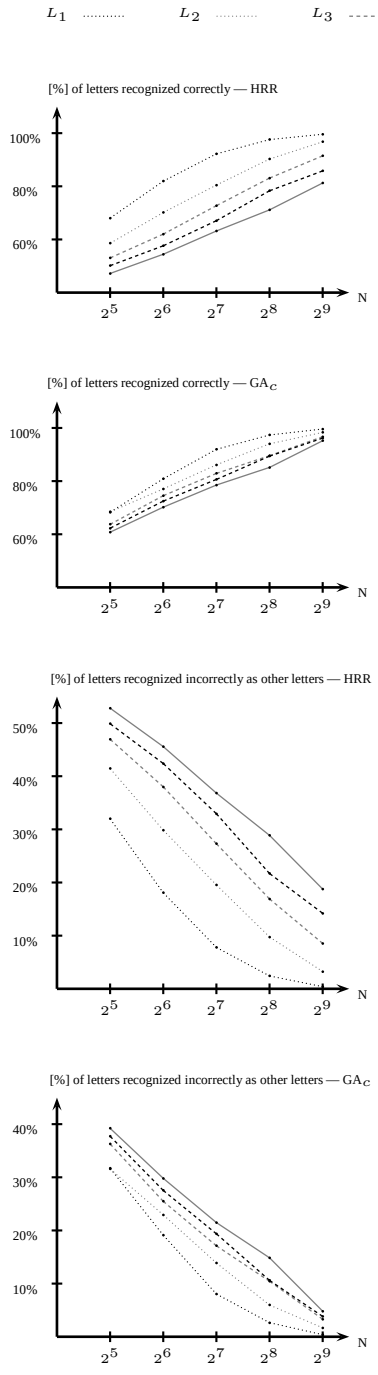
	$\mathbb{R}^4$	$\mathbb{R}^5$	$\mathbb{R}^6$	$\mathbb{R}^7$	$\mathbb{R}^8$
t	8.70%	5.76%	4.30%	2.44%	1.21%
t^2	9.08%	6.88%	5.13%	2.89%	1.33%
t^3	9.16%	7.01%	5.11%	3.10%	1.42%
t^4	9.37%	7.23%	5.49%	3.18%	1.42%
t^5	9.20%	6.99%	5.34%	3.51%	1.69%

Figure 2: Incorrect recognition of $S \sharp A \approx t^x$ in HRR and GA$_c$ using clean-up memory of $\{t, t^2, t^3, t^4, t^5\}$, 1000 trials.

and GA$_c$ models were capable of performing the following task:

> *Given only a set of letters A, B, C, D, E and an encoded sequence $S_{\bullet\bullet\bullet\bullet\bullet}$ comprised of those five letters find out the position of each letter in that sequence.*

We assumed that no direct access to t or its powers is given — they do belong to the clean-up memory, but cannot be retrieved “by name". One may think of this problem as a “black box" that inputs randomly chosen letter vectors and in return outputs a (multi)vector representing always the same sequence, irrespectively of the dimension of data. Inside, the black box generates (multi)vectors t, t^2, t^3, t^4, t^5. Their values are known to the observer but their names are not. Since we can distinguish letters from non-letters, the naive approach would be to try out all 120 alignments of letters A, B, C, D and E using all possible combinations of non-letters as the powers of t. Unfortunately, powers of



	N=2^5	N=2^6	N=2^7	N=2^8	N=2^9
L_1	68.03%	81.94%	92.21%	97.58%	99.59%
L_2	58.56%	70.15%	80.46%	90.25%	96.75%
L_3	53.05%	62.00%	72.71%	83.09%	91.46%
L_4	50.13%	57.61%	67.06%	78.31%	85.79%
L_5	47.20%	54.43%	63.18%	71.13%	81.24%

	$\mathbb{R}^5$	$\mathbb{R}^6$	$\mathbb{R}^7$	$\mathbb{R}^8$	$\mathbb{R}^9$
L_1	68.30%	80.88%	91.96%	97.36%	99.61%
L_2	68.43%	77.08%	86.12%	94.03%	98.33%
L_3	63.73%	74.52%	82.89%	89.56%	96.70%
L_4	62.28%	72.47%	80.63%	89.39%	96.18%
L_5	60.78%	70.19%	78.52%	85.15%	95.19%

	N=2^5	N=2^6	N=2^7	N=2^8	N=2^9
L_1	31.98%	18.07%	7.79%	2.42%	0.41%
L_2	41.44%	29.85%	19.54%	9.75%	3.25%
L_3	46.95%	38.00%	27.29%	16.91%	8.54%
L_4	49.87%	42.39%	32.94%	21.69%	14.21%
L_5	52.80%	45.57%	36.82%	28.87%	18.76%

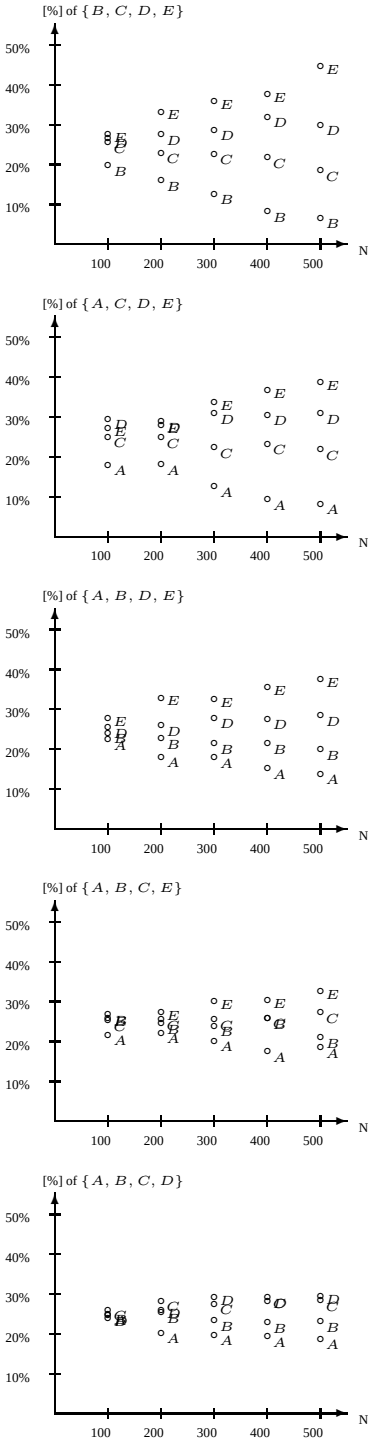
	$\mathbb{R}^5$	$\mathbb{R}^6$	$\mathbb{R}^7$	$\mathbb{R}^8$	$\mathbb{R}^9$
L_1	31.70%	19.12%	8.04%	2.65%	0.39%
L_2	31.57%	22.93%	13.88%	5.97%	1.68%
L_3	36.27%	25.48%	17.11%	10.44%	3.30%
L_4	37.72%	27.53%	19.37%	10.61%	3.83%
L_5	39.22%	29.81%	21.48%	14.85%	4.81%

Figure 3: Recognition of $S_{L_1L_2L_3L_4L_5} \sharp\, t^x \approx L_x$ in HRR and GA$_c$ using clean-up memory containing letters only, 1000 trials.

t are different each time the black box produces a sequence. We will use an algorithm based on the assumption that t^x, if not recognized correctly, is more similar to highest powers of t as shown in Section 4. The second assumption is that letters lying closer to the end of the sequence are often offered as the incorrect answer to questions concerning letters (Section 5). The clean-up memory $\mathcal{C}$ for this experiment consists of all five letters and the five powers of t. We will also use an auxiliary clean-up memory $\mathcal{L}$ containing letters only.

The algorithm for finding out the position of each letter begins with asking a question described by equation (1) — "Where in the sequence $S_{\bullet\bullet\bullet\bullet\bullet}$ is the letter L_x?":

$$S_{\bullet\bullet\bullet\bullet\bullet} \sharp\, L_x = \left\{ \begin{array}{ll} S_{\bullet\bullet\bullet\bullet\bullet} \circledast (L_x)^* & \text{in HRR} \\ (L_x)^+ S_{\bullet\bullet\bullet\bullet\bullet} & \text{in GA}_c \end{array} \right\} = (t^x)' \approx t^x \quad (1)$$



Asking with A

	N=100	N=200	N=300	N=400	N=500	
non A	1643	960	616	506	393	= 100%
B	19.90%	16.04%	12.66%	8.30%	6.62%	
C	25.56%	22.92%	22.73%	21.94%	18.58%	
D	26.78%	27.71%	28.73%	32.02%	30.03%	
E	27.75%	33.33%	35.88%	37.75%	44.78%	

Asking with B

	N=100*	N=200*	N=300	N=400	N=500	
non B	2474	1588	1133	837	628	= 100%
A	18.03%	18.14%	12.71%	9.44%	8.28%	
C	25.14%	24.87%	22.51%	23.30%	21.97%	
D	29.59%	28.90%	30.98%	30.59%	30.89%	
E	27.24%	28.09%	33.80%	36.68%	38.85%	

Asking with C

	N=100	N=200	N=300	N=400	N=500	
non C	3003	2214	1694	1346	1121	= 100%
A	22.51%	18.16%	18.00%	15.30%	13.83%	
B	24.18%	22.81%	21.55%	21.47%	20.07%	
D	25.51%	26.06%	27.86%	27.64%	28.55%	
E	27.81%	32.97%	32.59%	35.59%	37.56%	

Asking with D

	N=100*	N=200	N=300	N=400	N=500	
non D	3432	2686	2264	1808	1564	= 100%
A	21.71%	22.23%	20.19%	17.64%	18.67%	
B	26.84%	24.72%	24.03%	25.77%	21.16%	
C	25.44%	25.58%	25.66%	26.05%	27.37%	
E	26.02%	27.48%	30.12%	30.53%	32.80%	

Asking with E

	N=100*	N=200*	N=300	N=400*	N=500	
non E	3754	3116	2617	2353	1921	= 100%
A	24.77%	20.31%	19.79%	19.46%	18.84%	
B	25.12%	25.48%	23.42%	22.90%	23.27%	
C	26.08%	28.34%	27.59%	29.20%	28.37%	
D	24.03%	25.87%	29.19%	28.35%	29.52%	

Figure 4: Finding letter alignment in a sequence S_{ABCDE} in HRR, 10000 trials.

for each letter $L_x \in \mathcal{L}$. Next, we need to find the item in the clean-up memory $\mathcal{C} \setminus \mathcal{L}$ that is most similar to $(t^x)'$. Let us denote this item by z. With high probability, z is the power of t associated with the position of the letter L_x in the sequence $S_{\bullet\bullet\bullet\bullet\bullet}$, although, if recognized incorrectly, z will most likely point to some other $t^{y>x}$. Now let us ask a second question (eq. (2)) — "Which letter is situated at the z'th position in the sequence $S_{\bullet\bullet\bullet\bullet\bullet}$?":

$$S_{\bullet\bullet\bullet\bullet\bullet} \sharp z = \left\{ \begin{array}{ll} S_{\bullet\bullet\bullet\bullet\bullet} \circledast z^* & \text{in HRR} \\ \langle S_{\bullet\bullet\bullet\bullet\bullet}\, z^+ \rangle_1 & \text{in GA}_c \end{array} \right\} = L' \approx L_x. \quad (2)$$

We use the projected product in GA_c because we are looking for a letter vector placed on the position indicated by

z. In HRR the resulting L' should be compared with letters only. In most cases L' will point to the correct letter. However, in a small fraction of test results, L' will point to letters surrounding L_x, because z has been mistakenly decoded as t^y for some $y \neq x$. Also, letters preceding L_x should come up less often than letters proceeding L_x.

Figure 4 presents test results for HRR. The data in Figure 4 should be interpreted as follows: the first row of each table next to a graph contains the vector lengths of the data used in 5 consecutive experiments (10000 trials each). The second row contains the number of faulty answers within those 10000 trials. The next 4 rows present the percentage of occurence of a "faulty" letter within all faulty answers presented in the second row.

Faulty alignments (i.e. those, for which the percentages corressponding to letters do no align increasingly within a single column) have been marked with a "*" in the table headings. We used S_{ABCDE} as the mysterious encoded sequence $S_{\bullet\bullet\bullet\bullet\bullet}$. In each case we crossed out the most frequently occurring letter and we concentrated on the frequency of the remaining letters. In HRR, for sufficiently large vector sizes, the frequencies f_L of all letters $L \in \mathcal{L}$ aligned correctly

$$\begin{aligned} f_B < f_C < f_D < f_E \quad & \text{asking with } A, \\ f_A < f_C < f_D < f_E \quad & \text{asking with } B, \\ f_A < f_B < f_D < f_E \quad & \text{asking with } C, \\ f_A < f_B < f_C < f_E \quad & \text{asking with } D, \\ f_A < f_B < f_C < f_D \quad & \text{asking with } E. \end{aligned}$$

It was straightforward that these inequalities lead to $f_A < f_B < f_C < f_D < f_E$ and correctly identify the encoded sequence as S_{ABCDE}. Test results are less accurate when we asked about letters lying closer to the end of a sequence, therefore the size of the vector should be adequately long. Moreover, the longer the vector, the larger the difference between the frequencies.

GA_c was expected to perform worse in this experiment, because we can construct powers of a mulivector t^{i-1} by multiplying it with t from one side only. Another reason for poor performance was that the frequencies of the powers of t recognized incorrectly tend to cluster in GA_c (Figure 2). Indeed, Table 1 shows that letter frequencies do not align correctly at all. We therefore needed to slightly modify the algorithm for finding letter alignment in GA_c. Since powers of t are more similar to each other in GA_c than in HRR (Section 5), we concentrated on two largest frequencies in each series of asking questions — the largest frequency represents the letter L that was used to ask the question and the second largest frequency indicates letter $\hat{L}$ that most likely proceeds letter L.

Table 1 presents the frequencies of letters recognized as the most probable answer to Equation (2), the second largest frequency in each row is printed in bold. Partial letter alignments have been placed next to each table and contradictory alignments have been preceded with a "*". When being asked with the last letter of the sequence, HRR provided less accurate answers and so did GA_c by yielding more contradictions than in case of previous letters. It is impossible to avoid contradictory alignments in GA_c because we do not know which letter is the last one and the algorithm for recovering letter alignment in GA_c instructs us to write down the partial alignment with that letter being proceeded by another letter. The remaining alignments point correctly to the sequence S_{ABCDE}

$$\left.\begin{array}{c} A \prec B \\ C \prec D \prec E \\ A \prec B \prec C \prec D \\ A \prec C \\ B \prec D \prec E \\ A \prec E \end{array}\right\} \Rightarrow A \prec B \prec C \prec D \prec E.$$

7 Conclusion

We have shown that multivector powers in GA_c have properties similar to convolutive powers of HRR vectors

- (multi)vectors t^{i-r} and t^i are similar in much the same way as t^i and t^{i+r},
- items placed near the beginning of a sequence are remembered more prominently and thus, are recognized correctly more often,
- items placed near the end of a sequence are remembered less precisely and often come up as the most probable answer when the correct item is not recognized.

We have used the last two properties to find the alignment of sequence items without the explicit knowledge of (multi)vector powers. While HRR retrieved the original alignment without greater problems, GA_c left us with an easily soluble logical puzzle providing fragmentary alignments.

These properties can be used to build holographic lexicons, dictionaries and other structures that require storing order information and word meaning in the same pattern.

Acknowledgement

This work was supported by *Grant G.0405.08* of the *Fund for Scientific Research Flanders*.

References

[1] D. Aerts and M. Czachor (2004), "Quantum aspects of semantic analysis and symbolic artificial intelligence", *J. Phys. A*, vol. 37, pp. L123-L13.

[2] D. Aerts and M. Czachor (2007), "Cartoon computation: Quantum-like algorithms without quantum mechanics", *J. Phys. A*, vol. 40, pp. F259-F266.

Table 1: Finding letter alignment in a sequence S_{ABCDE} in GA_c, 10000 trials.

$\mathbb{R}^5$	f_A	f_B	f_C	f_D	f_E
asking with A	71.60%	**8.22**%	6.44%	2.72%	6.47%
asking with B	8.98%	65.92%	9.16%	6.76%	**9.18**%
asking with C	7.28%	9.52%	67.25%	**9.67**%	6.28%
asking with D	8.74%	6.75%	8.80%	66.83%	**8.88**%
asking with E	8.03%	**9.96**%	6.83%	9.28%	65.90%

$A \prec B$
$^*B \prec E$
$C \prec D$
$D \prec E$
$^*E \prec B$

$A \prec B$
$C \prec D \prec E$

$\mathbb{R}^6$	f_A	f_B	f_C	f_D	f_E
asking with A	80.34%	**5.34**%	4.61%	5.08%	4.63%
asking with B	6.56%	73.91%	**7.48**%	4.67%	7.38%
asking with C	6.71%	7.47%	72.83%	**7.68**%	5.31%
asking with D	6.79%	5.37%	7.42%	72.30%	**8.12**%
asking with E	5.52%	7.77%	5.58%	**8.54**%	72.59%

$A \prec B$
$B \prec C$
$C \prec D$
$^*D \prec E$
$^*E \prec D$

$A \prec B \prec C \prec D$

$\mathbb{R}^7$	f_A	f_B	f_C	f_D	f_E
asking with A	89.78%	2.42%	**2.91**%	2.37%	2.52%
asking with B	3.78%	83.92%	4.04%	**4.44**%	3.82%
asking with C	4.30%	4.79%	80.54%	5.11%	**5.26**%
asking with D	4.35%	5.07%	5.41%	79.09%	**6.08**%
asking with E	4.57%	5.07%	**5.85**%	5.68%	78.83%

$A \prec C$
$B \prec D$
$^*C \prec E$
$D \prec E$
$^*E \prec C$

$A \prec C$
$B \prec D \prec E$

$\mathbb{R}^8$	f_A	f_B	f_C	f_D	f_E
asking with A	95.33%	0.88%	1.26%	1.20%	**1.30**%
asking with B	1.34%	92.27%	1.72%	**2.93**%	1.74%
asking with C	2.15%	2.28%	88.99%	2.35%	**4.23**%
asking with D	1.93%	**3.75**%	2.82%	88.19%	3.31%
asking with E	2.60%	2.36%	**5.09**%	3.32%	86.63%

$A \prec E$
$^*B \prec D$
$^*C \prec E$
$^*D \prec B$
$^*E \prec C$

$A \prec E$

[3] M. Czachor (2007), “Elementary gates for cartoon computation", *J. Phys. A*, vol. 40, pp. F753-F759.

[4] D. Aerts and M. Czachor (2008), “Tensor-product versus geometric-product coding", *Physical Review A*, vol. 77, id. 012316.

[5] D. Aerts, M. Czachor, and B. De Moor (2009), “Geometric Analogue of Holographic Reduced Representation", *J. Math. Psychology*, vol. 53, pp. 389-398.

[6] D. Aerts, M. Czachor, and B. De Moor (2006), “On geometric-algebra representation of binary spatter codes". preprint arXiv:cs/0610075 [cs.AI].

[7] D. Aerts, M. Czachor, and Ł. Orłowski (2009), “Teleportation of geometric structures in 3D ", *J. Phys. A* vol. 42, 135307.

[8] W.K. Clifford (1878), “Applications of Grassmann’s extensive algebra", *American Journal of Mathematics Pure and Applied*, vol. 1, 350–358.

[9] R. W. Gayler (1998), “Multiplicative binding, representation operators, and analogy", *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, K. Holoyak, D. Gentner, and B. Kokinov, eds., Sofia, Bulgaria: New Bulgarian University, p. 405.

[10] S. Deerwester et al. (1990), “Indexing by Latent Semantic Analysis", *Journal of American Society for Information Science*, vol. 41, 391.

[11] H. Grassmann (1877), “Der Ort der Hamilton’schen Quaternionen in der Ausdehnungslehre”, *Mathematische Annalen*, vol. 3, 375–386.

[12] M.N. Jones & D.J.K. Mewhort (2007), “Representing Word Meaning and Order Information in a Composite Holographic Lexicon", *Psychological Review*, vol. 114, No. 1, pp. 1-37.

[13] G. E. Hinton, J. L. McClelland and D. E. Rumelhart (1986), “"Parallel distributed processing: Explorations in the microstructure of cognition", vol. 1, 77Ű109, *"Distributed representations"*, The MIT Press, Cambridge, MA.

[14] P. Kanerva (1996), “Binary spatter codes of ordered k-tuples”. In C. von der Malsburg et al. (Eds.), *Artificial Neural Networks ICANN Proceedings, Lecture Notes in Computer Science* vol. 1112, pp. 869-873.

[15] P. Kanerva (1997), “Fully distributed representation”. *Proc. 1997 Real World Computing Symposium (RWCŠ97, Tokyo)*, pp. 358-365.

[16] E. M. Kussul (1992), *Associative Neuron-Like Structures*. Kiev: Naukova Dumka (in Russian).

[17] E.M. Kussul and T.N. Baidyk (1990), "Design of Neural-Like Network Architecture for Recognition of Object Shapes in Images", *Soviet J. Automation and Information Sciences*, vol. 23, no. 5, pp. 53-58.

[18] N.G. Marchuk, and D.S. Shirokov (2008), "Unitary spaces on Clifford algebras", *Advances in Applied Clifford Algebras*, vol 18, pp. 237-254.

[19] A. Patyk (2010), "Geometric Algebra Model of Distributed Representations", in *Geometric Algebra Computing in Engineering and Computer Science*, E. Bayro-Corrochano and G. Scheuermann, eds. Berlin: Springer. Preprint arXiv:1003.5899v1 [cs.AI].

[20] A. Patyk-Łońska (2011), "Distributed Representations Based on Geometric Algebra: the Continuous Model", submitted to *Informatica*.

[21] R. Pike (1984), "Comparison of Convolution and Matrix Distributed Memory Systems for Associative Recall and Recognition", *Psychological Review*, vol. 91, No. 3, pp. 281-294.

[22] T. Plate (1995), "Holographic Reduced Representations", *IEEE Trans. Neural Networks*, vol. 6, no. 3, pp. 623-641.

[23] T. Plate (2003), *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. CSLI Publications, Stanford.

[24] D.A. Rachkovskij (2001), "Representation and Processing of Structures with Binary Sparse Distributed Codes", *IEEE Trans. Knowledge Data Engineering*, vol. 13, no. 2, pp. 261-276.

[25] P. Smolensky (1990), "Tensor product variable binding and the representation of symbolic structures in connectionist systems". *Artificial Intelligence*, vol. 46, pp. 159-216.

Grammar Checking with Dependency Parsing: A Possible Extension for LanguageTool

Maxim Mozgovoy
University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, 965-8580 Japan
E-mail: mozgovoy@u-aizu.ac.jp

Keywords: grammar checking, dependency parsing, LanguageTool

Received: October 25, 2011

This paper describes a possible extension for a well-known open source grammar checking software LanguageTool. The proposed extension allows the developers to write grammatical rules that rely on natural language parser-supplied dependency trees. Such rules are indispensable for the analysis of word-word links in order to handle a variety of grammatical errors, including improper use of articles, incorrect verb government, and wrong word form agreement.

Povzetek: Članek opisuje razširitev programa LanguageTool s pravili, ki uporabljajo odvisnostna drevesa.

1 Introduction

Grammar checking is a well-recognized problem of natural language processing. Grammar checkers are helpful in a variety of scenarios, such as text authoring and language learning. The purpose of such tools is to find grammatical errors in the input text: incorrect use of person, number, case or gender, improper verb government, wrong word order, and so on. A grammar checker normally works in combination with a spellchecker — a module that detects spelling errors in individual words. As a rule, spell checker cannot correct even basic grammatical flaws, such as erroneous choice of article (like in the expression “an box”).

While a spellchecker is already an essential part of a modern text authoring system, a grammar checking module is still found only in large commercial packages like Microsoft Office or WordPerfect Office. Certain grammar checkers are also available as additional software packages or online services, offered by independent companies [1-3].

This situation is slowly changing nowadays. With the growing popularity of open source software, more natural language processing systems should become available for wider use. Open spellchecking libraries, such as JOrtho and GNU Aspell already exist, and anyone can extend own software with their capabilities. Grammar checking is a more challenging task, and most open projects are still far beyond well-established proofing tools, such as offered by MS Word.

*Supported in part by the Fukushima Prefectural Foundation, Project F-23-1, FY2011.

This paper is based on M. Mozgovoy, *Dependency-Based Rules for Grammar Checking with LanguageTool* published in the proceedings of the 1st International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS’2011 conference).

1.1 Rule-Based Grammar Checking

Probably, the predominating approach to grammar checking today consists in testing the input text against a set of handcrafted rules [4, 5]. For example, the rule

```
I + Verb (3rd person, singular form)
```

corresponds to the incorrect verb form use, as in the phrase “I has a dog”. In order to emphasize the nature of such rules as erroneous patterns, they are often called “mal-rules”.

This method has several attractive features: (a) rules can be easily added, modified or removed; (b) every rule can have a corresponding extensive explanation, helpful for the end user; (c) the system is easily debuggable, since its decisions can be traced to a particular rule; (d) the rules can be authored by the linguists, possessing limited or no programming skills. An obvious disadvantage of a rule-based system is a large amount of manual work, needed to build an extensive rule set.

An alternative approach is represented with several varieties of statistical systems that analyze existing collections of grammatically correct and incorrect texts, attempting to find word patterns and/or text features that correspond to correct sentences [6, 7]. The simplest statistical grammar algorithm consists in analyzing N-grams — chains of N consecutive words [8]. If a certain word chain is common in the master text corpus, it is considered correct.

Statistical grammar checkers have their own advantages and drawbacks, but their analysis is beyond the scope of this article.

1.2 Introducing LanguageTool

The purpose of the present work is to design a possible extension for the LanguageTool grammar checker [9]. LanguageTool is a modern rule-based open source

grammar checking system, available both as a plug-in for OpenOffice.org and as a downloadable library, which makes it ready for use in any software projects. Currently LanguageTool supports 21 languages, though the number of ready grammatical rules ranges from 4 for Lithuanian to 1994 for French (as of November, 2011). The rules can be authored by any interested contributors.

Unfortunately, the syntax of rules in LanguageTool does not allow formulating certain grammatical phenomena. In the subsequent sections, we will consider these limitations and a possible method to reduce them.

2 Basic Design Principles of LanguageTool

LanguageTool defines an XML-based language for describing mal-rules. In its simplest form, a mal-rule is just a sequence of tokens to be matched in the text:

```
<!-- "all be it" instead of "albeit" -->

<pattern>
  <token>all</token>
  <token>be</token>
  <token>it</token>
</pattern>
<message>Did you mean 'albeit'?</message>
```

The syntax of the rules is flexible and powerful: it is possible to use OR and NOT logic operations (“match token A or token B”; “match any token except C”), to skip optional tokens, and, to some extent, to use regular expressions.

Several syntactic elements are backed with additional linguistic modules — *sentence splitter* and *part-of*-speech *tagger*. Sentence splitter determines the boundaries of each sentence, thus allowing the user to find certain tokens exactly at the beginning or at the end of a sentence:

```
<!-- "another words," instead of
     "in other words,"
     at the beginning of a sentence -->

<pattern>
  <token postag="SENT_START"></token>
  <token>another</token>
  <token>words</token>
  <token>,</token>
</pattern>
<message>Did you mean
         'in other words'?</message>
```

Part-of-speech tagger determines every word’s part of speech, helping the user to find tokens that belong to a certain class:

```
<!-- "ca" + [personal pronoun] instead of
     "can" + [personal pronoun] -->

<pattern>
  <token>ca</token>
  <token postag="PRP"></token>
</pattern>
<message>Did you mean 'can'?</message>
```

LanguageTool makes use of third-party libraries for splitting and tagging the input text. Fortunately, a number of ready solutions are available for this purpose (e.g., Ratnaparkhi’s MXTERMINATOR and MXPOST [10, 11]).

3 Introducing Dependency-Based Rules

Despite high expressive power and flexibility, LanguageTool’s rule system has a notable shortcoming: it treats the input text as a sequence of tokens, ignoring tree-like nature of natural language sentences.

Consider, for example, the following problem. In English, `a/an` article should never be used with a noun in a plural form. The current LanguageTool rule to detect such a case is defined as follows:

```
<!—"a/an" article, then a plural noun -->

<pattern>
  <token regexp="yes">a|an</token>
  <token postag="NNS|NNPS"</token>
</pattern>
<message>Don't use indefinite articles
         with plural words.</message>
```

However, this rule ignores the fact that there can be any number of words between `a/an` and the corresponding noun (“a box”, “a wooden box”, “a simple wooden box”). The rule definition can be improved if we allow any number of optional adjectives between the article and the noun, but in general case this solution is inadequate.

In order to handle such problems, the grammar checker should analyze nonlinear structure of the phrase. An article is logically linked with a noun, regardless of any words between them. This nonlinear structure can be obtained with an additional module, known as *dependency parser*. This instrument represents the structure of every sentence with a *parse tree*, having words as nodes and logical links between them as edges (see Fig. 1).

As it can be seen, the article “a” is linked directly to the noun “box”. Having such a tree, it is possible to extend the syntax of LanguageTool grammatical rules, enabling the developers to analyze word-word relationships.

4 Technical Approach

In order to achieve our goals, we had to solve three subproblems: 1) select a suitable dependency parsing instrument; 2) develop an appropriate syntax for dependency-based rules; 3) design the corresponding rule-matching algorithm.

4.1 Selecting a Practical Dependency Parser

After examining currently available solutions, we decided to use one of two parsers: MaltParser [12] or

LDPar [13]. Both of them are high-quality dependency parsers, available as open source.

MaltParser is written in Java, and thus suits better for the use in combination with the current implementation of LanguageTool, also made with Java. LDPar distribution contains cross-platform C++ code, providing compilable efficient implementation. Both parsers are based on machine learning: the parser first has to be trained with a collection of correctly parsed sentences (a *treebank*). MaltParser and LDPar also share the same format of input and output data.

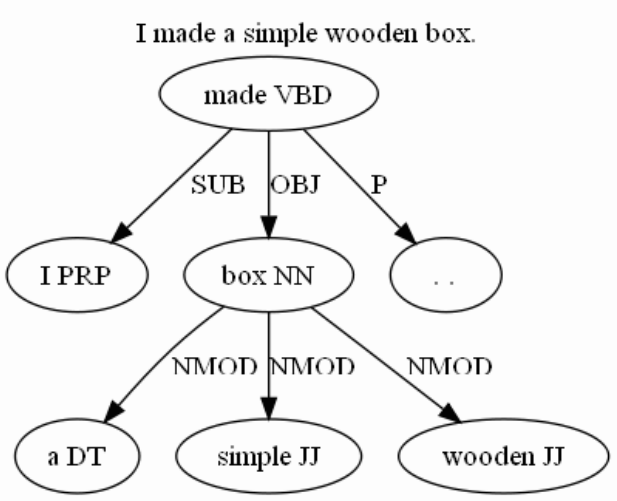


Figure 1: Parse tree for the phrase "I made a simple wooden box."

4.2 Syntax for Dependency-Based Rules

Dependency-based rules should provide syntactic means for the following basic functions:

1) Match a link between two given words, optionally labeled with a given label. This function should be generalizable to the matching of the whole subtree.
2) Check whether a certain word appears before or after another word, in order to control word precedence.
3) Check for the absence of the given subtree in the parse tree.

In order to satisfy these requirements, we suggest the following syntax for an individual dependency-based rule. The rule definition is split into chunks, each representing a separate subtree to be matched:

```
CHUNK1
CHUNK2
...
CHUNKN
```

Every `CHUNKi` is represented with a sequence of tokens, defined with `token` XML tag:

```
<token [attributes]>token-value</token>
```

Currently our system supports the following attributes:

- **pos="text"**: the token should belong to the specified part-of-speech class;
- **label="text"**: the link to the token's parent (according to the parse tree) should has the specified label;
- **parent="number"**: the token should have the specified token as a parent (according to the parse tree);
- **except**: the token's value should not match `token-value`;
- **before="number"**: the token should appear in the sentence before the specified token;
- **after="number"**: the token should appear in the sentence after the specified token;
- **chunk_start**: start-of-chunk marker;
- **inverse**: the current chunk (subtree) should not be found in the parse tree;
- **set_anchor="text"**: the token will be marked with a symbolic "anchor" (see below);
- **anchor="text"**: the token should be found at the specified anchor position in the parse tree.

Attributes **parent**, **before**, and **after** expect a token's cardinal number within the current chunk as an argument. By default, every chunk of the rule has to be matched in the parse tree in order to satisfy the rule.

Anchors are introduced to simplify the analysis of the subtrees. For example, suppose that the first chunk of a rule matches an object of the sentence's root verb:

```
<token label="ROOT"></token>
<token parent="1" label="OBJ"
       set_anchor="anchor1"></token>
```

Now, suppose that a certain chunk later in the chain needs to check that this object has dependent adjectives. By using the anchor, it is possible to start matching directly from the right place:

```
<token anchor="anchor1"></token>
<token parent="1" pos="ADJ"></token>
```

4.3 Examples

The following examples illustrate the capabilities of dependency-based rules:

```
<!-- Example 1:
     in non-interrogative sentences
     the subject should be placed before
     the predicate -->

<token pos="VB|VBP|VBZ|VBD"
       label="ROOT"></token>
<token after="1" label="SUB"></token>

<token chunk_start="" inverse=""
       label="ROOT"></token>
<token parent="1">?</token>
```

The first chunk ensures that the system has found a subject (labeled SUB), placed after the main verb. The second chunk asserts the absence of '?' mark, linked to the tree root.

```
<!-- Example 2:
     "a/an" should not be used
     with plural nouns -->

<token>a|an</token>
<token pos="NNS|NNPS" parent="1"></token>
```

This mal-rule finds a/an articles, linked to plural nouns (marked as NNS or NNPS by a part-of-speech tagger). Note that the determiner (such as an article) is always directly connected to the corresponding word, even if they are not adjacent in the original sentence.

```
<!-- Example 3:
     the gerund should be used in
     conjunction with auxiliary verbs -->

<token pos="VBG" label="ROOT"></token>
```

If a gerund (verb ing-form) is recognized as a parse tree root, this means the absence of an obligatory auxiliary verb (such as "is", "was"). If an auxiliary verb is present, it becomes a root element of the tree.

```
<!-- Example 4:
     improper personal verb form used -->

<token pos="VBZ"></token>
<token parent="1"
       label="SUB">I|we|you|they</token>
```

If the subject of a certain verb is I/we/you/they, the verb should not be in the 3rd person singular form.

4.4 Implementation

Each chunk of a rule is matched separately. The chunk-matching algorithm is a straightforward implementation of the depth-first search routine:

```
// token_index : integer
// used_tokens: set of integers
bool matchSubtree(token_index,
                  used_tokens)
{
  if(token_index > MAX_INDEX_IN_CHUNK)
    return true;

  for_each(token k in the tree)
    if(used_tokens.contains(k) == false)
      if(tokens_match(token_index, k)
        if(matchSubtree(token_index + 1,
                  union(used_tokens, k))
          return true;

  return false;
}

// first call:
// b = matchSubtree(0, empty());
```

5 RuleDesigner Tool

In order to simplify rule authoring, we have also created a specialized RuleDesigner tool that provides a centralized interface for the development, debugging and testing of grammatical rules.

Our approach to the rule creation process can be compared with test-driven development. Each rule has a special "self-tests" section that contains an arbitrary text fragment. Self-test is passed if the system finds the given number of errors in this text, using the current rule:

```
<test matches="3">I loves London.
We eats in London.
John and I loves London.</test>
```

RuleDesigner automatically runs self-tests and displays all the information needed to check and correct grammatical rules:

- editable rule definition;
- editable list of self-tests;
- the results of self-tests.

Furthermore, for each sentence in self-tests RuleDesigner shows its parse tree[1], part-of-speech markup, explanation to the user, and a list of possible text corrections[2] (see Figure 2).

It is also possible to run tests on a user-specified text block with all the rules turned on. It helps to identify false positives, not revealed with isolated rule-level self-tests.

6 Discussion

LanguageTool is a good example of an extensible rule-based grammar checker. Basic grammatical rules can be expressed by means of standard regular expressions. If their expressive power is insufficient to describe a certain rule, one can make use of additional natural language processing-powered syntactic elements, backed with sentence splitter and part-of-speech tagger.

This architecture can be extended further by incorporating other language processing modules. An obvious candidate for this role is a natural language parser that shows immediate word-word relationships. We have demonstrated several examples of grammatical errors, detectable with parser-powered mal-rules.

Since we consider rule-based grammar checking to be an established technology, the discussion of its advantages and drawbacks is beyond the scope of our work. However, our experiments have revealed weak points of the language tools we use (parser and part-of-speech tagger, mainly).

Normally, these tools, being based on machine learning algorithms, need initial training on annotated text data. Most such training collections are represented

[1] Obtained with AT&T GraphViz tool.

[2] Discussion of text-correction functionality is outside the scope of this paper.

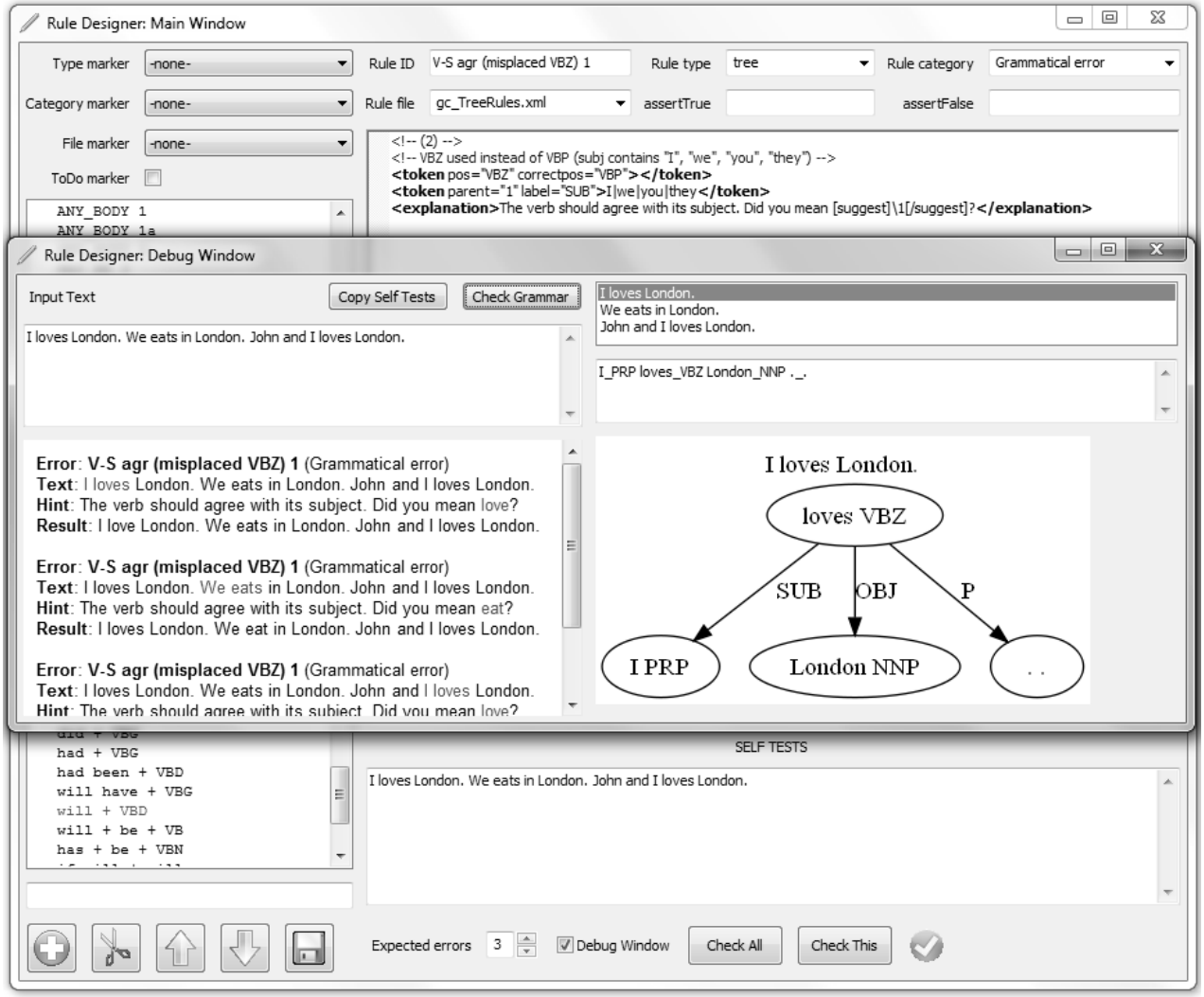


Figure 2: RuleDesigner.

with grammatically correct sentences. Thus, ungrammatical phrases may contain previously unseen patterns, causing incorrect results. For example, a part-of-speech tagger cannot reliably determine a tag for the word "like" in the phrase "he like dogs", since such a pattern never appears in the training collection.

Since processing ungrammatical sentences is a crucial feature for a grammar checking module, this issue needs further research. One of the possible solutions would be to extend the training collection with ungrammatical sentences. Our preliminary experiments have indeed shown that the inclusion of ungrammatical sentences into the training collection increases the quality of part-of-speech tagging.

7 Conclusion

We have designed and implemented the mechanism of natural language parser-backed rules for a LanguageTool-based grammar checking module. Our syntax allows designing the rules that analyze word-word dependencies in a given phrase. We have shown real examples of language phenomena, where such rules are much more helpful than built-in LanguageTool instruments.

Dependency-based rules are typically more complicated than the rules, based on regular expressions. Therefore, we developed RuleDesigner — a visual tool for rule authoring. It shows the user how language tools (sentence splitter, part-of-speech tagger, and dependency parser) process the input text, thus assisting debugging. We also included a system of self-tests, useful to keep the grammar checker consistent during the process of development.

References

[1] J. Burston (2008). Bon Patron: An Online Spelling, Grammar, and Expression Checker. *CALICO Journal*, vol. 25(2), pp. 337-347.

[2] H.J. Chen (2009). Evaluating Two Web-based Grammar Checkers-Microsoft ESL Assistant and NTNU Statistical Grammar Checker. *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 14(2), pp. 161-180.

[3] B. O'Regan, A. Mompean and P. Desmet (2010). From Spell, Grammar and Style Checkers to Writing Aids for English and French as a Foreign Language: Challenges and Opportunities. *Revue francaise de linguistique appliquee*, vol. 15(2), pp. 67-84.

[4] E.M. Bender, D. Flickinger, S. Oepen, A. Walsh, and T. Baldwin (2004). Arboretum: Using a precision grammar for grammar checking in CALL. *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy, pp. 83-86.

[5] M. Milkowski (2010). Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, vol. 40(7), pp. 543-566.

[6] M.J. Alam, N. UzZaman and M. Khan (2006). N-gram based statistical grammar checker for Bangla and English. *Proceedings of ninth International Conference on Computer and Information Technology (ICCIT 2006)*, Dhaka, Bangladesh.

[7] J. Wagner, J. Foster and J. van Genabith (2006). Detecting grammatical errors using probabilistic parsing. *Workshop on Interfaces of Intelligent Computer-Assisted Language Learning*. Columbus, Ohio, USA.

[8] J. Sjobergh (2006). The Internet as a Normative Corpus: Grammar Checking with a Search Engine. *Technical Report,* KTH Nada, Sweden.

[9] D. Naber (2003). A rule-based style and grammar checker. *Master's thesis,* University of Bielefeld, Germany.

[10] A. Ratnaparkhi (1996). A maximum entropy model for part-of-speech tagging. *Proceedings of the conference on empirical methods in natural language processing*, Philadelphia, Pennsylvania, USA, pp. 133-142.

[11] J.C. Reynar and A. Ratnaparkhi (1997). A maximum entropy approach to identifying sentence boundaries. *Proceedings of the fifth conference on Applied natural language processing*, Washington D.C., USA, pp. 16-19.

[12] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, vol. 13(2), pp. 95-135.

[13] P. Jian and C. Zong (2009). Layer-Based Dependency Parsing. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*, Hong Kong, China, pp. 230-239.

Learning Predictive Qualitative Models with Padé

Jure Žabkar, Martin Možina, Ivan Bratko and Janez Demšar
University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, Ljubljana
E-mail: jure.zabkar|martin.mozinalivan.bratkoljanez.demsar}@fri.uni-lj.si

Keywords: qualitative modelling, machine learning

Received: April 18, 2010

Qualitative models are similar to regression models, except that instead of numerical predictions they provide insight into how a change of a certain input variable affects the output within a context of other inputs. Although people usually reason qualitatively, machine learning has mostly ignored this type of model. We present a new approach to learning qualitative models from numerical data. We describe Padé, a suite of methods for estimating partial derivatives of unknown sampled target functions. We show how to build qualitative models using standard machine learning algorithms by replacing the output variable with signs of computed derivatives. Experiments show that the developed methods are quite accurate, scalable to high number of dimensions and robust with regard to noise.

Povzetek: Predstavljena je nova metoda za učenje iz kvalitativnih podatkov, imenovana Padé. Temelji na ocenjevanju parcialnih odvodov neznane vzorčene ciljne funkcije.

1 Introduction

Qualitative models describe quantitative relations in qualitative terms, for instance, *the more it rains and the longer I stay in the rain, the wetter I will get (unless I have an umbrella)*. Although seemingly inferior to the more accurate numerical models, there are many reasons why qualitative models are interesting for artificial intelligence.

One of the goals of artificial intelligence, according to one of its founding fathers Alan Turing, is to mimic the natural, human intelligence. In everyday life we intuitively use qualitative models, not numerical equations. For instance, the complete, realistic equation for behaviour of a child swing would be extremely complicated, yet a five year child knows how to "operate" the swing, and can describe her actions *qualitatively*, *e.g.* when to lean forward and backward to regulate the amplitude.

Induced qualitative models can offer more insight into the domain than numerical ones. The standard approach to regression modelling is fitting the data to a polynomial or another chosen function template. Although such models are sometimes considered symbolic, they do not offer any useful insight. For instance, the true and insightful numerical symbolic model for swinging of a simple pendulum would be a sine function like the one we get by solving the corresponding differential equations. Such solutions are difficult to induce from data using the current regression modelling tools. In contrast, qualitative model can provide a simple, but correct conceptual description: the pendulum swings back and forth. Given enough data, we can discover that the amplitude of the pendulum eventually decreases until the pendulum stops. Given data on multiple pendulums, we find out that longer strings yield longer periods, and that changing the weight has no effect. While these descriptions are insufficient for computing any actual periods, they often provide all the insight we need. For instance, a practical task may be to make the period of a pendulum match that of another one. The guidance provided by the qualitative model – *increase the length if the period is too long* and vice versa – would suffice to accomplish the task. Even when the final goal is to have a quantitative model, the qualitative one can be helpful in its construction [12].

Finally, such models can also be more applicable than numerical ones. For a simple example from economics, consider the law of demand: "the higher the price, the shorter the queue (other things left unchanged, less people are willing to buy things for a higher price)." Any numerical description of this relation would fail to give exact predictions since it would include variables which are not measurable with sufficient precision. Qualitative models, on the other hand, deliver what they promise, that is, correct qualitative predictions. The above simple qualitative rule is routinely, although not necessarily consciously, used to control the market prices.

The field of machine learning, which developed many methods for induction of (numerical) regression models, showed surprisingly little interest in learning of qualitative models from data. We will describe a suite of new machine learning algorithms with a common name Padé (an acronym for "partial derivative", and the name of a famous French mathematician). Padé first computes partial derivatives with respect to all independent attributes for all examples appearing in the data. Then it discards the quantitative information, the magnitude, which is difficult to estimate precisely, and only keeps the signs of derivatives, which represent qualitative relations between the independent and dependent attributes. Afterwards, we can use standard ma-

chine learning algorithms to induce predictive qualitative models, or venture into exploratory analysis and visualisation techniques.

We will continue the introduction with a formal definition of the problem and an overview of the related work. The following section describes the algorithms for computation of partial derivatives from the data. In the section on experiments we test the algorithms on artificial data sets specifically constructed to explore particular properties of the algorithms. We conclude with discussion of the experimental results and some remarks.

1.1 Problem definition

We define *qualitative partial derivative* of function $f(x_1, \ldots, x_n)$ with respect to attribute x_i as the sign of partial derivative,

$$\frac{\partial_Q f}{\partial_Q x_i} = \mathrm{sgn}\frac{\partial f}{\partial x_i} \tag{1}$$

The qualitative derivative can be increasing ($+$), decreasing ($-$) or steady ($\circ$). We will write the fact that a function is increasing, decreasing or steady with respect to x_i as $f = Q(+x_i)$, $f = Q(-x_i)$ and $f = Q(\circ x_i)$, respectively.

Qualitative models are models which describe how the qualitative behaviour of the function with respect to one attribute depends upon values of other attributes. For instance, function $f(x_1, x_2) = x_1 x_2$ increases with x_1 if x_2 is positive, and decreases with x_1 if x_2 is negative. If x_2 is zero, x_1 has no effect on the value of the function. This model can be written down in form of the following three rules:

$$\begin{aligned}&\text{if } x_2 > 0 \text{ then } f = Q(+x_1)\\ &\text{if } x_2 < 0 \text{ then } f = Q(-x_1)\\ &\text{if } x_2 = 0 \text{ then } f = Q(\circ x_1).\end{aligned}$$

Function arguments can also be discrete, as in the following qualitative relations between the price of a product and its type and size:

$$\begin{aligned}&\text{if ProductType = car}\\ &\qquad\text{then Price} = Q(+\mathrm{ProductSize})\\ &\text{if ProductType = computer}\\ &\qquad\text{then Price} = Q(-\mathrm{ProductSize})\end{aligned}$$

The task of qualitative modelling is to construct such models. In our case, we will induce them from data given as a set of learning examples. Each example is described by values of discrete or continuous attributes and with a continuous outcome. The outcome represents the value of some unknown function. The task is to describe the qualitative behaviour with respect to one or more attributes, conditioned by values of these or other attributes.

The method proposed in this paper solves the problem in two steps. First, we compute qualitative partial derivatives for each example. This translates modelling the function's behaviour into the training of classifiers which predict the qualitative derivative in different parts of attribute space. For instance, the above rules can be acquired by running the CN2 rule learning algorithm on examples labelled by qualitative partial derivatives.

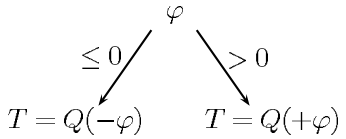
Separate models can be built for each attribute with respect to which we observe the function's behaviour. Alternatively, one can also build a classifier which predicts all qualitative derivatives at once.

$$T = Q(+l)$$

(a) Relation between the period and the length of the rope.

$$T = Q(\circ m)$$

(b) Relation between the period and the mass of the bob.



(c) Relation between the period and the initial amplitude.

Figure 1: Qualitative models describing the relations between the period T and the experimentally controlled variables.

1.2 Introductory example

Consider a set of experiments with a simple pendulum. The task is to learn qualitative relations between the period of the pendulum's first swing (T), the length of the rope, l, the mass of the bob, m, and the initial angle of displacement, φ. A sample of data collected in such an experiment is shown in Table 1 (first four columns).[1]

We can then use Padé to compute the qualitative relations for each measurement: $\partial_Q T/\partial_Q m$, $\partial_Q T/\partial_Q l$, $\partial_Q T/\partial_Q \varphi$, and append them to the original data (Table 1, last three columns). Finally, an algorithm for induction of classification trees is used to construct a qualitative tree for each qualitative relation, where examples are represented by original attributes and the partial derivative (*e.g.* one of the last three columns from Table 1) plays the role of the class. The resulting trees are shown in Fig. 1. Two of them have only a single leaf: the period always increases with the length of the rope ($T = Q(+l)$) and does not depend on the mass of the bob ($T = Q(\circ m)$). The tree describing the relation between T and φ says that for negative angles, T decreases with increasing angle while for positive an-

[1] A part of this experiment was actually performed using a Nao robot.

m	l	φ	T	$\partial_Q T/\partial_Q m$	$\partial_Q T/\partial_Q l$	$\partial_Q T/\partial_Q \varphi$
3.61	0.69	37.23	1.70	○	+	+
5.49	0.71	46.52	1.74	○	+	+
9.19	0.84	-48.91	1.91	○	+	−
7.17	0.33	33.89	1.17	○	+	+
6.81	0.50	65.93	1.51	○	+	+
4.64	0.69	-78.89	1.7	○	+	−

Table 1: A sample of data collected by experimenting with a simple pendulum.

gles, T increases when φ increases. We can reinterpret this as $T = Q(+|\varphi|)$.

1.3 Related work

Mathematical foundations of qualitative reasoning were established by the work of Kalagnanam and Simon [6, 7, 5], but building on the much older work of Samuelson [9] in economics, as well as the work on qualitative stability in ecology [4, 8].

Many algorithms have, in one way or another, tackled the problem of qualitative model induction from observation data. Algorithms QUIN and epQUIN [11, 10, 2] learn qualitative trees similar to those in Figure 1, except for a somewhat different definition of the relation in the leaf. Qualitative relations in QUIN are not based on partial derivatives, as in Padé, but on qualitative constraints. A constraint $z = M^{+-}(x, y)$ would state that for every pair of examples in which x increases and y decreases (or stays the same), the function value z increases. This also implies that the function value depends on no other attributes than x and y. QUIN constructs such trees by computing the qualitative change vectors between all pairs of examples in the data and then recursively splitting the space into regions which share common qualitative properties, such as the one given above. Although Padé combined with a tree learning algorithm can produce a similar tree as QUIN, the two methods are fundamentally different. Besides Padé being merely a preprocessor which can be used with any other machine learning algorithm or visualization technique, the crucial difference is that Padé considers individual examples while QUIN operates on pairs. As one of the consequences, Padé can compute qualitative (or numerical) derivatives for a particular point in the attribute space, while QUIN observes properties of regions of space.

Gerçeker and Say [3] fit polynomials to numerical data and use them to induce qualitative models. LYQUID is designed for modelling dynamic systems, where the data consists of traces sampled in time. We believe that this system could be adapted to also work in static systems.

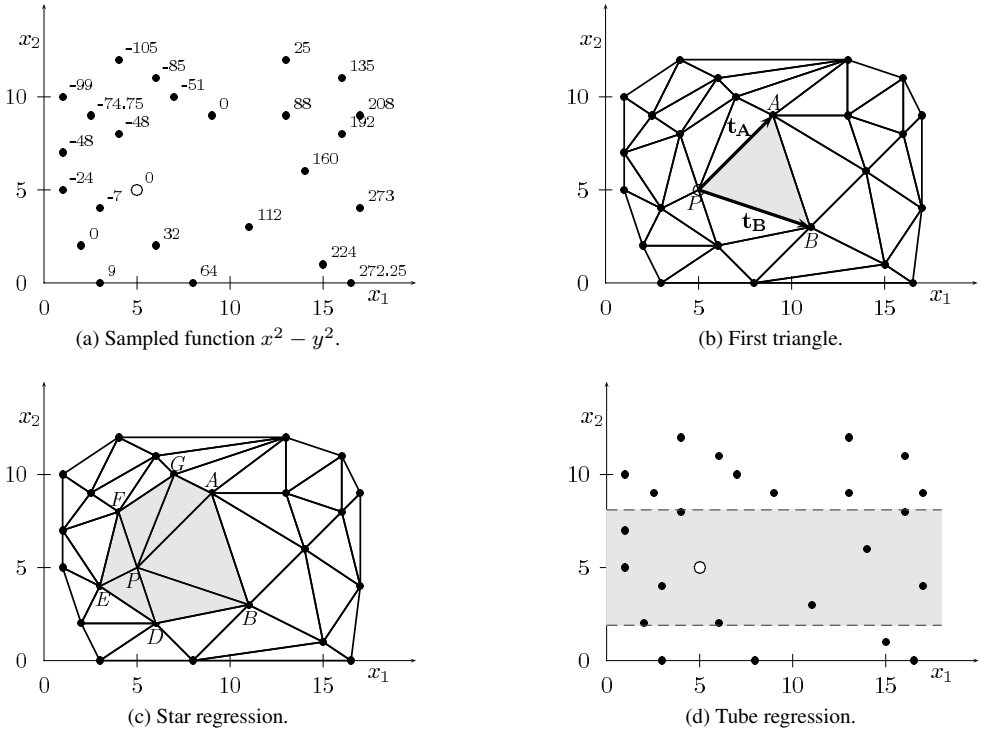
Padé differs from past methods in being, to our knowledge, the only algorithm for computing (partial) derivatives on point cloud data. An important difference between these algorithms and Padé is also that Padé is essentially a preprocessor while other algorithms induce a model. Padé merely augments the learning examples with additional labels, which can later be used by appropriate algorithms for induction of classification or regression models, or for visualization. This results in a number of Padé's advantages. For instance, to our knowledge, most other algorithms for learning qualitative models only handle numerical attributes, except for QDE learners which already take qualitative behaviours as input. One variant of Padé can use discrete attributes while computing the derivative, while with others we can use them later, when machine learning algorithms are applied to Padé's output.

The major contribution of this work, besides the idea of transforming the problem of qualitative modelling to standard induction of classifiers, are methods for computing partial derivatives of an unknown sampled function. In this respect it is related to numerical analysis for estimation of partial derivatives. Numerical analysis methods for computing partial derivatives are only useful for a function which is *known* in the sense that we can compute its value at any values of arguments which the algorithm requires. These methods are not appropriate for learning from data, where the function is sampled only in a limited number of points.

2 Algorithms

Let f be a continuous function of n arguments, $y = f(x_1, x_2, \ldots, x_n)$. The function is sampled in N points; the point's coordinates together with a function value represent a learning example. In machine learning terminology, each example is described by a list of *attributes*, $(a_1, a_2, \ldots a_n)$ and the outcome $f(a_1, a_2, \ldots, a_n)$. The basic task of Padé is to compute a partial derivative at each point (learning example) P in direction x_i.

Fig. 2a shows an example of such data for a function $f(x, y) = x^2 - y^2$. Each point represents a learning example, and the numbers beside the examples give the function values. We will compute the derivative w.r.t. x_1 at $P = (5, 5)$ marked by a hollow symbol. Other points used in computation will be marked as A, B, C and so on. We will treat these points as elements of affine space, and use $\mathbf{t_A} = A - P$, $\mathbf{t_B} = B - P$,... to denote the vectors from P to the corresponding points. We will also extend the definition of f to these vectors, *i.e* $f(\mathbf{t_A}) = f(A - P) := f(A) - f(P)$ and so forth. These linear transformations simplify the computation by setting up a coordinate system



(a) Sampled function $x^2 - y^2$. (b) First triangle.

(c) Star regression. (d) Tube regression.

Figure 2: An artificial, sampled function (a) and an illustration of Padé's methods (b-d).

in which the point P lies in the centre, $\mathbf{t_P} = 0$ and the corresponding function value $f(\mathbf{t_P})$ equals 0.

Formally, the partial derivative with respect to x_i at point P is defined as

$$\frac{\partial f}{\partial x_i}(P) = \lim_{h \to 0} \frac{f(P + h\mathbf{x_i}) - f(P)}{h}, \qquad (2)$$

where $\mathbf{x_i}$ is the i-th base vector.

This definition cannot be used directly on data since it involves infinitesimally small h and, furthermore, since we cannot compute the function value at arbitrary points. According to Taylor's theorem, the function can be treated as approximately linear in small neighbourhoods of P,

$$f(x_1, \ldots x_n) = b_0 + \sum_{i=1}^{n} b_i x_i + \epsilon, \qquad (3)$$

where ϵ represents the remainder in Taylor expansion (the function's non-linearity within the neighbourhood) and also any noise in the data.

The derivative with respect to x_i equals b_i in (3). The task is then to define a suitable neighbourhood and estimate the coefficient b_i accordingly. We will present three different ways for solving this problem. The first method determines the linear function f by simple linear interpolation over the simplex while the other two methods use linear regression.

2.1 First Triangle method

First triangle method models the function's behaviour by dividing the attribute space into simplices (our two-dimensional illustrations show them as triangles) by using the standard Delaunay triangulation [1] as shown in Fig. 2b. Let us assume that there is no noise in the data and that the sample is sufficiently dense so the function is approximately linear within each triangle.

Since the number of unknown coefficients b_i in (3) equals the number of vertices of the simplex, the coefficients can be found analytically by setting $\epsilon = 0$. Performing the calculation in vector space instead of in the affine space of points also eliminates the free term b_0 and point P.

Let $\mathbf{t_1}, \ldots, \mathbf{t_n}$ be the vectors from P to the vertices of the simplex which lies in direction x_i. We look for $b_1, \ldots, b_n$ which satisfy

$$[b_1 \ldots b_n][\mathbf{t_1} \ldots \mathbf{t_n}] = [f(\mathbf{t_1}) \ldots f(\mathbf{t_n})] \qquad (4)$$

(note that $\mathbf{t_i}$ are n-dimensional vectors) and thus

$$[b_1 \ldots b_n] = [f(\mathbf{t_1}) \ldots f(\mathbf{t_n})][\mathbf{t_1} \ldots \mathbf{t_n}]^{-1} \qquad (5)$$

For our two-dimensional example (Fig. 2b), we interpolate over the triangle PAB and compute the coefficients as

$$[b_1, b_2] = [f(\mathbf{t_A}), f(\mathbf{t_B})][\mathbf{t_A}, \mathbf{t_B}]^{-1},$$

which equals

$$[b_1, b_2] = [f(A) - f(P), f(B) - f(P)][A - P, B - P]^{-1}.$$

2.2 Star Regression

Star Regression is based on similar assumptions as the First triangle method, but improves its noise resistance by assuming the function's linearity across the entire star (the set of simplices surrounding a point) around the point P instead of just across a single simplex.

We can no longer use interpolation as in the First triangle, as it would result in a system with more equations than unknowns and would usually have no solution. We therefore allow non-zero error terms ϵ and translate the problem into computation of univariate linear regression over the vertices in the star.

If $\mathbf{t_1},\ldots,\mathbf{t_n}$ are the vectors from P to the vertices of the star, we compute b_i, which equals the derivative, as

$$b_i = \frac{\sum_j t_{ji} f(\mathbf{t_j})}{\sum_j t_{ji}^2}, \tag{6}$$

where t_{ji} represents the i-th component of the vector corresponding to the j-th point in the star, $\mathbf{t_j}$.

In our illustration (Fig. 2c), we compute the univariate linear regression over examples A, B, C, D, E and F, and use the first coefficient as derivative.

2.3 Tube Regression

Tube Regression adds even more noise resilience. Instead of triangulating, it considers a certain number of examples in a (hyper)tube passing through point P in direction parallel to the axis of differentiation (Fig. 2d; the tube is represented by the shaded area). We now assume that the function is approximately linear within short parts of the tube and again estimate the derivative from the corresponding coefficient computed by the univariate regression, this time over the examples in the tube.

Since the tube can also contain examples that lie quite far away from P, we weight the examples by their distances from P along the tube (that is, ignoring all dimensions but x_i). The weight of the j-th example in the tube equals

$$w_j = e^{-t_{ji}^2/\sigma^2}, \tag{7}$$

where t_{ji} is the i-th component of vector $\mathbf{t_j}$ (that is, the distance between P and the j-th example in direction parallel to the axis of differentiation, x_i). Parameter σ is chosen so that the farthest example in the tube has a user-set negligible weight. As a rule of thumb, we use tubes with 30 examples, with the farthest (*e.g.* the right-most point in the tube in Fig. 2d) having a weight of $w_{30} = 0.001$.

We then use the standard weighted univariate linear regression to compute the coefficient of the linear term b_i,

$$b_i = \frac{\sum_j w_j t_{ji} f(\mathbf{t_j})}{\sum_j w_j t_{ji}^2}. \tag{8}$$

The Tube regression is computed from a larger set of examples, so we can use the t-test to estimate the significance of the derivative. Significance together with the sign of b_i can be used to define qualitative derivatives in the following way: if the significance is above the user-specified threshold (*e.g.* $p \leq 0.7$), the qualitative derivative equals the sign of b_i; if significance is below the threshold we define the qualitative derivative to be *steady*, disregarding the sign of b_i.

2.4 Time complexity

Let N be the number of examples and n the number attributes (function arguments, dimensions).

First triangle and *Star regression* methods are based on Delaunay triangulation. The time complexity of its computation is difficult to assess without making any strong assumptions about the data. The state of the art qhull library needs, roughly, $O(2^n N \log N)$ to compute the triangulation (a detailed analysis can be found in [1]).

For each data point, the First triangle algorithm needs to find the triangle lying in the desired direction, which requires computing the determinant of a n-dimensional matrix for every triangle in the star. The time complexity is $O(Nn^3t)$, where t is the maximal number of triangles in any star. The value of t is again difficult to estimate, but it usually rises exponentially with the number of dimensions, which makes the time complexity $O(Nn^3 2^n)$. The total time complexity, including the triangulation, is thus $O(N2^n(\log N + n^3))$

Star regression computes univariate regression at every data point and has a time complexity of $O(Ns)$, where s is the number of points in the star. As s generally rises exponentially with the number of dimensions, the theoretical time complexity of this part of Star regression is $O(N2^n)$. The total complexity is dominated by that of triangulation and thus equals $O(2^n N \log N)$.

Tube regression finds the nearest neighbours of each of N examples, which takes $O(nN^2)$, followed by linear regression over the k examples in the tube. The total time complexity is $O(nN^2 + k) \approx O(nN^2)$.

Since these theoretical time complexities do not offer much insight into the algorithms' actual running times, we conducted a set of experiments with different number of examples and attributes. The goal function was random since it does not affect the time complexity. All experiments were run on a 2 GHz laptop with 2 GB of RAM. Results (Table 2) indicate that the time complexity of triangulation-based methods is indeed exponential in number of attributes and log linear in number of examples, while the Tube regression is linear in number of attributes and quadratic in number of examples. The exponential time complexity prevents the use of triangulation-based methods with more than four attributes.

3 Experiments

We evaluated Padé on a set of artificial data sets to observe the correctness of derivatives, its scalability with respect to the number of attributes, and its treatment of noise and of

	1000			2000			5000			10000		
	FT	SR	TR	FT	SR	TR	FT	SR	TR	FT	SR	TR
2	3	1	10	5	1	41	14	6	259	42	29	1159
4	242	5	22	449	19	86	1082	114	532	2280	473	2444
6	33049	1387	32	–	3908	134	–	–	857	–	–	3924
8	–	–	45	–	–	176	–	–	1163	–	–	5143
10	–	–	60	–	–	232	–	–	1627	–	–	6694

Table 2: Running times (in seconds) of First triangle (FT), Star regression (SR) and Tube regression (TR) methods for calculation of derivatives w.r.t. each attribute on data sets with 1000, 2000, 5000 and 10000 examples and 2, 4, 6, 8, 10 attributes. Symbol – denotes that the program ran out of memory.

discrete attributes. The accuracy is measured by comparing the predicted qualitative behaviour with the analytically derived true relation.

3.1 Accuracy

We observed the accuracy of Padé on a few mathematical functions. We estimated partial derivatives using Padé and compared them with the analytically obtained correct answers, except for the functions in Fig. 3d and Fig. 3e for which we computed numerical approximations of partial derivatives by Mathematica [14]. Note that this procedure does not require cross-validation or a similar form of data sampling since the known ground truth (the correct derivatives) is not used in the induction process.

Functions $f(x, y) = x^2 - y^2$ and $f(x, y) = xy$, with x and y sampled from $[-10, 10]$ were used as simple examples of functions which are continuous and differentiable in the whole interval. The heavily oscillating $f(x, y) = \sin x \sin y$, with x and y from $[-7, 7]$ represents a function whose qualitative behaviour changes frequently, so the partial derivatives are more difficult to compute and model. Functions $Im(\arcsin(x + iy)^4)$ and $Im(\text{arctanh}(x + iy)^3)$ in $[-2, 2]\times[-2, 2]$ are two examples of discontinuous functions. All functions are visualized in Fig. 3.

We computed the derivatives and trained the classifiers on ten random samples of 1000 points. Average proportions of correctly calculated qualitative derivatives are shown in Table 3. First triangle and Tube regression perform equally well, except for the Tube regression's failure on $f(x, y) = \sin x \sin y$. A visual exploration of predicted derivatives using a scatter plot clearly shows that this is due to the tube being too long and thus covering multiple periods of the function. Star regression's performance lags behind those of the other two algorithms.

To estimate the dependence of classification accuracy on data set size we conducted these same experiments on samples of 100 to 2000 data points. We found out that learning curves tend to flatten out at around 500 examples (Fig. 4). The general order of methods w.r.t their accuracy remains the same for all sample sizes, except for the last two functions, which are discontinuous and where the First Triangle method seems to suffer the least at very small samples.

3.2 Scalability to high number of dimensions

We checked the scalability of Padé to high dimensional spaces with an experiment with function $x^2 - y^2$, in which we added 98 attributes with random values from $[-10, 10]$ to the data. We use the Tube regression to calculate the derivatives since First triangle and Star regression cannot handle such high dimensional data due to their use of triangulation. We analysed the results by inducing classification trees with the computed qualitative derivatives as classes. Trees for derivatives by x and y agree well with the correct results (Fig. 5).

3.3 Robustness to noise

We sampled the function $f(x, y) = x^2 - y^2$ in 1000 points with x and y from $[-10, 10]$, and introduced uniform random noise of up to ± 20 to the function value. Since the First triangle and Star regression methods suppose no or little noise, we again tested only the Tube regression.

Induced models (Fig. 6) are correct and the split thresholds are surprisingly accurate given the huge relative amount of noise at around $x = 0$ and $y = 0$.

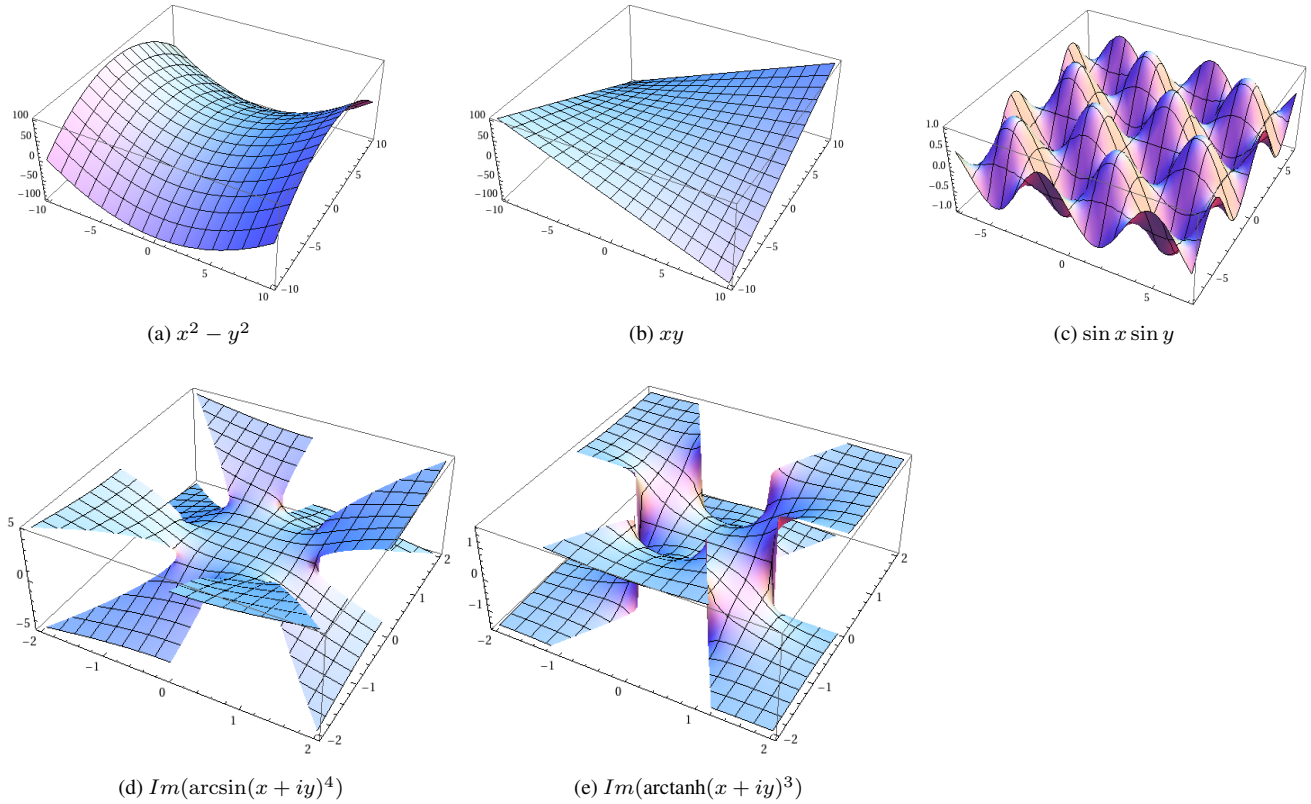
3.4 Handling of discrete attributes

We explored Padé's handling of discrete attributes on a function defined as

$$f(x, s) = \begin{cases} x/10 & ; s = 1 \\ 10x & ; s = 0 \end{cases}$$

Besides the continuous attribute x and boolean attribute s, the data set also included an attribute r with random values and no influence on f. Variables x and r were from the same definition range, $[-10, 10]$. The function was sampled in 400 points.

Tube Regression, whose results we used to construct a classification tree, found the correct solution (Fig. 7). Other methods failed to recognize the role of s, which they were given as a continuous attribute. Using dummy variables, like in statistical regression methods, does not work for triangulation-based Padé's methods.



(a) $x^2 - y^2$ (b) xy (c) $\sin x \sin y$

(d) $Im(\arcsin(x+iy)^4)$ (e) $Im(\text{arctanh}(x+iy)^3)$

Figure 3: Functions used in experiments.

$f(x,y)$	First Triangle		Star Regression		Tube Regression	
	$\partial_Q f/\partial_Q x$	$\partial_Q f/\partial_Q y$	$\partial_Q f/\partial_Q x$	$\partial_Q f/\partial_Q y$	$\partial_Q f/\partial_Q x$	$\partial_Q f/\partial_Q y$
$x^2 - y^2$	98%	98%	98%	97%	95%	95%
xy	99%	99%	92%	92%	99%	99%
$\sin x \sin y$	89%	89%	73%	72%	53%	53%
$Im(\arcsin(x+iy)^4)$	93%	93%	87%	86%	93%	93%
$Im(\text{arctanh}(x+iy)^3)$	91%	93%	76%	79%	86%	90%

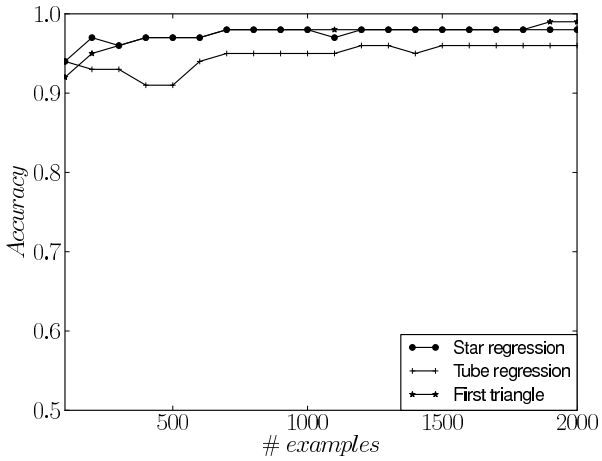
Table 3: The comparison of accuracies of Padé's methods on artificial datasets.

4 Conclusion

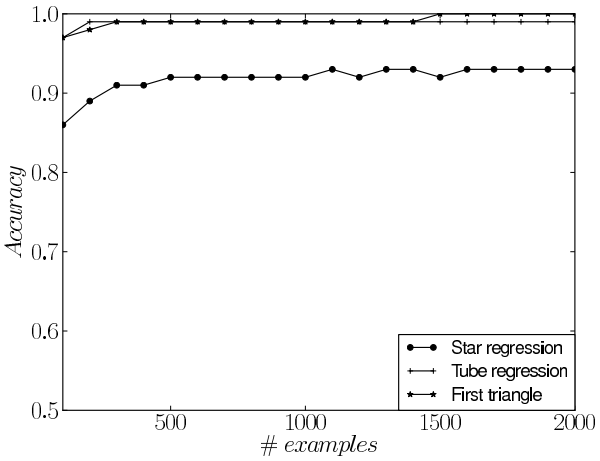
We described a new approach to induction of qualitative models whose advantage over (rare) existing similar algorithms is that it translates the problem into a standard supervised learning problem, which is one of the most researched fields in machine learning. The proposed translation requires computation of qualitative partial derivatives, which we defined simply as signs of ordinary partial derivatives. The biggest problem – and with that the core of this paper – is computation of partial derivatives of the function which is being modelled. Standard methods from numerical analysis cannot be applied here since they require a known function whereas in our case the function value is known only in a finite number of sampled examples.

We proposed three methods for this task. Two are based on triangulation and suppose either no noise or a small amount of noise. Besides this, the two methods are not likely to be useful in real-world scenarios which often contain more attributes than triangulation can handle. Experiments with the time complexity of the methods clearly show that the triangulation-based methods are unable to handle more than 6 attributes. Nevertheless, in absence of noise these two methods provide very good accuracy for functions with complex qualitative behaviour and low number of arguments, such as $f(x,y) = \sin x \sin y$. Tube regression on the other hand offers robustness to noise, scales well to high dimensional spaces and can also handle discrete function arguments with proper definition of metrics.

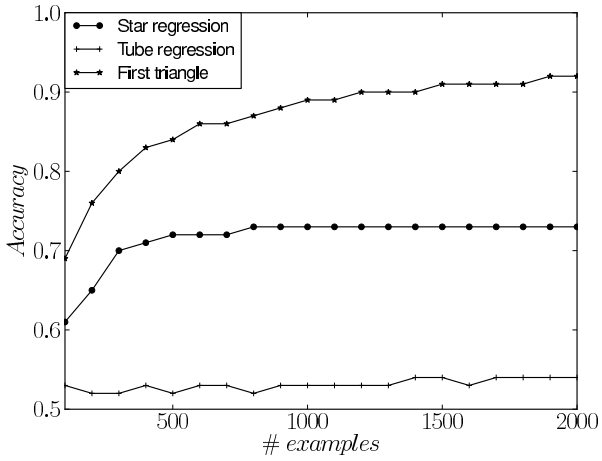
In general, the triangulation-based methods may be mostly of theoretical interest, while the Tube regression has all the features required of a practically useful machine



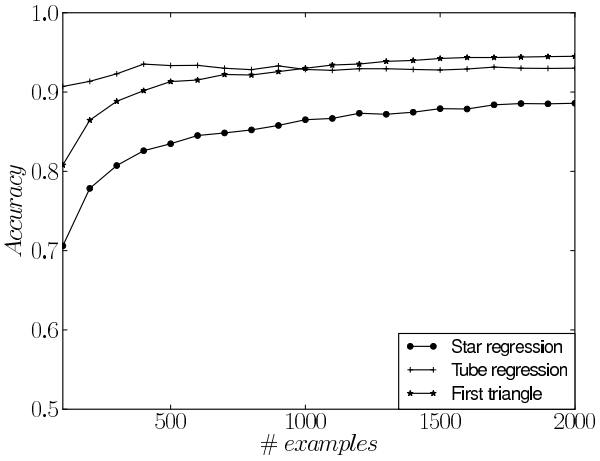
(a) $f(x, y) = x^2 - y^2$



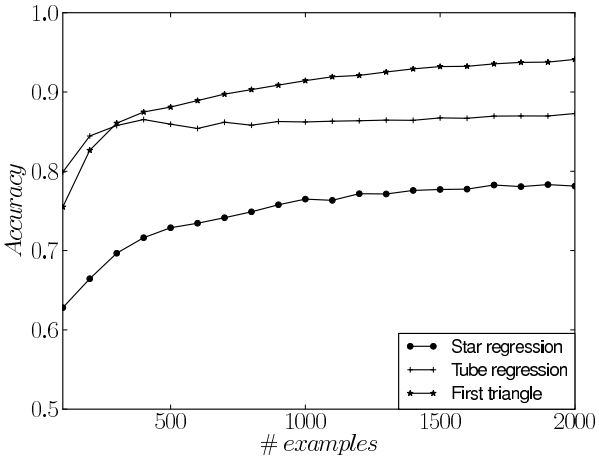
(b) $f(x, y) = xy$



(c) $f(x, y) = \sin x \sin y$

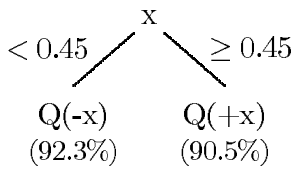


(d) $f(x, y) = Im(\arcsin(x + iy)^4)$

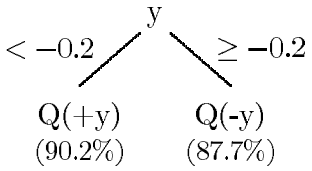


(e) $Im(\text{arctanh}(x + iy)^3)$

Figure 4: Accuracy of $\partial_Q f/\partial_Q x$ over different sizes of data sets. Results for $\partial_Q f/\partial_Q y$ are similar.

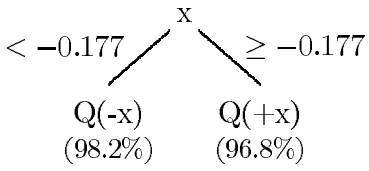


(a) Derivative w.r.t. x.

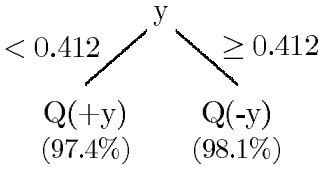


(b) Derivative w.r.t. to y.

Figure 5: Qualitative models of function $x^2 - y^2$ with 98 additional random attributes.



(a) Derivative w.r.t. x.



(b) Derivative w.r.t. to y.

Figure 6: Qualitative models of function $x^2 - y^2$ with added random uniform noise.

learning method.

We have put the methods at a practical test within European project XPERO (IST-29427). Our goal was to provide a robot with an algorithm for autonomous learning. We found qualitative models most suitable for this task. For example, a particular case was to discover the relation between the area of the ball in the image from robot's camera, and the robot's angle and distance from the ball [13]. The robot learnt that the area of the ball is increasing with decreasing distance and decreasing with increasing angle (the robot turning away from the ball, so it gradually vanishes from the robot's field of view).

Since the field of learning qualitative models from data is rather unexplored, the paper opens more new interesting questions than it answers. Pioneers of qualitative modelling who constructed the models manually were able to describe real phenomena using simpler models, not unlike the classification trees and rules presented here. Is this generally the case? Do simple learning algorithms like tree induction, suffice, or will actual problems require more sophisticated algorithm, such as, for instance, support vector machines?

This paper follows the mathematical definition of partial derivative which is essentially univariate. Partial derivatives are linear and do not interact: the effect of changing two quantities at the same time equals the sum of effects of changing each of them separately. The exception to this rule are certain kinds of singularities. Does this happen in practice, especially in qualitative descriptions of problems? Can it happen, for instance, that two economic measures used separately decrease the inflation while using both together would increase it? Is treating each attribute separately indeed appropriate? We leave these questions open for further research.

Acknowledgements

This work was supported by the Slovenian research agency ARRS (J2-2194, P2-0209) and by the European project XMEDIA under EC grant number IST-FP6-026978.

References

[1] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.

[2] Ivan Bratko and Dorian Šuc. Learning qualitative models. *AI Magazine*, 24(4):107–119, 2003.

[3] R. K. Gerçeker and A. Say. Using polynomial approximations to discover qualitative models. In *Proc. of the 20th International Workshop on Qualitative Reasoning*, Hanover, New Hampshire, 2006.

[4] Clark Jeffries. Qualitative stability and digraphs in model ecosystems. *Ecology*, 55(6):1415–1419, 1974.

[5] Jayant Kalagnanam and Herbert A. Simon. Directions for qualitative reasoning. *Computational Intelligence*, 8(2):308–315, 1992.

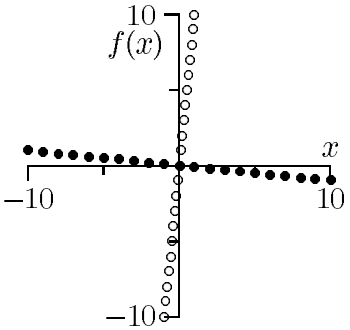
[6] Jayant Kalagnanam, Herbert A. Simon, and Yumi Iwasaki. The mathematical bases for qualitative reasoning. *IEEE Intelligent Systems*, 6(2):11–19, 1991.

[7] Jayant Ramarao Kalagnanam. *Qualitative analysis of system behaviour*. PhD thesis, Pittsburgh, PA, USA, 1992.

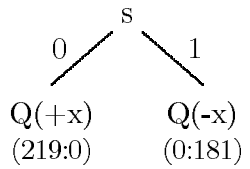
[8] Robert M. May. Qualitative stability in model ecosystems. *Ecology*, 54(3):638–641, 1973.

[9] Paul A. Samuelson. *Foundations of Economic Analysis*. Harvard University Press; Enlarged edition, 1983.

[10] Dorian Šuc. *Machine Reconstruction of Human Control Strategies*, volume 99 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, The Netherlands, 2003.



(a) Modelled function; filled and hollow circles represent examples with $s = 1$ and $s = 0$, respectively.



(b) Qualitative tree based on Tube Regression.

Figure 7: The function used in the experiment with discrete attributes and the corresponding qualitative model.

[11] Dorian Šuc and Ivan Bratko. Induction of qualitative trees. In L. De Raedt and P. Flach, editors, *Proceedings of the 12th European Conference on Machine Learning*, pages 442–453. Springer, 2001. Freiburg, Germany.

[12] Dorian Šuc, Daniel Vladušič, and Ivan Bratko. Qualitatively faithful quantitative prediction. *Artificial Intelligence*, 158(2):189–214, 2004.

[13] Jure Žabkar, Ivan Bratko, and Janez Demšar. Learning qualitative models through partial derivatives by Padé. In *Proceedings of the 21th International Workshop on Qualitative Reasoning*, Aberystwyth, U.K., 2007.

[14] Wolfram Research, Inc. Mathematica, version 7.0. Wolfram Research, Champaign, Illinois, 2008.

"Must-Work": A Scalable Model for Parallel Recursive Problems on P2P Networks

Mourad Amad, Djamil Aïssani
Laboratory LAMOS, University of Bejaia, Algeria
E-mail: amad.mourad@gmail.com, lamos_bejaia@hotmail.com

Toufik Bellal and Hameza Amrioui University of Bretagne Occidentale (Brest), France
E-mail: toufik.Bellal@univ-brest.fr, hameza.amrioui@univ-brest.fr

Keywords: P2P, on demand dynamic hierarchical master worker, recursive problems, parallel computing

Received: July 9, 2010

Most of real world problems are cpu-time intensive; their solutions take too much time to compute using a single machine. The grid and cloud computing offer a potential solution to this problem. However, such solutions are in general expensive. An alternative solution uses P2P networks, a set of machines in the Internet which collaborate to perform the same task. Branch-and-Bound is a model of such solution, but most of parallel applications developed on P2P networks are based on the Master Worker paradigm, particularly for divide and conquer problems (D&C), where the Master divides the tasks and sends them to Workers, while the Workers execute received tasks as in the client server model. This solution suffers from the scalability problem, and, as a consequence, hierarchical Master Worker model was introduced. Scalability is improved, but it still remains a critical issue. In this paper, we propose a new generic and scalable model for parallelizing the resolution of the recursive problem (a type of divide and conquer problems) on an existing P2P architecture. As opposed to the existing hierarchical Master-Worker models, in our solution, each network node is both a Master and Worker called "Must-Work". The proposed solution uses a dynamic tree for tasks distribution; it is constructed according to a node requestor. We have evaluated our solution using The Quicksort method under the MPI platform. The results are globally satisfactory in term of time execution compared to the sequential solution.

Povzetek: Članek opisuje način reševanja rekurzivnih problemov v omrežjih vsak-z-vsakim.

1 Introduction

At the present time, the mathematical resolution of the major optimization problems (*eg. vehicle routing problem, travelling salesman, minimum spanning tree, eight queens puzzle, Knapsack, Cutting stock, ...*) remains extremely complex and/or expensive in terms of machine time. It is however possible in the majority of the cases (*some problems are excluded, eg. TSP*) to distribute calculation on several machines, each one treating one part of the problem, under the aegis of a main authority. The multiprocessor machines or the material parallelization of machines single processors constitutes the first application of this concept. This involves nevertheless a high cost of sharing the number of processors to be acquired. An alternative solution is to distribute calculation on a network, but the put question is how to distribute this calculation with efficiency, optimization and fault tolerance?

Divide and conquer (D&C) is an important design paradigm based on multi-branched recursion. It works by recursively breaking down a problem into two or more sub-problems of the same type, until they become simple enough to be solved directly. The solutions to the sub-problems are then combined to give a solution to the original problem [8]. In fact, recursive problems (*our case study*) can be resolved by D&C type algorithms.

The recursivity is a very interesting field of data processing; it makes it possible to solve certain problems in a very fast way and very simple. The recursive problems require many means and resources, because enormous data must be stored in the stack of machine, which causes an overflow of this stack in case of one processor. Whereas, on a network (*eg. P2P*), this problem is not posed. As examples of the famous recursive problems, we can quote: the problem of the eight reins, Hanoi turn (*see figure 2*) and the Quicksort sorting (*our case study*).

Recursive problem belongs to the divide on conquer problem type, where Master-Worker models have been proposed for their solutions. In the conventional Master-Worker paradigm, a single supervisor process manages and controls a sets of processes. Distribution of tasks is performed in two phases: **1)** the distribution from supervisor to Master process, **2)** and that from Master process to Worker processes. The computed results are performed on the reverse way. The main problem of this solution is the central supervisor, and the overload of the Master.

The hierarchical Master Worker type solutions are then introduced to ameliorate the former one. In the hierarchical Master-Worker models, Masters are required to partition the problem-space (*preparing work for the Workers*), schedule work, balance the load of the Workers to maintain efficiency [11], and correlating their output into a global result. Workers simply perform given operations. This paradigm sustained good performance [3]. However, scalability is always an open issue [12].

In this paper, we consider the parallel resolution of the recursive problems (*divide and conquer type*). We propose a scalable and generic hierarchical Master Worker model based on a distributed tree diffusion which is constructed on demand by any Master node in any P2P network. It is characterized by high dynamicity (*node can join and leave the network at any time*), scalability (*hight number of node is supported*) and genericity (*for any type of P2P network and any recursive problem*).

The remainder of this paper is organized as follows: In the second section, we present a background and related works on parallel computing techniques, especially for divide and conquer problem type, with more importance to the hierarchical Master Worker models. In the third section, we give our contribution for parallelizing the recursive problem solutions on any existing P2P architecture. We give a performance evaluation of the proposed model for Quicksort method in section 4. Finally, we conclude and give some perspectives.

2 Background and Related Works

The resolution of the numerical problems which are expensive in terms of computing is a challenge. To solve a given problem more quickly, a natural idea consists of making simultaneously several agents cooperating for its resolution, which will thus work in parallel. Parallel calculation is a technique in which several actions are carried out simultaneously, so that the time of resolution is reduced. In addition to the material components intended for parallel calculation, a support by software components is also necessary, in order to coordinate the simultaneous execution of several lines of computer program code. Such dependence is necessary, because of the existing interdependencies between the various program codes [2].

Grid computing has become an alternative to traditional supercomputing environments for developing parallel applications in recent years [9]. Master-Worker paradigm is a common model to evaluate a pool of tasks, it is used by many scientific and engineering applications like tree search algorithms, genetic algorithms, training of neural networks, stochastic optimization, parameter analysis for engineering design, Monte Carlo simulation [10].

Master Worker paradigm is an adequate solution for divide and conquer problem such as our case study (*recursive problems*). It consists of two entities: a Master and multiple Workers. The Master decomposes the problem into small tasks and distributes these tasks among a farm of Worker processes, as well as for gathering the partial results in order to produce the final result of the computation. Worker processes execute in a very simple cycle: receive a message from the Master with the next task, process the task, and send back the result to the Master. This paradigm is well adapted to the first generation of P2P networks (*eg. Napster*[1]), because Napster is composed of one server which store a resource's index (*Master*). All other peers (*Workers*) are connected to the server in order to publish/search a resource (*execute tasks*). However, the most proposed Master-Worker solutions suffer from the scalability problem (*in terms of Master bottleneck but also security: the master is busy all time*) [11], more generally they are specific for only some problems and using also specific underlying architectures.

Given this, hierarchical Master-Worker model has been proposed; it uses submasters to decrease the workload of the Master, but a problem still exists: if the number of Workers or submasters grows, the submasters also will be bottleneck because many communications appear between Workers and their submasters. Hierarchical Master Worker using a shared memory space for work managing at the submasters was introduced by GhasemiGol et all (*Linda model*) [11]. In this model, Workers execute a task and put the results in a shared memory space on the submaster. Effectively, the solution reduces communication cost, but accessing shared memory is complicated and it is not practical in large scale network. Some Workers do not work voluntary, such as free rider on file sharing P2P applications which is opposed to our contribution, where nodes are called Must-Work.

In [14], the authors propose GVGE, a shared and interactive virtual collaborative geographic environment for solving geographic problems. It is composed of three layers: resource layer, service layer and application layer. GVGE is a platform that can manage grid applications. In [13], the authors propose a java environment for developing parallel programs limited to small clusters. In [15], ParCop is proposed using Master Worker model where each node manages two types of links: permanent and temporary pathways. The former makes connections with its neighbours, and the later makes connections between the Master and Worker during computational. Pathways can be more than one hop and then message transmissions consume more CPU. In [16], the authors propose how to build new types of groups called “similarity groups" into the JNGI project [17], in order to increase the relevance of task dispatching and therefore to increase the performance of JNGI. In [19], desktop grids inspired from biological systems, a large computational tasks are broken down into sufficiently small subtasks. Each subtask is encapsulated into a mobile agent. The management of the mobile agents is not shown in the paper. A similar work has been proposed in [20]. In their paper, the authors propose a middleware for parallel-based computations across a P2P network. It is different

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

from our work on the tree diffusion construction process, but also the fault tolerance consideration. In [22], the authors propose a nice construction recursive model that can complete our work by integrating it on each node.

The second generation of P2P network (*eg. Gnutella*) and the third generation (*eg. Chord*) can be a good candidates for hierarchical Master-Worker models; we just need to construct a tree with efficiency and optimization.

In this paper, we propose a generic and scalable model for parallel computing which can be implemented on any existing underlying P2P architectures (*structured and unstructured*), because it is based on the construction of a tree in a connex graph (*by exchange messages between neighboring nodes*) from any node without need of global knowledge of the underlying P2P architecture. In graph theory [23], the construction of a tree from source node to a set of receivers is already feasible if the graph is connected. The overlay P2P network is a connected graph.

P2P systems know an explosive growth in the few last years [1]; they progressively replace the Client/Server architecture. P2P systems are a good solutions for problems which need higher scalability, they resist to denial of service attacks, and held the top of the paving stone on Internet (*application layer*). Unlike the centralized systems such as Client/Server, in P2P systems, it is the hosts who provide the resources which will be available on the network [21]. These resources are those of the computers, the users and the connection they have. Parallel computing is an interesting example of P2P applications; it is also one of the most important solutions for the optimization problems. Many P2P architectures have been proposed for file sharing applications (*eg. Chord [6], CAN [5] and Tapestry [7]*). Figure 1 is an example of P2P architecture, we use it in performance evaluation section.

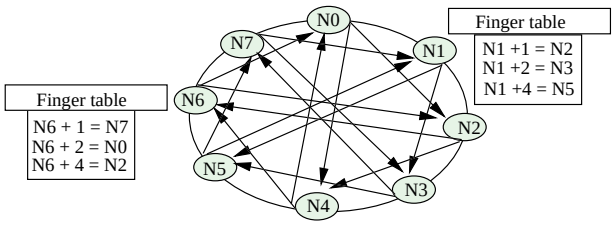


Figure 1: Example of P2P network (*Chord architecture*)

Our contribution aims to improve the scalability and contributes significantly to the genericity of the hierarchical Master Worker models by its important characteristics, it can be easily implemented on any existing P2P architecture, and then benefits from their advantages (*scalability, self organization, fault tolerance, ...*). The standard operational interpretation of a recursive program forms a tree of recursive calls spreading out from the node of the initial invocation [21]. Our solution is fundamentally based on dynamic task distribution tree constructed on demand under an existing P2P overlay network. The next section describes and analyzes the proposed solution.

3 Proposed Solution: "Must-Work"

In this paper, we propose a scalable and hierarchical Master Worker model based on a dynamic on demand tree construction, where each node is sometimes Master for a given tree, sometimes Worker for another tree, it is called Must-Work. When a node receives calculus request (*task*), it can't refuses it. If the task is complex, it plays the role of Master (*divides and conquers*). Otherwise, it is Worker (*executes task*). In some specific situations, a given node can be simultaneously Master for some Workers and Worker for some Masters. As illustrated on figure 2, the node N3 is a Master in `Hanoi(3,A,B,C)` problem resolution when N0 is a source, and it is a Worker in `Hanoi(3,A,B,C)` problem resolution when N2 is a source.

3.1 Functional principal

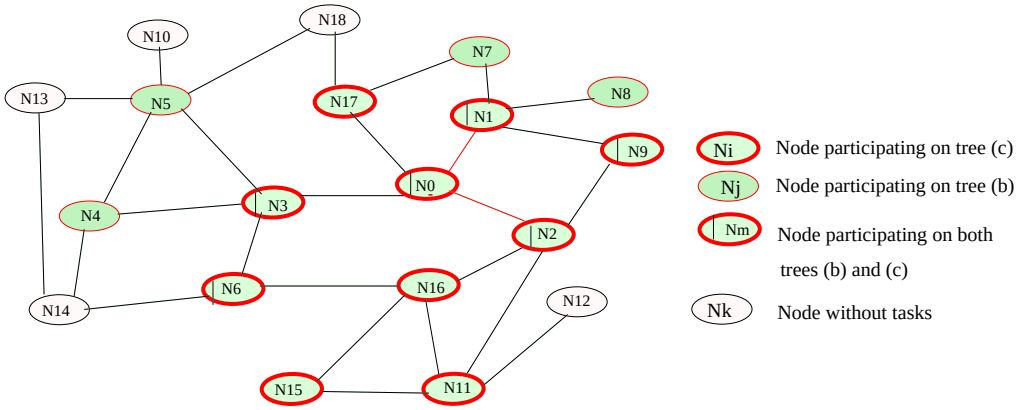
The parallelization of the recursive problems consists to break up the problem into independent calculation portions (*tasks*), and to diffuse each one on a dynamic on demand constructed tree branch from P2P underlying architecture, as long as the portion problem is decomposable. Let us arrive at an evident problem (*where its resolution is easy and does not need complex treatment or dividing*). The considered node makes calculation and returns result to the immediate requestor, and so on, until the machine initiator of calculation receives all the required calculation portions. With other words, the resolution of a recursive problem on existing P2P architecture can be transformed to a dynamic on demand diffusion tree construction, then to diffuse calculation and recovering it thereafter. The depth of the constructed diffusion tree depends on the recursive problems to solve. Figure 2 shows an example of the calculus distribution (*diffusion*) tree for the Hanoi tours problem. The network is composed of 19 nodes, the degree of the Hanoi problem is 3. Each Master divides the calculus on three parts, and sends them to their successors in the constructed tree, and so on.

Figure 2 illustrates an example for resolving simultaneously two Hanoi problems. The node N0 calculates `Hanoi (3,A,B,C)` and the node N2 calculates also `Hanoi(3,A,B,C)`, each one constructs it own tree on the same network. Node N3 participates on both resulting trees, it is a Master on the first tree (*figure 2.b*) and Worker on the second tree (*figure 2.c*).

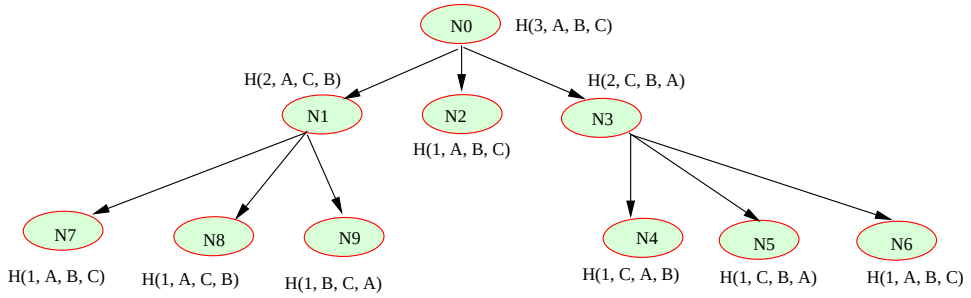
For illustration purpose, let consider the following notations:

3.2 On demand diffusion tasks tree construction

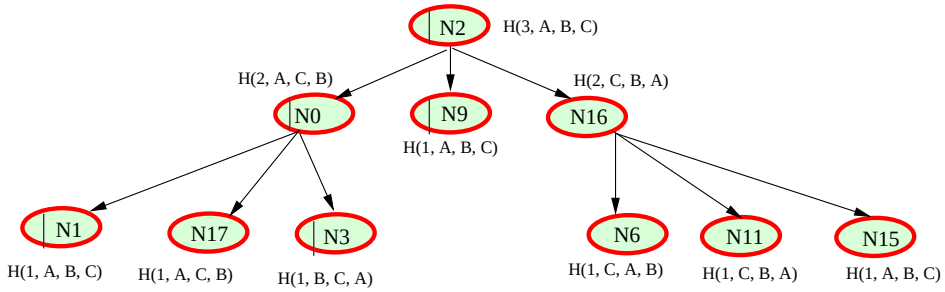
At the time of launching calculation decision, the initiator node N_i puts its state to *initiator*, and plays the role of Master, it builds a set of receiving children (*immediate successors at one hop in the dynamic on demand constructing tree*) by executing the function `Child(i)`. Before



(a): P2P network



(b): Resulting tree of Hanoi tours solution initiating by node N0



(c): Resulting tree of Hanoi tours solution initiating by node N2

Figure 2: Diffusion tree construction for Hanoi tours problem

Notation	Designation
N_i	Node identifier.
Child(j)	Function which returns child order j of the current node (j^{th} *successor*)
Nb_nd	Number of nodes in the network.
Nb_fi	Number of children of node N_i
Ei	Set of children of node N_i (*set of successors*).
Nb_aqi	Number of acquirements for node N_i (*number of recipients which have received the request from node* N_i).
D	Number of portions (*degree*) of the recursive problem to parallelized (*Parameter of the problem to be resolved*).
state	State of node N_i (*free, occupied, initiator*) initialized at free.
Parent(i)	Predecessor of node N_i.

Table 1: Conventional notations

sending a task to any node, it divides calculation into several portions (*tasks*) according to the degree of the problem (*eg. The Hanoi turn problem shown on figure 2 is with degree 3*). The requested node sends the message: '`are you free?`' to each neighbouring. If it receives a number (NPR) of positive responses higher than the number of tasks (D) to distribute, it diffuses them towards its children (*successors*) according to algorithm 1 (*the sender node should save the set of its neighboring nodes to which it has send a task*). Otherwise, it sends the first tasks to free nodes and the remainder tasks to occupied nodes if they exist, or diffuse a second work for the same children.

As illustrated on algorithm 2, at the reception of a calculation portion from node N_j, node N_i puts its state to *occupied*, it tests if the portion of the received calculus (*task*) is not decomposable, so yes (*node is a Worker*): it makes the last calculation, and returns back the result to its parent (*transmitting of calculation result using function parent(j)*), then it changes its state to *free*. Otherwise (*node is a Master*), it breaks up this portion into a set of sub portions (*eg. k sub portions*) and diffuses them towards the set of its children (*k successors*) as illustrated on algorithm 1.

When a node N_i receives a calculus result from one child, it decreases the number of the acquirements (*Nb_aqi–*), and when it receives all the sent calculus por-

Algorithm 1 : Starting the calculus by the initiator N_i (*root*)

1: Begin
2: Ei = ;
3: statei = *initiator*;
4: Nb_fi = card({neighboring nodes}) ;
5: Ei← {neighboring nodes};
6:If ($Nb\_fi \geq D$) **Then**
7: **For** m:=1 **To** D **Do**
8: k← identifier of the neighboring node (Ei) with higher capability;
9: Send the k^{th} calculus portion to child (k);
10: Ei←Ei - child (k);
11: **End**
12:Else
13: **For** k:=1 **To** Nb_fi **Do**
14: Send the K^{th} calculus portion to child (k);
15: **End**
16: **For** k:=Nb_fi +1 **To** D **Do**
17: Random (N_p);// here, we can choose the better nodes in terms of calculus capabilities;
18: Send the K^{th} calculus portion to child (N_p);
19: **End**
20: Nb_aqi= D;
21: End
22: statei = *free*;
23: End.

Algorithm 2 :At the reception of a calculus portion by node N_i from node N_j

1: Begin
2: Nb_aqi = 0;
3: Ei = ∅;
4: statei = *occupied*;
5: perei = j;
6: **If** (*the calculus portion is not decomposable*) **Then**
7: Do the calculus and return the result to perei (*requestor*);
8: statei = *free*;
9: **Else**
10: Execute Algorithm 1;
11: **End**
12: End.

Algorithm 3 : At the reception of each portion calculus result

1: Begin
2: Nb_aqi–;
3: **If** (*Nb_aqi=0*) **Then**
4: Do its calculus portion;
5: **If** (*statei= initiator*) **Then**
6: Return the final result
7: **Else**
8: Send the result to its parent (*parent(i)*);
9: **End**
10: statei= *free*;
11: **Else**
12: Wait the other results;
13: **End**
14: End.

tion results (*Nb_aqi=0*), it makes its calculation, and returns the final result if it is the initiator. Otherwise, it sends the result to the requestor node (*parent(i)*), and so on. This process is illustrated on algorithm 3.

3.3 Some considerations on the solution

In this section, we give some detail illustrations on the scalability, fault tolerance, load balancing of our proposed solution.

- **Scalability:** The nodes in the network are sometimes Masters, sometimes Workers; they have the same responsibility (*function*). They divide the calculus and distribute it to their successors (*neighboring nodes with one hop*), which forms progressively and dynamically a calculus diffusion tree, where only the leaf nodes (*Workers*) in the tree which does operational calculus. The inner nodes divide and conquer tasks (*Masters*). In fact, the solution does not suffer from the scalability problem, but it depends on the scalability of the underlying P2P overlay network on which is implemented (*It is more scalable on Chord architecture than on Gnutella, because Chord is more scalable than Gnutella*). On the other hand, the proposed solution is generic relatively to both underlying architecture and recursive problem to be solved.

- **Fault tolerance:** When a Master node distributes the calculus portion to its children (*successors*), it activates a predefined time-out (T), it waits until time out expiration, if it doesn't receives all calculus portion results, then it sends request to children (*Workers*) which have not returned results yet, asking them to give results. If they respond to the request and they does not finished their tasks, the requestor node increments the time-out (T). Otherwise (*the direct successor who has not given response is failed*), the requestor node redistributes their calculus portions (*supposed failed*) to other successors, and so on. The calculus is still continuous even with node failures.

  At each node, we associate a queue for storing the requested calculus (*tasks*) which have not served yet with FIFO service politic. Perhaps, the failure of the first Master (*root*) blocks the system, and the results can not be recuperated. However, in practical applications, root Master is managed by the users of application themselves, and then we assume that is a reliable node.

- **Load balancing:** In the proposed hierarchical Master-Worker solution; each node is a Master for some Workers, and a Worker for some Masters (*"Must Work"*). No node is more important than other nodes. The Master divides the complex calculus (*we assume here that each node is enough intelligent to do this*

treatment for any recursive problem), and then diffuses the calculus portions (*tasks*) to its children (*successors at one hop*). When it receives the result portions, it makes its treatment, and then sends the result to the requestor. When a node receives a calculation request which is not decomposable, it executes it locally, and returns the result to the requestor. In the case where all the children of a given Master are on state *occupied*, and this last has not achieved the distribution of all the calculus portions, it can send them to those with high capabilities using a scheduling strategy such as in [4].

4 Performance Evaluation

In order to evaluate the proposed solution, we tested it for sorting a vector with different sizes and using the Quicksort method. We use the MPI platform (*Message Passing Interface*)[2] under linux, which is a library of C/Fortran functions based on the message communications (*the most important are: MPI_send and MPI_recv*). Simulations are down on machine carried out in a personal computer with the following characteristics: 2.16 GHz and 1GB of RAM. A specific tool was developed for simulation purposes.

Figures 3 and 4 represent respectively the sorting execution times of two vectors with dimensions 15 and 500 by the Quicksort method, using both parallel program and sequential one implemented on Chord architecture.

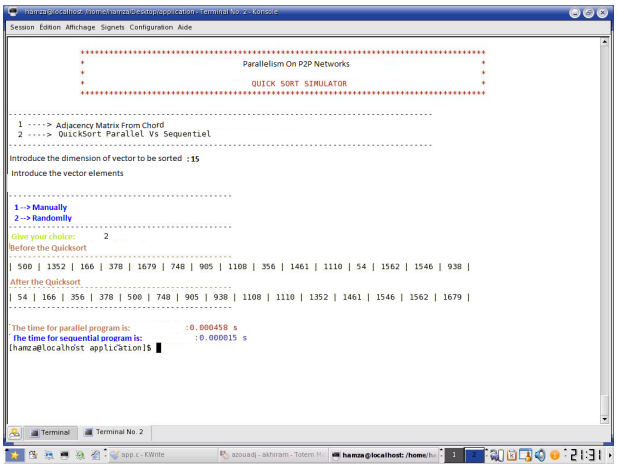


Figure 3: Sorting a vector of 15 elements using Quicksort method

For the vector of size 15 (*Figure 3*), the execution time for the sorting using the sequential solution is 0.000015s, and that using the parallel solution is 0.000458s. Sequential solution is better than parallel one. This is du to the high cost of message communication time comparatively to the execution time.

For a vector with 500 elements (*Figure 4*), the execution time for the sequential solution is 0.00326s. Whereas, the

[2]http://www.mcs.anl.gov/research/projects/mpi/

Figure 4: Sorting a vector of 500 elements using the Quicksort method

execution time for the parallel solution is 0.00292s. The results show that the proposed parallel computing is very interesting for the large scale problems, where communication time is negligible in front of calculation time. Table 2 gives a first idea of the average acceleration rate.

–	Sequential solution	Parallel solution
15 elements	0.000015s	0.000458s
500 elements	0.00326s	0.00292s
Rate	0.46	15.68

Table 2: Average acceleration rate

Figure 5 represents two curves of the execution time according to the vector dimension to be sorted using the two programs (*sequential and parallel*). It is seen clearly that for the vectors with small size, the execution time of the sequential program is better than that of the parallel program, because the inter nodes communication time (*exchanged messages*) is significantly important relatively to calculus time. Starting from a dimension of 500 (*500 values*), we note a significant difference in favour of the parallel program as illustrated on table 2.

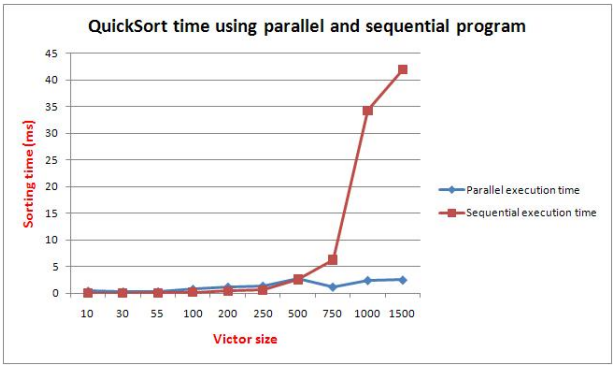


Figure 5: Execution time as function of the vector size

Figure 6 shows the execution time percent for both se-

quential program and parallel program from total execution time as function of the sorted vector size. We can clearly observe that for complex problem; the execution time for parallel solution is very improved compared to that of the sequential one. This improves the scalability of the proposed solution. From figure 6, we can see also that when the vector size reaches 500, the parallel solution becomes more interesting.

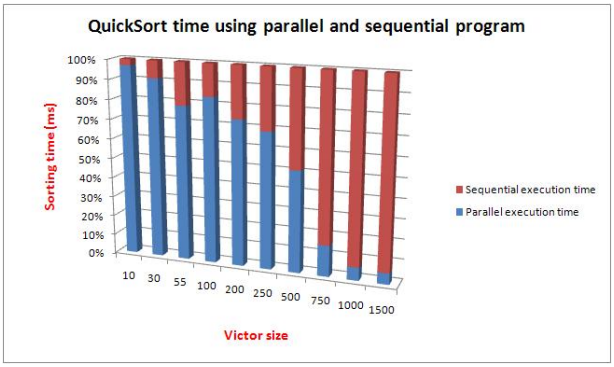


Figure 6: Percent of sequential and parallel execution times for sorting vector with the Quick sort method

Figure 7 shows the number of the generated messages du to the tree construction process (*for considering the free neighbouring nodes with higher capabilities*) as function of the degree of the Hanoi problem. When the degree of the Hanoi problem increase, the generated messages increase also. The most important number of messages is due to the diffusion tree construction process. When the diffusion tree is constructed, the number of overhead messages becomes stable.

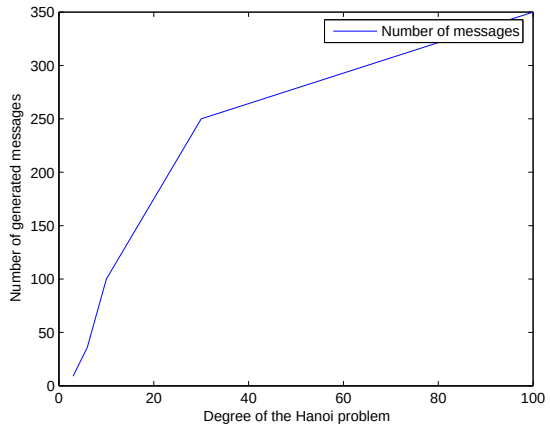


Figure 7: Number of generated messages as function of the degree of the Hanoi problem

5 Conclusions and Future Works

Master-Worker is a high-level programming framework that has been proposed to simplify the development of large scale parallel applications for computational grids. Many recent works are based on this concept. The hierarchical Master Worker paradigm is a very interesting approach to solve Divide and Conquer problem type , where the recursive problems are a good representative examples. The most difficult tasks are the fault tolerance, load balancing and scalability of the model.

In this paper, we have proposed a scalable hierarchical Master Worker model based on a dynamic on demand construction tree for tasks distribution. The solution is doubly generic; it can be used for any given parallel solutions of a recursive problem, but also it uses any existing P2P network as underlying architecture because P2P networks are a connected graphs. Each node in our model is both Master and Worker called "Must-Work".

By analyzing the various algorithms of the solution through validation examples, we can draw the following characteristics: the proposed solution minimizes the execution time in the case of great calculations, because parallelism allows the simultaneous execution of several tasks. It cures to the problems due to the recursivity, like the overflow of the stack as well as the capacity overshooting. It minimizes also the communication cost; each parent node assigns the calculation tasks directly to its children in its finger table (*using only one hop*). If a machine finishes its calculation, it will be released to carry out another calculation. In our model, the tasks are often affected to nodes with high capabilities.

The proposed solution guarantees the load balancing of the machines (*nodes*) in their affecting the tasks, since they are released according to the need of calculation. It thus minimizes the number of machines which participate at the same time on the resolution of the given recursive problem, because a child (*successor*) can be present in several finger tables, then since it will be released, it can be contacted by another parent (*predecessor*) for another task.

In terms of future works, first, we envision to test the proposed solution on a real specialized P2P platform, such as: Proactive[3] , XtremWeb[4] or JXTA[5]. Secondly, as opposed to Must-Work type node, we consider another node type which can refuses task execution, because it is belonging to an open P2P network (*contains free rider nodes*), we call it CAN-Work.

Acknowledgement

The authors would like to thank Mr A. Bendjoudi from CERIST and Miss W. Azzeghagh from University of Bejaia for their comments.

References

[1] C. Shirky, What is p2p..and what isn't, *O' Reilly Network*, 2001.

[3] http://proactive.inria.fr/
[4] http://www.xtremweb.net/
[5] https://jxta.dev.java.net/

[2] M. J. Flynn, Some computer organizations and their effectiveness, *IEEE Trans. Computers*, 21(9):948-960, 1972.

[3] K. Aida, W. Natsume and Y. F. Kata, Distributed Computing with Hierarchical Master-worker Paradigm for Parallel Branch and Bound Algorithm, *In Proceedings of the 3st International Symposium on Cluster Computing and the Grid (CCGRID)*, Tokyo, Japan, 2003.

[4] J.-P. Goux, S. Kulkarni, J. Linderoth and M. Yoder, "Master-Worker": An enabling framework for master-worker applications on the computational grid, *Cluster Computing*, Vol. 4, pp. 63-70, www.cs.wisc.edu/condor/doc/camera.doc, 2001.

[5] S. Ratnasamy, P. Francis, M. Handley, R. Karp and S. Shenker, A scalable content addressable network, *in ACM SIGCOMM, New York*, 2001.

[6] I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. F. Kaashoek, F. Dabek and H. Balakrishnan, Chord: A scalable peer-to-peer lookup service for internet application, *IEEE/ACM Transactions on networking*, Vol 11, No. 1, January 2003.

[7] B. Y. Zhao, J. Kubiatowich and A. D. Joseph, Tapestry: an infrastructure for fault-tolerant wide-area location and routing, *Technical report, No. UCB/CDS-01-1141*, Computer Science Division, University of California, Berkeley, April 2001.

[8] Z. Dai, F. Viale, X. Chi, D. Caromel and Z. Lu, Task-Based Fault-Tolerance Mechanism to Hierarchical Master/Worker with Divisible Tasks, *in Proceedings of the 11th IEEE International Conference on High Performance Computing and Communications*, Seoul, Korea, 2009.

[9] I. Foster and C. Kesselman, The Grid: Blueprint for a New Computing Infrastructure, Morgan-Kaufmann, 1999.

[10] J. Pierre, G. J. Linderoth and M. Yoder, Metacomputing and the Master-Worker Paradigm, *ANL/MCS-P792-0200, Mathematics and Computer Science Division, Argonne National Labroratory*, 2000.

[11] M. GhasemiGol, M. Sabzekar, H. Deldari and A. H. Bahmani, A Linda-based Hierarchical Master-Worker Model, *International Journal of Computer Theory and Engineering*, Vol. 1, No. 5, pp. 1793-8201, December, 2009.

[12] C. Banino, Scalability Limitations of the Master-Worker Paradigm for Grid Computing, *in proceedings of workshop on state-of-the-art in scientific computing (para'04), Denmark*, 2004.

[13] E. S. Manolakos and D. G. Galatopoulos *JavaPorts: An Environment to facilitate parallel computing on a heterogenous cluster of workstations*, Informatica, Vol. 23, pp. 97-105, 1999.

[14] J. Zhu, J. Gong, W. Liu, T. Song and J. Zhang, A collaborative virtual geographic environment based on P2P and Grids technologies , *Journal of Information Sciences*, Elsevier, Vol. 177, pp. 4621-4633, 2007.

[15] N. A. Al-Dmour and W. J. Teahan, ParCop: A Decentralized Peer-to-Peer Computing System, *In Proceedings of the ISPDC/HeteroPar'04, Cork, Ireland* , 2004.

[16] J. B. Ernst-Desmulier, J. Bourgeois, F. Spies and J. Verbeke, Adding New Features In A Peer-to-Peer Distributed Computing Framework, *in Proceedings of the 13th Euromicro Conference on Parallel, Distributed and Network-Based Processing (Euromicro-PDP'05), Lugano, Switzerland* , 2005.

[17] J. Verbeke, N. Nadgir, G. Ruetsch and I. Sharapov, Framework for peer-to-peer distributed computing in a heterogeneous, decentralized environment, *In Proceedings of GRID 2002, Baltimore, Sun Microsystems, Inc.*, Palo Alto, CA 94303, USA, January 2002.

[18] I. Podnar, M. Rajman, T. Luu, F. Klemm and K. Aberer, Beyond Term Indexing: A P2P Framework forWeb Information Retrieval, *Informatica, Vol. 30, pp. 153-161, 2006*.

[19] A. J. Chakravarti, G. Baumgartner and M. Lauria, The Organic Grid: Self-Organizing Computation on a Peer-to-Peer Network, *IEEE Transactions on systems, man, and cybernetics*, Vol. 35, No. 3, may 2005.

[20] W. Wadge, Distributed Application Reliability on Unstable, Dynamic, P2P-based platforms, *CSAW, Kalkara, Malta, 2004*.

[21] M. Gupta, S. Mukhopadhyay and N. Sinha, Automatic Parallelization of Recursive Procedures, *Int. Journal of Par. Prog., Vol. 28(6):537-562, 2000*.

[22] M. Haveraaen, Efficient parallelisation of recursive problems using constructive recursion, *Euro-Par 2000 - Parallel Processing, volume 1900, Lecture Notes in Computer Science, Springer Verlag, 2000, pp. 758-761*.

[23] Gasper Fijavz, ColoringWeighted Series-Parallel Graphs, *Informatica Vol. 30, pp. 321-324, 2006*.

Geographic Knowledge Discovery from Web 2.0 Technologies for Advance Collective Intelligence

Ickjai Lee and Christophcer Torpelund-Bruin
School of Business (IT), Cairns Campus, QLD 4870, Australia
E-mail: {Ickjai.Lee, Christopher.Torpelund}@jcu.edu.au

Keywords: geographic knowledge discovery, collective intelligence, Web 2.0 technologies, data mining, decision support

Received: February 24, 2011

Collective intelligence is currently a hot topic within the Web and Geoinformatics communities. Research into ways of producing advances with collective intelligence is becoming increasingly popular. This article introduces a novel approach to collective intelligence with the use of geographic knowledge discovery to determine spatially referenced patterns and models from the Geospatial Web which are used for supporting decisions. The article details the latest Web 2.0 technologies which make geographic knowledge discovery from the Geospatial Web possible to produce advanced collective intelligence. The process is explored and illustrated in detail, and use cases demonstrate the potential usefulness. Finally, potential pitfalls are discussed.

Povzetek: Članek se ukvarja z obdelavo geografskih podatkov s tehnologijo spleta 2.0.

1 Introduction

The second generation of web development and design have been coined as Web 2.0 technologies, where 2.0 refers to the historical context of web businesses coming back after the dot-com collapse [16]. Web 2.0 incorporates the move from Web-as-information-source architecture to the concept of Web-as-participation-platform, whereby users are encouraged to add value to the application as they create and collaborate information. The openness and freedom of user participation paves the way for Collective Intelligence (CI) which allows applications to be continuously improved to deeper the relationship with the users. This cycle of improvement is known as the perpetual beta, where a final version of the application is never reached - it simply continues to become better by offering more targeted experiences for each user according to their personal need [2].

As users of Web 2.0 services have grown, the functionality that the services provide have evolved into real world oriented human functions [17, 19, 22]. This implies the merging of geographical information with the abstract information that currently dominates the Internet. Often is the case that a user will search for something based on added spatial and temporal constraints. For example, "what is the best restaurant closest to a location x?", or, "how long will it take to get to the nearest hospital?". The merging of information with the real world has been dubbed as the Geospatial Web or Geoweb for short. The current explosion of digital geographic and geo-referenced datasets is said to be the most dramatic shift in the information environment for geographic research since the Age of Discovery [14]. Virtual globes such as Google Earth and NASA World Wind as well as mapping websites such as Google Maps, Live Search Maps and Yahoo Maps have been major factors in raising awareness towards the importance of geography and location as a means to index information. However, current collective intelligence techniques often fail to take into account these added spatial and temporal dimensions on user interactions and contributions. By considering these added dimensions, particular patterns and knowledge could be discovered about the users which could improve the accuracy of collective intelligence techniques.

Producing collective intelligence is a difficult challenge with the already vast amounts of user generated datasets on the Internet. The problem can become complicated when dealing with datasets with added geographic dimensions. The following are potential challenges associated with Geographic Knowledge Discovery (GKD) on spatially referenced datasets: 1) data access (inaccessibility) challenge; 2) diverse data types (inconsistency) challenge; 3) user interface (unavailablity) challenge. What has been proposed to overcome some of the challenges related to GKD is the need for a solid geographic foundation that accommodates the unique characteristics and challenges presented by geospatial data. Current national and global geospatial data lacks a proper infrastructure whereby contributed data can be aggregated and fully utilized for CI.

We propose to explore the use of GKD as a new technique for generating CI. GKD is an extension of Knowledge Discovery from Databases (KDD) and is based on a belief that there is novel and useful geographic knowledge hidden in the unprecedented amount and scope of digital geo-referenced data being collected, archived and shared by researchers, public agencies and the private sector [14].

Previous work [20] briefly explores the GKD model from Geoweb whilst current work extends it to extensive collective intelligence through GKD from Geoweb. Using the Voronoi diagram for Geoweb for emergency management [21] has been reported and algorithmic aspect of web map segmentation has been reported [11]. In this article, we investigate new emerging Web 2.0 technologies and whether they provide a means for generating the foundation that can be used to conduct GKD effectively to produce highly accurate CI for profitable traffic. The main aim of this article is not to empirically evaluate the performance of proposed framework with recommender systems and other visualization approaches, but to illustrate how Web 2.0 and Geoweb technologies could be used for GKD processes and advanced CI. Our proposed advanced CI process is abstractly described in Figure 1. Web 2.0 technologies are particularly used for user-oriented data selection and visualization. The proposed CI process can be used as an exploratory tool rather than a confirmatory tool. The main aim of this article is not to quantitatively compare and contrast with recommender systems, but to illustrate GKD processes from Web 2.0 and Geoweb technologies for advanced collective intelligence. Case studies show that how the generalized Voronoi diagrams and clustering can be combined and used with user-oriented datasets available through Web 2.0 and Geoweb. They demonstrate the potential usefulness and applicability of our proposed framework. The structure of the remainder of this article is as follows: Section 2 provides an overview of the latest Web 2.0 technologies which can be used to integrate the process. Section 3 describes the KDD processes along with the GKD for CI process. Section 4 demonstrates case studies using the GKD for CI process. Finally, section 5 concludes the article by listing potential pitfalls of the proposed process.

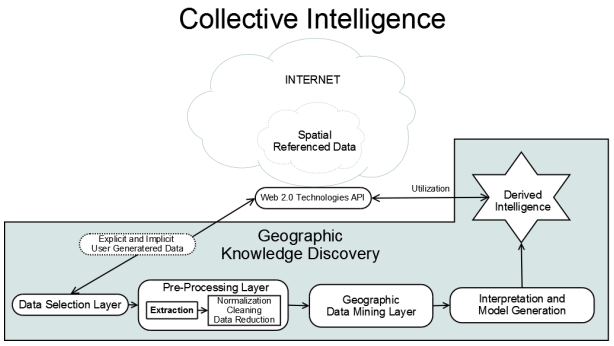


Figure 1: GDM on Web 2.0 technologies for CI process structure.

2 Web 2.0 Technologies

Web 2.0 technologies cannot be summed up and generalized but are instead a complex and continually evolving technology infrastructure which can include server-software, content-syndication, messaging-protocols, standards oriented browsers with plug-ins and extensions, and various client-applications. The encapsulating services may use just one or a combination of technologies, as the models defining the technologies are designed for hackability and remixability following open standards [2]. This necessitates fewer restrictions and allows for wider adoption and reuse. This infrastructure of complementary technologies provide services with information-storage, creation, and dissemination challenges and capabilities that go beyond what the public formerly expected in the environment of the so-called "Web 1.0". With the capabilities come the possibilities CI, but only if the challenges are overcome by the wide spread adoption of open Web 2.0 standards. Some of the common and standard Web 2.0 technologies used in the paper include:

- **Folksonomies:** The ability to allow collaborative tagging, social classification, social indexing, and social tagging.

- **Extensible Markup Language (XML) and/or Java Script Object Notation (JSON):** A general purpose specification for creating custom markup languages. Its primary purpose is to help share structured data based on user defined elements. XML document are compiled with a particular schema/Data Type Definition (DTD) in order to become well-formed and valid. JSON is a lightweight data interchange format for representing objects.

- **Really Simple Syndication (RSS) or Atom Feeds:** An extension of XML, allows the syndication, aggregation and notification of data. The feed can contain headlines, full text articles, summaries, metadata, data and various multimedia.

- **Simple API for XML (SAX), Document Object Model (DOM) and Extensible Stylesheet Language (XSL):** Not only is storage and distribution of data important, but so is the ability to extract useful information from the data. For this, both the data and multiple schema/DTD which define the data are required. SAX and DOM are Application Program Interfaces (APIs) for inspecting the entire contents of the data. XML Path Language (XPATH) and XML Query Language (XQUERY) act as filters designed as XSL which transform the XML document and allow specific queries.

- **Asynchronous JavaScript and XML (AJAX), Adobe Flex, JavaFX and Microsoft Silverlight:** Allowing development and deployment of cross platform rich Internet applications with immersive media and content. These applications utilize Remote Method Invocations (RMI) and Remote Procedural Calls (RPC) to servers to allow distributed interprocess communications. Web 2.0 application layer

protocols that allow this functionality include Simple Object Access Protocol (SOAP), Representational State Transfer (REST) and XML-RPC.

– **Mashups:** The merging of content from different sources, both client- and server-side.

CI is based on derived intelligence extracted from explicit and implicit user generated data, and therefore data representation is a core component for CI. XML is recommended by the World Wide Web Consortium (W3C) and is a fee-free open syntax which can be used to share information between different kinds of computers, different applications, and different organizations. This openness is highly important because it allows accessible-by-all data without needing to pass through many layers of conversion. Without XML, core components of Web 2.0 technologies would not be able to collaborate and achieve CI. The list of collaborating technologies exchanging information in XML is constantly varying - which reflects the precise character of the perpetual-beta. Current XML-based technologies which can be used for CI include Web Services Description Language (WSDL), Web Ontology Language (OWL), Linguistics Markup Language (LGML), Attention Profiling Markup Language (APML), Geography Markup Language (GML), and Predictive Model Markup Language (PMML). The languages defining various information can each be separated into the groups of: collaboration-based, explicit-based, implicit-based and intelligence-based. Section 3 describes each of these components working together for CI. The following sub sections briefly introduce each of these technologies which are then combined into the GKD from the Geoweb for CI process.

2.1 Web services description language (WSDL)

The first part of collaborating services is to provide a way for the services to communicate and describe what services they offer. The Services Description Language Version 2.0 (WSDL 2.0), is a W3C recommended XML language for describing Web services. The WSDL describes Web services in two fundamental stages. The first being abstract or document driven, which describes a Web service in terms of the messages it sends and receives; messages are described independent of a specific wire format using a type system, typically XML schema. The way messages are exchanged defines an operation which is defined by a message exchange pattern which identifies the sequence and cardinality of messages sent and/or received as well as who they are logically sent to and/or received from. The second stage defines the concrete or procedural-oriented level of the service, which defines how a service accepts bindings and associates with network endpoints, or ports [5]. The data exchanged by the Web service are defined as elements and are described with a unique name, and data type. Elements can be of simple types, complex types or be defined in an XML Schema Definition (XSD), DTD, REgular LAnguage for XML Next Generation (RelaxNG) and Resource Description Framework (RDF) file.

2.2 Web services choreography description language

The Web Services Choreography Language (WSCL) is a W3C candidate recommendation targeted for composing interoperable, peer-to-peer collaborations between any type of participant regardless of the supporting platform or programming model used by the implementation of the hosting environment [8]. The WSCL is a collection of components which builds an architecture stack targeted for integrating interacting applications which consists of:

– Defining the basic formatting of a message and the basic delivery options (SOAP);

– Describing the static interface and data types of the Web service end points (WSDL);

– Allows publishing the availability of a Web Service and its discovery from service requesters (Registry);

– Allows authentication of participants to ensure that exchanged information are legitimate and not modified or forged (Security layer);

– Allows reliable and ordered delivery between participants (Reliable Messaging layer);

– Allows the use of protocols for long-lived business transactions and enables participants to meet correctness requirements (Context, Coordination and Transaction layer);

– Describes the execution logic of Web services and rules for consistently managing non-observable data (Business Process Languages layer);

– Defines a common viewpoint of the collaborating participants describing their complementary observable behavior (Choreography layer);

The draft insists that the future of E-Business applications requires the ability to perform long-lived, peer-to-peer collaborations between the participating services, within or across the trusted domains of an organization. The WSCDL is the means by which technical multi-participant contracts can be created and viewed from a global perspective.

2.3 Attention profiling markup language (APML)

The Attention Profiling Mark-up Language (APML) is an XML-based portable file format containing a description of the user's rated interests. The APML also attempts

to contain other forms of attention data such as Attention.XML, Instant Messaging (IM) conversations, browser history, emails and other documents. The APML promises to make it easier for Web services to collect attention information of individual users to cater for the needs of individual and general users. The most compelling reason for the adoption of APML is that it defines an open and public standard of profiling that the user has direct access to. This means the user can directly be aware of what information is being shared about them and certain that Web services can provide exactly what they want. This differs from traditional captured user information by companies which tends sometimes be regarded as private and sacred. Attention information is kept up-to-date because APML is a lossy format, which maintains only the current trends and styles of the user.

2.4 The semantic web: web ontology language (OWL) and resource description framework (RDF)

The OWL and RDF are considered as the core technologies underpinning the Semantic Web; a collaborative effort led by W3C with participation from a large number of researchers and industrial partners with the aim to separate data from specific applications and making it possible for the web to understand and satisfy the requests of people and machines to use the Web content. The Semantic Web is not only concerned about the integration and combination of data drawn from diverse sources, but also how the data relates to real world objects so that both people and machines may understand and analyze the data on the Web. The OWL and RDF achieve this by publishing in languages specifically designed for data rather than just documents and the links between them. The network of linked data has been described as the Giant Global Graph (GGG), as opposed to the HTML-based World Wide Web (WWW) [4].

The OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans [3]. OWL 1.0 is currently a W3C recommendation and is currently being updated to OWL 2.0 though a working draft. On top of the features of OWL 1.0, OWL 2.0 is designed to facilitate ontology development providing classes, properties, individuals, and data values stored as Semantic Web documents, with the ultimate goal of making Web content more accessible to machines. The Multimedia Web Ontology Language (MOWL) is a further refinement by the W3C which has been designed to facilitate semantic interactions with multimedia contents. The MOWL was also merged with the Knowledge Description Language (KDL) to allow semantic processing of media data calls for perceptual modeling of domain concepts with their media properties. A further extension of MOWL allows semantics for spatio-temporal relations across media objects and events. The OWL and MOWL are most commonly serialized using RDF/XML syntax.

The RDF is a W3C recommended extension and revision of XML for conceptually describing and modeling information implemented in web resources. The fundamental aim is to identifying information using Web identifiers (using Uniform Resource Identifiers, or URIs), and describe it in the form of a subject-predicate-object triple expression so that machine intelligence can store, exchange, and use machine-readable information distributed throughout the Web. The information is represented as a graph of nodes and arcs, with each node being referenced by a unique URI. This allow data to be processed outside the particular environment in which it was created, in a fashion that can work at Internet scale [9]. The triple describes the relationship of the subject and the object of the information given the conditional predicate. An example from the W3C RDF primer describes the statement: "http://www.example.org/index.html has a creator whose value is John Smith", as the following RDF statement:

- a subject http://www.example.org/index.html;
- a predicate http://purl.org/dc/elements/1.1/creator;
- and an object http://www.example.org/staffid/85740.

The URI references are used to identify not only the subject of the original statement, but also the predicate and object, instead of using the words "creator" and "John Smith", respectively [1]. Another particular format might be more direct and easily understood, however the RDF's generality and potential for collaborative intelligence through sharing gives it great value. Another advantage to the RDF is the URIs can define real locations of the referenced information. In this sense, the RDF can also provide a means for geospatial indexing of the information which can be used by the GKD process to identify particularly interesting patterns.

2.5 Geography markup language (GML)

GML serves as a modeling language for geographic systems as well as an open interchange format for geographic transactions on the Internet. It is an extension to XML that allows the ability to integrate all forms of geographic information (discrete, areal and sensor) onto data. It does this by allowing a rich set of primitives that include features, geometry, coordinate reference system, time, dynamic features, coverage, unit of measure and map presentation styling rules. The way that data is represented by GML is defined by a GML profile namespace which defines restricted subsets of GML. These profiles can be built on specific GML profiles or use the full GML schema set. The GML can be used as a standalone data format or be included as an extension to other XML-based formats to give added spatial dimensions.

2.6 The predictive model markup language (PMML)

The Predictive Model Markup Language (PMML) is an application and system independent interchange format for statistical and data mining models [18]. It is an XML-based language developed by the Data Mining Group (DMG) and allows models to be created within one vendor's application, and use other vendors' applications to visualize, analyze, evaluate or otherwise use the models. Previously, the exchange of fully trained or parameterized analytic models was very difficult, but PMML allows effective utilization between applications and is complementary to many other data mining standards. The PMML also defines the input and output format of data and how, in terms of standard data mining terminology, to interpret their results. This kind of intelligence sharing is critical between collaborating CI servers and clusters and allows for an ensemble of different models which can be used to increase the accuracy of classification [10].

3 Geographic Knowledge Discovery for Collective Intelligence

3.1 Knowledge discovery process

Knowledge discovery is the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data. It is often described as deriving knowledge from the input data that can be used for further usage and discovery in the process. The process generally consists of several steps that can be executed in a non-linear order. The generic steps include:

- **Data Selection**: Creating a subset of the total data that focuses on chosen foci for concentrating the data mining activities.
- **Pre-Processing**: The cleaning of the selected data to remove noise, eliminate duplicate records, filling in missing data fields and reducing both the dimensionality and numerosity of the data in order to build and an efficient representation of the information space.
- **Data Mining**: The attempts to uncover interesting patterns.
- **Interpretation and Reporting**: The evaluation and attempted understanding of the results of the data mining process.
- **Utilization**: The use of the learned knowledge to provide accurate decision support for the utilizing industry.

Data mining is an ongoing popular research topic that focuses on the algorithms for revealing hidden patterns and information in the data. These include segmentation, dependence analysis, deviation and outlier analysis, regression and cluster analysis [7, 13]. The possible types of features of the data can be nominal, ordinal, interval-scaled, ratioed and any combination of all these types. Depending on the type of data, a distance metric is used to measure similarity and dissimilarity between the objects. By comparing these similarity and dissimilarity metrics, interesting patterns can be found within the data [6].

3.2 Collaborating web services for CI

In order to perform GKD for CI, a well formed system must be established in order to coordinate contributing services. Using Web 2.0 technologies and the Semantic Web, we define a process whereby Web services may share information in order to improve decision-making. This process is defined in Figure 2. A Web service can collaborate

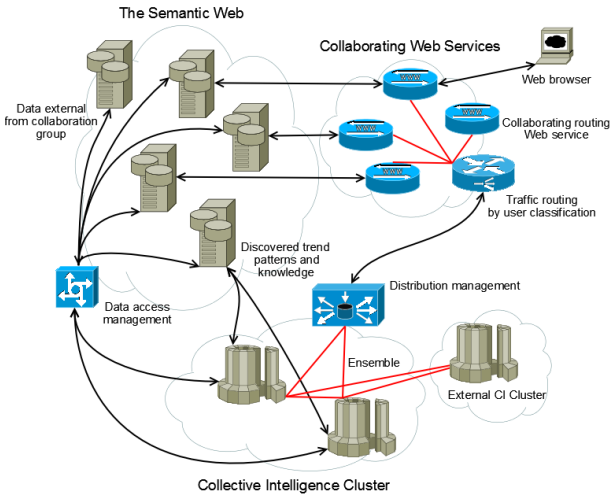


Figure 2: Collaborating Web services for CI.

with others in various ways. If the users of a transaction-based Web service are willing, then information regarding the users, products and transactions can be shared and used within the GKD process to discover patterns and knowledge for effective decision-making. However, a Web service can also not contribute information and yet still collaborate to achieve effective decision-making. These kinds of Web services interact with the discovered patterns and knowledge to route to the contributing transaction-based Web services. If traffic is routed from a collaborating Web service to an eventual profitable action, then both Web service share the profits. In this configuration, a Web service can still make a profitable action when a user does not initiate a transaction with them by effectively routing the user to a collaborating partner. A non-transaction-based Web service can profit by being popular among Web users and routing profitable traffic to transaction-based Web services. How the user is routed is determined by the discovered patterns and knowledge by the GKD process. In order for a Web service to begin collaboration, the WSCDL is used for

determining Web service information which can consist of:

- Messaging format and service end-points agreement with WSDL;
- If contributing information or purely routing-based;
- Determine security rules and how to access information provided (if any);
- Determine business rules and action of successful profitable traffic.

Information being gathered by Web services on the current user can be given to the traffic router which classifies the user and provides routing to potentially profitable destinations. The information can be anything from user gathered details, shopping basket analysis or blog, forum, tagging and rating analysis to determine APML-based information. How the user is classified is determined by the previously discovered patterns and knowledge by the CI clusters running the GKD process.

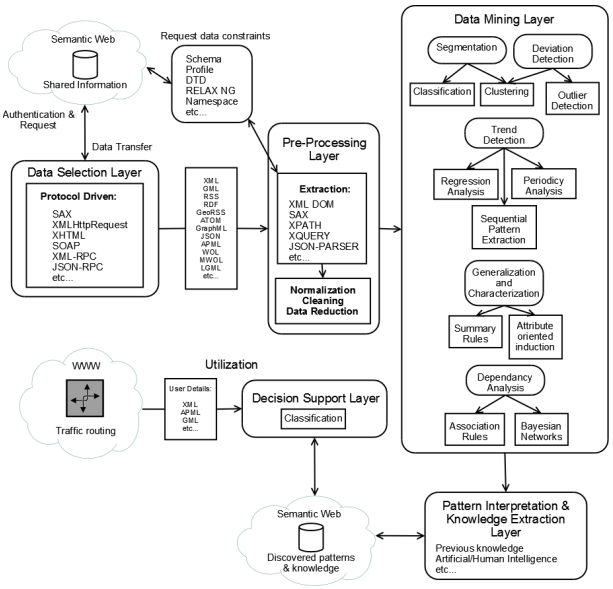


Figure 3: GKD framework for CI.

3.3 GKD for CI

Figure 3 describes GKD framework for CI. The GKD process produces useful patterns and knowledge from data retrieved from the Geoweb. The data does not just come from the collaborators, but also from publicly available linked semantic information via RDF and information collected via Web crawlers. The power behind patterns being discovered from diverse sources is that they represent global trends, as opposed to finding local trends from a singular source. The greater the diversity and number of sources - the greater the accuracy of user classification and decision support. The CI cluster is made up of multiple local CI servers with possible connections to external CI servers. This configuration is an ensemble method which aims to increase the accuracy of classification at the expense of increased complexity [10].

The first stage of the GKD process is the acquisition of data. The data is segmented and processed in various ways depending on the data mining model being used. What is common among all methods is the type of data which is available from the Semantic Web, which will be some form of XML or JSON data. When the WSCDL is used to setup collaboration, the Web service data formats and end-points are detailed so the CI server knows exactly what it is retrieving. Data constraints are determined by requesting the schema/profile information from the Web services. This allows the data to be extracted into a refined information space suitable for the data mining layer. The data mining layer can be one or a combination of models. New data mining models are constantly being discovered and refined and it is important that this layer be modular in order to easily adapt to changes.

Discovered patterns are analyzed and the models determined as novel and potentially useful are stored back onto the Semantic Web. The patterns can be stored in any format deemed suitable, but there are quite a number of advantages of producing results in a format which can be easily shared between applications, such as the PMML. This is especially important with external CI cluster collaboration when the discovered models must be compared in order to achieve greater accuracy. The discovered models represent current trends within analyzed data and are constantly changing. The advantage of determining trends from multiple sources is that certain trends will become prevalent in certain areas before other areas. Once a new trend is detected then collaborating Web services can take advantage of this information and maximize the potential profitable traffic to the trend. Web services can also submit user tracking information, such as APML-based information which can be used with the pre-determined models to classify the user to a certain group. This allows user specific traffic routing which again can greatly increase the chance of profitable traffic.

However, traffic routing need not only be Web based. Decision support can also make use of the increasing amounts of spatially reference information in order to determine real-world geospatial routing. If a user is classified as a particular group and their geospatial location is known then real-world profitable traffic can be achieved by suggesting Web services associated with the real-world elements of their group. An example of this would be to classify a particular user as a fan of a football club and to suggest products and the location of sports stores from Web services associated with that particular football club. Another possibility would be to offer current trend information related to the football club in order to deepen the relationship with the club which may eventually lead to profitable traffic. Another example scenario would be to determine that a user is interested in a particular food group and

to offer information and links to nearby restaurants of that food group. With the addition of spatial dimensions, recommendations to the user can take on new aspects and be represented as real locations on a Web map. This is the real power of using the GKD process to aid decision making. The collaborated data from the Semantic Web can then also be seen as a spatially indexed Geoweb which can be used for segmentation queries to determine potential profitable traffic for the classified user. The geospatial information can even play an important part when generating the models for classification; depending on various trend regions. The possibilities for profitable traffic from GKD from collaborating Web services is literally endless.

4 Case Studies

The following subsections give case studies using the GKD for CI process to demonstrate the potential usefulness of the system. In this study, we utilize the generalized Voronoi diagram for space tessellations and clustering for user profile segmentation. Datasets are retrieved from various mashups and visualized with mapping websites.

4.1 Restaurant recommendation case study

In this case study, let us assume a food and wine information Web service has been recording a user and building a profile using the APML. The Web server records the user searching for Château Pétrus information and sends the APML to the CI cluster for recommendations to suggest to the user. The APML contains the information related to the Château Pétrus in the RDF format and we are able to retrieve machine understandable information via its URI. With the aid of the Semantic Web, it determines that Château Pétrus is a beverage originating from France which is consumed by humans usually when dining out. The APML also contains location information relating to the user in the GML. The CI server processes these attributes with learned models to determine matches to products of the collaborating Web services that are near the user's location. It determines a number of wine distributers and restaurants which have the Pétrus in stock. Details of the distributers and restaurants which include stock number, price and location information are returned back to the Web service and used to generate the Web map as shown in Figure 4. To entice the user further, the average ratings and small snippets of reviews can be added to each location on the Web map. If the user follows any of the suggestions and either buys from a distributer or books at one of the restaurants, then the distributer or restaurant profit from the sale and the originating food and wine information Web service receives a portion from the total profit. This case shows how Geoweb and the Semantic Web are able to connect online interactions to result in real world transactions.

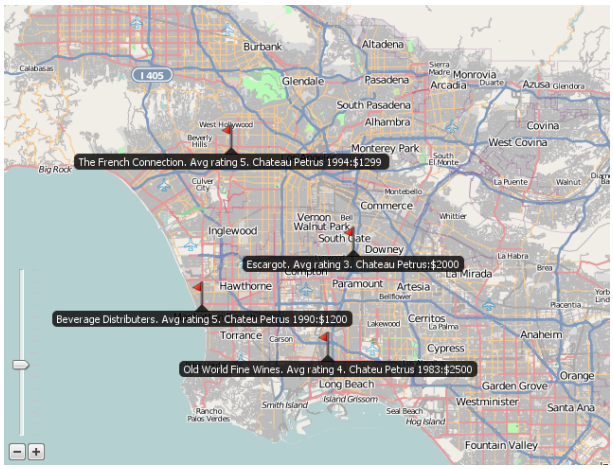


Figure 4: Recommended locations for Château Pétrus: including average rating and prices.

4.2 Recommending discovered trends

In this case study, let us assume that the CI clusters are updating their models as well as determining emerging, continuing, and fading trends from the current conditions on the collaborated Web services information. A trend is considered emerging when it has been newly detected for the current update period; continuing if it still remains from the previous period; and fading if the trend is no longer detected. The detection of trends can be useful for maximizing the possible profitable traffic to the Web services associated with current trends. The trends can be detected by the increase of sales for a particular product or even by digesting information from contributed user information to blogs, news and review Web services which might come from many various locations. Let us assume that the CI cluster have discovered an emerging trend of Brazilian coffee in the area of New York from a subset of collaborating Web services. The CI cluster finds associations of the collaborating Web service data with Brazilian coffee and updates the learned models. Now let us assume that a user is searching for classy coffee shops around New York. This tracked information is sent from a Web service to the CI cluster which determines, through association pattern mining, that classy coffee shops are linked to popular coffee. The CI server classifies popular coffee in New York as Brazilian coffee, which was pre-determined as an emerging trend. The CI server then searches the collaborated data for coffee shops with Brazilian coffee and returns a list of matches back to the original Web service which are nearest to New York, which can be generated as the Web map described in Figure 5. Using determined trends as recommendations can increase the probability of profitable traffic. The ability for GKD to determine trends for real locations from spatially referenced data and model them effectively to the user are the results of the Geoweb and Semantic Web information.

The known user location is rather vague and the exact location cannot be determined. To overcome this, the results

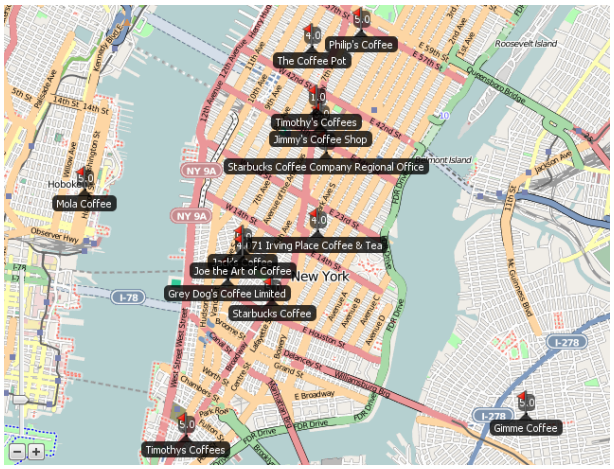
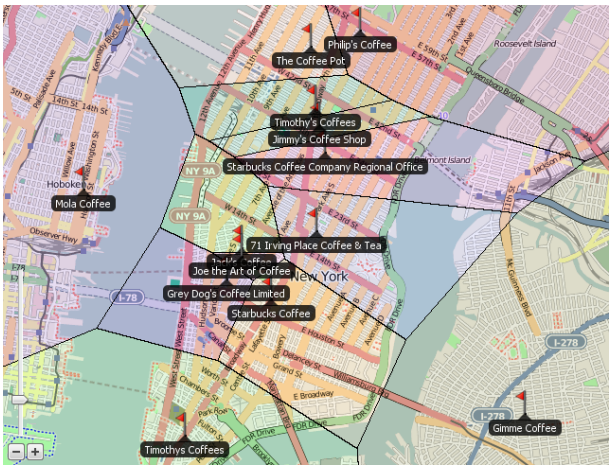
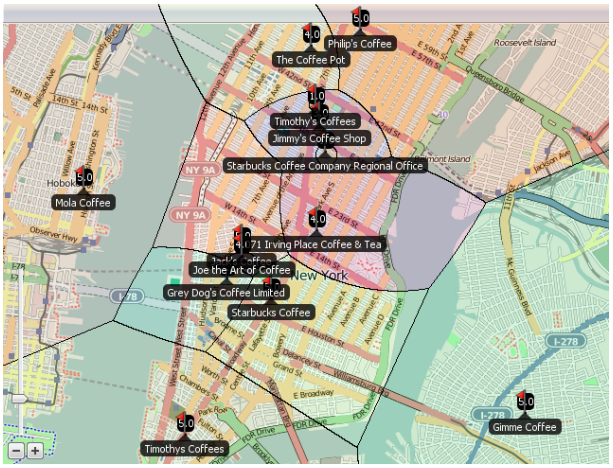


Figure 5: Recommended coffee shops with ratings.



(a)



(b)

Figure 6: Examples of recommendation links displayed as locations on a Web map: (a) Recommended coffee shops with segmentation; (b): Recommended coffee shops with weighted segmentation based on user ratings.

can undergo Web map segmentation; an effective visualization technique to help aid decision support. The results of this are shown in Figure 6 (a). Web map segmentation is not just limited to ordinary distriction but can also produce many various visualizations which could be used to further entice the user. Figure 6 (b) shows average user ratings used to create weighted regions associated with the coffee shops. The generalized Voronoi diagram has been used for space tessellations [15]. The user need only to determine which region they are located in to decide the closest highly rated coffee shop. Added spatial dimensions to information can be used to greatly enhance the depth of the information. This kind of specific information might just be what users require in order to engage them into profitable actions.

5 Final Remarks

In reality there could be many different number of the cases described in Section 4 because GKD from the Semantic Web produces extremely versatile patterns and models which can be used to determine potentially profitable traffic in a vast number of ways. In order for a true collaborative infrastructure to exist, Internet developers must try and implement their Web services using Web 2.0 technologies which conform to W3C and Open Geospatial Consortium (OGC) protocols to build the required foundation which collaborating Web services can exist on. At the moment, the Internet still consists of dominantly network as information content. There exists great potential for the increase of profitable traffic with the more collaboration that is achieved. However, there are still many components related to CI and GKD which can still be improved.

Because Web 2.0 content is forever changing and increasing, GKD techniques must be developed that can handle diverse data types which does not only consider the size of the data - but also the throughput which streaming information must be processed. A user interaction with a Web service may be near instantaneous - but the same is not necessarily so for the processing required to analyze and update current models produced by GKD. Better techniques into producing dynamically updating models under heavy streaming loads needs to be explored in the future. However, the processing time is not the only issue related to discovering models. Current GKD techniques are still relatively new and considered as an emerging research field. How can new techniques be made that can cope with the extremes of massive streaming Web data [7]?

Another problem which does not focus on the technical issues is the one regarding privacy [12]. Researchers need to ask themselves if discovering new knowledge about individuals is breaching ethical privacy. We cannot observe people going to work, seeing what they do, what they like to buy, how they invest their money, finding about their personal views without their permission. Is it okay to use this same kind of information about people that is distributed on the Internet? However, users who share private information about themselves allows target marketing with a higher accuracy - which can benefit the user because it allows them to get exactly what they want. In the future, users and Web services must be allowed the freedom to collaborate with-

out the fears of breaches of privacy and laws. This means collaboration technologies must be designed to easily allow collaboration while also circumventing fraudulent activity.

References

[1] Rdf primer. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/, Februar 2004. Stand: 15.4.2009.

[2] S. Alag. *Collective Intelligence in Action*. Manning Publications, 2008.

[3] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah McGuinness, Peter Patel-Schneijder, and Lynn Andrea Stein. OWL Web Ontology Language Reference. Recommendation, World Wide Web Consortium (W3C), February10 2004.

[4] Tim Berners-Lee. Giant global graph. Blog, 11 2007.

[5] Roberto Chinnici, Jean-Jacques Moreau, Arthur Ryman, and Sanjiva Weerawarana. Web services description language (wsdl) version 2.0 part 1: Core language. World Wide Web Consortium, Recommendation REC-wsdl20-20070626, June 2007.

[6] V. Estivill-Castro and I. Lee. Argument Free Clustering via Boundary Extraction for Massive Point-data Sets. *Computers, Environments and Urban Systems*, 26(4):315–334, 2002.

[7] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, C.A., 2000.

[8] Nickolas Kavantzas, David Burdett, Greg Ritzinger, Tony Fletcher, Yves Lafon, and Charlton Barreto. Web services choreography description language version 1.0. World Wide Web Consortium, Candidate Recommendation CR-ws-cdl-10-20051109, November 2005.

[9] Graham Klyne, Jeremy J. Carroll, and Brian McBride. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation, Feb 2004. Available at: http://www.w3.org/TR/rdf-concepts, last access on Dez 2008.

[10] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, New Jersey, 2004.

[11] I. Lee, K. Lee, and C. Torpelund-Bruin. Voronoi Image Segmentation and Its Application to Geoinformatics. *Journal of Computers*, 4(11):1101–1108, 2009.

[12] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. *Journal of Cryptology*, 15(3):177–206, 2002.

[13] H. J. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery: An Overview*. Cambridge University Press, Cambridge, UK, 2001.

[14] Harvey J. Miller. *Geographic Data Mining and Knowledge Discovery*. Handbook of Geographic Information Science, 2004.

[15] A. Okabe, B. N. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, West Sussex, second edition, 2000.

[16] Tim O'Reilly. What is web 2.0? design patterns and business models for the next generation of software., 2005.

[17] V. Podgorelec, L. Pavlic, and M. Hericko. Semantic Web Based Integration of Knowledge Resources for Supporting Collaboration. *Informatica*, 31(1):85–91, 2007.

[18] Stefan Raspl. Pmml version 3.0 - overview and status. In *KDD-2004 Workshop on Data Mining Standards, Services and Platforms (DM-SSP 04), KDD-2004 The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[19] I. Svetel and M. Pejanovic. The Role of the Semantic Web for Knowledge Management in the Construction Industry. *Informatica*, 34(3):331–336, 2010.

[20] C. Torpelund-Bruin and I. Lee. Geographic Knowledge Discovery from Geo-referenced Web 2.0. In *Proceedings of 2008 International Workshop on Geoscience and Remote Sensing*, pages 291–294, Shanghai, China, 2008. IEEE Computer Society.

[21] C. Torpelund-Bruin and I. Lee. When Generalized Voronoi Diagrams Meet GeoWeb for Emergency Management. In H. Chen, C. C. Yang, M. Chua, and S-H. Li, editors, *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics*, Lecture Notes in Computer Science 5477, pages 64–75, Bangkok, Thailand, 2009. Springer.

[22] H. Wang, X. Jiang, L-T. Chia, and A-H. Tan. Wikipedia2Onto. Building Concept Ontology Automatically, Experimenting with Web Image Retrieval. *Informatica*, 34(3):297–306, 2010.

Improving Amazon-like Review Systems by Considering the Credibility and Time-Decay of Public Reviews

Bo-Chun Wang
Department of Computer Science, University of Southern California, USA
E-mail: bochunwa@usc.edu

Wen-Yuan Zhu
Department of Computer Science, National Chiao Tung University, Taiwan
E-mail: wyzhu@cs.nctu.edu.tw

Ling-Jyh Chen
Institute of Information Science, Academia Sinica, Taiwan
E-mail: cclljj@iis.sinica.edu.tw

Keywords: review systems, credibility, time-decay, e-commerce

Received: February 12, 2011

In this study, we investigate the review system of Amazon.com, which is regarded as one of the most successful e-commerce websites in the world. We believe that the review results provided by Amazon's review system may not be representative of the advertised products because the system does not consider two essential factors, namely the credibility and the time-decay of public reviews. Using a dataset downloaded from Amazon.com, we demonstrate that although the credibility and time-decay issues are very common, they are not handled well by current public review systems. To address the situation, we propose a Review-credibility and Time-decay Based Ranking (RTBR) approach, which improves the Amazon review system by exploiting the credibility and time-decay of reviews posted by the public. We evaluate the proposed scheme against the current Amazon scheme. The results demonstrate that the RTBR scheme is superior to the Amazon scheme because it is more credible and it provides timely review results. Moreover, the scheme is simple and applicable to other Amazon-like review systems in which the reviews are time-stamped and can be evaluated by other users.

Povzetek: Članek predlaga izboljšavo Amazonovega sistema recenzij.

1 Introduction

`Amazon.com` is regarded as one of the most successful online vendors in the world. In addition to spearheading online retail sales of a variety of products (such as books, music CDs, DVDs, software, and consumer electronics) and providing various useful web services, `Amazon` features review and recommendation systems that provide candid comments and recommendations for customers. Recent surveys have reported that 50% of online shoppers spend at least ten minutes reading reviews before making a decision about a purchase, and 26% of online shoppers read reviews on `Amazon` prior to making a purchase [1].

Review systems have been implemented on a number of popular Web 2.0-based e-commerce websites (e.g., `Amazon`[1] and `eBay`[2]), product comparison websites (e.g., `BizRate`[3] and `Epinions`[4]), and news websites (e.g., `MSNBC`[5] and `SlashDot`[6]). Generally, a review system is a kind of reputation system that facilitates the development of trust in Internet interactions [24]. Unlike recommendation systems, which seek to personalize each user's web experience by exploiting item-to-item and user-to-user correlations [20, 26], review systems give an average rating for an item based on other customers' opinions about the item.

`Amazon.com` allows users to submit their reviews to the web page of each product, and the reviews can be accessed by all users. Each review consists of the reviewer's name (either the real name or a nickname), several lines of comment, a rating score (ranging from one to five stars), and the timestamp of the review. All reviews are archived

[0] A preliminary version of this study was published in the Proceedings of International Workshop on Web Personalization, Reputation and Recommender Systems, 2008 [28]. In this extended version paper, we adopt the *Shallow Syntactic Features (ShallowSyn)* method [33] to estimate the credibility of public reviews that have not received a sufficient number of reviews. Moreover, we apply the *Kendall* test [16] to evaluate the correlation between the ranking results of our approach and the Amazon review system. We evaluate our proposed approach using a rich set of datasets, and discuss the deployment issue of the proposed approach in real systems. Hence, this manuscript is a much more thorough and authoritative presentation of our study on the Amazon review system.

[1] Amazon. http://www.amazon.com/
[2] eBay. http://www.ebay.com/
[3] BizRate. http://www.bizrate.com/
[4] Epinions. http://www.epinions.com/
[5] MSNBC News. http://www.msnba.msn.com/
[6] SlashDot. http://slashdot.org/

in the system, and the aggregated result, derived by averaging all the received ratings, is reported on the web page of each product. It has been shown that such reviews provide basic ideas about the popularity and dependability of the corresponding items; hence they have a substantial impact on cybershoppers' behavior [6, 9].

However, since the Amazon review system is an open forum, the anonymity of web reviewers increases the chances of abuse, such as unfair/biased ratings, ballot stuffing, and bad mouthing [7]. As a result, the review results may be misleading and untrustworthy [15, 21]. To mitigate the problem, Amazon incorporates a feature that allows users to evaluate other users' product reviews by stating whether they think a review is useful or not; however, the *discriminating capability* of the Amazon review system is generally considered limited because 1) the review results have the tendency to be skewed toward high scores [6]; 2) the *aging* issue of the reviews is not considered [32]; and 3) it has no means to assess the reviews' helpfulness if the reviews are not evaluated by a sufficiently large number of users (unless additional machine learning techniques could be applied [17, 33]).

To improve the discriminating capability of the Amazon review system, we propose a *Review-credibility and Time-decay Based Ranking* (RTBR) approach. Specifically, RTBR enhances the Amazon system by exploiting the *credibility* and *time-decay* of public reviews. Using data downloaded from the bookstore department of Amazon.com, we compare the proposed scheme with the current Amazon scheme, and show that it is superior because it is more credible and provides timely ranking results in all test cases. Moreover, the proposed scheme is simple and applicable to other Amazon-like rating systems, as long as each product's review is time-stamped and it can be evaluated by other users.

The remainder of this paper is organized as follows. In Section 2, we discuss related works on review systems. In Section 3, we present the proposed RTBR approach. In Section 4, using the real data downloaded from Amazon.com, we compare the proposed scheme with the current Amazon scheme and analyze the results. We also discuss the feasibility issue of the proposed scheme. We then summarize our conclusions in Section 5.

2 Related Work

A review system provides an average rating for each item based on other users' opinions of the item; hence, it is a kind of reputation system that facilitates the development of trust in Internet interactions [4, 8, 24, 30]. Review systems are used by a number of Web 2.0 sites (such as Amazon, BizRate, eBay, Epinions, and SlashDot). Though the systems differ in how they aggregate users' opinions and present the results, recent studies have shown that such systems have a strong impact on cybershoppers' purchase decisions [1, 6, 9].

Amazon and eBay, two of the most successful Web 2.0 e-commerce stores, pioneered the use of review systems by aggregating user-contributed content. On the eBay website, buyers and sellers are allowed to post reviews about each other after a transaction has been completed. A review can be positive (1), neutral (0), or negative (-1). The system aggregates the reviews of each user by summing all of his/her received ratings, and details the results on the user's profile page. Resnick et al. [23, 25] evaluated eBay's review system via controlled experiments and empirical analysis. In [23], they found more than half users were willing to provide feedback and it was almost all positive. In addition, [23] suggested that the users may reciprocate and retaliate. [25] found that eBay's reputation system had significant effect in the market, and sellers who had high reputation scores would sell their goods with higher prices. [15, 21] suggested that the eBay review system is likely to mislead users because it lacks a discriminating capability (for instance, the eBay review system has difficulty distinguishing between a user who receives 50 positive reviews and a user who receives 100 positives and 50 negatives, as the aggregated ratings of the two users are equal to +50). Moreover, it has been observed that ballot stuffing is common in the eBay review system; hence, this issue also needs to be resolved [3].

In contrast to eBay, the Amazon review system aggregates users' rating scores by averaging, instead of summing. As mentioned in the previous section, the Amazon review system allows users to submit their reviews to the web page of each product. It has been shown that the results of the Amazon review system are highly correlated to the prices of the corresponding products [6], and about 25% of online shoppers read reviews on Amazon before they make a purchase [1]. However, the shortcomings of the system are that it does not consider the aging issue of the reviews [32], and the review results are generally skewed toward high scores [6]. In addition, [11] shows the average score of 53% of the products does not reveal the true quality of product and may provide misleading recommendations. As a result, the discriminating capability of the Amazon review system is limited.

In addition to summing and averaging approaches, a number of other schemes have been proposed to improve the discriminating capability of review systems [13, 14, 22, 27, 29, 31]. For instance, [14, 22] propose Bayesian-based review systems that rate each product according to the feedback received. Specifically, each item of feedback is given either a positive (+1) or a negative (-1) rating. The Bayesian-based systems have been extended to filter out bad mouthing reviews [31]. However, the disadvantage of the Bayesian-based system is that it can not provide ratings with graded levels because it is a binomial model. Therefore, [13] proposes a generalization of the Bayesian-based systems, called *Dirichlet reputation systems*, which can support multiple value ratings. Finally, [27, 29] propose personalizing review results based on the *Personalized Similarity Measure* and users' preferences. However,

these approaches are rarely implemented in reality because the computation and storage overheads are prohibitive.

Since '*helpful*' reviews have stronger impacts on consumers' purchase decisions than other reviews [5], several studies have investigated how to assess reviews' helpfulness recently [17, 33]. For instance, [33] presents a *utility scoring* approach that computes three features of a given product review (namely the *Lexical Similarity Features*, *Shallow Syntactic Features*, and *Lexical Subjectivity Clues*) and then feeds the calculation results into a regression algorithm to measure the review's helpfulness. Similarly, [17] assesses review helpfulness using a SVM-based regression approach that considers five types of features, namely *Structural*, *Lexical*, *Syntactic*, *Semantic*, and *Metadata*.

Finally, [12] focuses on the analysis and detection of review spam. In [12], Jindal and Liu use a supervised learning and classification model to detect three types of review spam: *False Opinions*, *Reviews on brands only*, and *Non-reviews*. In [10], Hu and Liu propose a data mining and natural language processing based approach to facilitate mining and summarizing product reviews from a large number of customer reviews of a particular product. Since it is difficult and tedious for consumers to read hundreds or thousands of reviews for each product, the feature-based summary results provide consumers more concise information for purchase decisions.

3 The Proposed Approach: RTBR

In this section, we present the proposed review system, called *Review-credibility and Time-decay Based Ranking* (RTBR), for emerging Web 2.0-based applications. Unlike the current `Amazon` review system, the proposed scheme is expected to better represent the public's opinions about reviewed items, because it considers two additional factors of each review in the system, namely, 1) the quality of being convincing and believable, i.e., the *review-credibility* factor; and 2) the timeliness of being representative, i.e., the *time-decay* factor [7].

More precisely, we assume that there are n items in the system, and the i-th item has been reviewed r_i times. Let N_i denote the i-th item, $s_{i,j}$ denote the j-th rating score of N_i, and $t_{i,j}$ denote the length of time since $s_{i,j}$ was rated. For the j-th review of N_i, we define the review-credibility factor as $\omega_{i,j}$ and the time-decay factor as $\phi_{i,j}$, which we will discuss further in the following subsection.

Then, the proposed RTBR scheme calculates the score value of N_i (i.e., $\mathcal{S}_i$) by combining the review-credibility factor, the time-decay factor, and the review score of the received r_i reviews, as shown in Equation 1.

$$\mathcal{S}_i = \frac{\sum_{j=1}^{r_i} \omega_{i,j}\phi_{i,j}s_{i,j}}{\sum_{j=1}^{r_i} \omega_{i,j}\phi_{i,j}}. \quad (1)$$

Suppose $\Delta(\mathcal{S}_i, \mathcal{S}_j)$ is a comparison function that returns 1 when $\mathcal{S}_i \geq \mathcal{S}_j$, and it returns 0 otherwise. The RTSB scheme then reports the ranking of N_i by taking the complementary cumulative distribution function (CCDF) of $\mathcal{S}_i$. As shown in Equation 2, the ranking result indicates that N_i is in the top $\mathcal{R}_i^{RTBR}$ of all the compared products.

$$\mathcal{R}_i^{RTBR} = 1 - \frac{\sum_{j=1}^{n} \Delta(\mathcal{S}_i, \mathcal{S}_j)}{n}. \quad (2)$$

Note that, in the `Amazon` review system, the score value of N_i is obtained by averaging the received rating scores of r_i reviews (i.e., $\overline{\mathcal{S}}_i$), as shown in Equation 3, and the ranking results are derived in a similar manner to Equation 2, except that $\Delta(\mathcal{S}_i, \mathcal{S}_j)$ is replaced by $\Delta(\overline{\mathcal{S}}_i, \overline{\mathcal{S}}_j)$, as shown in Equation 4.

$$\overline{\mathcal{S}}_i = \frac{\sum_{j=1}^{r_i} s_{i,j}}{r_i}. \quad (3)$$

$$\mathcal{R}_i^{Amazon} = 1 - \frac{\sum_{j=1}^{n} \Delta(\overline{\mathcal{S}}_i, \overline{\mathcal{S}}_j)}{n}. \quad (4)$$

3.1 The review-credibility factor

As each product review may also be reviewed by other users, we use $k_{i,j}$ to denote the number of users that have reviewed the j-th review of N_i (i.e., $s_{i,j}$), and $u_{i,j}$ to denote the number of users (out of $k_{i,j}$) that think $s_{i,j}$ is useful. There are two cases for the definition of the review-credibility factor ($\omega_{i,j}$) for $s_{i,j}$, as shown in Equation 5:

- *Case 1:* If the j-th review of N_i has been reviewed by a sufficient number of people, i.e., $k_{i,j} \geq \gamma$, we define $\omega_{i,j}$ as the ratio of $u_{i,j}$ to $k_{i,j}$.
- *Case 2:* If $k_{i,j} < \gamma$, there may be a strong bias in the $k_{i,j}$ reviews when $k_{i,j}$ is small, or the value of $\frac{u_{i,j}}{k_{i,j}}$ cannot be calculated when $k_{i,j} = 0$. Thus, we define the value of $\omega_{i,j}$ using the *Shallow Syntactic Features* (ShallowSyn) method [33].

$$\omega_{i,j} = \begin{cases} \frac{u_{i,j}}{k_{i,j}} & , \text{if } k_{i,j} \geq \gamma \\ ShallowSyn_{\text{(the j-th review of } N_i\text{)}} & , \text{if } k_{i,j} < \gamma \end{cases} \quad (5)$$

Specifically, the *ShallowSyn* method employs the Support Vector Machine (SVM) approach to estimate the review-credibility of a give product review. The training dataset contains a sufficiently large number of reviews that have been reviewed by at least γ users. Using the training dataset, a SVM model is built in the off-line phase by considering a variety of features of each review, including the number of words, the number of sentences, and the number of the words of *shallow syntactic features* (i.e., proper nouns, numbers, modal verbs, interjections,

[7] A product may become popular (e.g., due to advertising, promotion, or marketing) or outdated (e.g., due to the release of a newer version) over time. Note that the time-decay factor should be weighted in accordance with the properties of the product types. For instance, it should be weighted higher for reviews of electronic products than that of books. We defer a detailed discussion and evaluation of this issue to a future work.

Algorithm 4 The algorithm for determining the value of the aging factor, λ, in the RBTR scheme.

```
1: Function Aging_Factor
2: i ⇐ −1 ; α1 ⇐ 1 − 10^i ; δ1 ⇐ ϒ(α1)
3: while true do
4:    α2 ⇐ 1 − 10^(i−1) ; δ2 ⇐ ϒ(α2)
5:    if |δ1−δ2|/δ1 ≤ 0.1 then
6:       return α1
7:    end if
8:    α1 ⇐ α2 ; δ1 ⇐ δ2 ; i ⇐ i − 1
9: end while
```

comparative and superlative adjectives, comparative and superlative adverbs, wh-determiners/adverbs/pronouns and possessive wh-pronouns). Then, the SVM model is used to estimate the review-credibility of the reviews that are reviewed by less than γ users.

3.2 The time-decay factor

For the j-th review of N_i, we define the time-decay factor ($\phi_{i,j}$) by

$$\phi_{i,j} = \lambda^{t_{i,j}}, \tag{6}$$

where λ is an aging factor ($0 < \lambda \leq 1$). The value of λ is calculated using the decision algorithm, as shown in Algorithm 4, where $\Upsilon(\alpha)$ is the comparison function that returns the average ranking distance of all the items when the value of λ is set to α in the RTBR scheme, i.e., the mean of $|\mathcal{R}_i^{RTBR} - \mathcal{R}_i^{Amazon}|$ for all i. Note that, as shown in Equation 6. the smaller the value of λ, the more emphasis we put on the time-decay factor of public reviews. Since each type of item has different sensitivity to the time-decay of reviews, the algorithm tries to determine the value of λ that will ensure the results of the proposed RTBR scheme are more representative and timely.

4 Evaluation

In this section, we evaluate the proposed RTBR scheme and compare it with the current `Amazon` review system. We present the properties of the dataset downloaded from the bookstore department of `Amazon.com` in subsection 4.1, and show the evaluation results in subsection 4.2. Moreover, we discuss the feasibility and implementation issues of the proposed scheme in subsection 4.3.

4.1 Data collection and analysis

We wrote a crawler program to download data from the bookstore department of Amazon.com at the end of June 2011[8]. The downloaded data relates to books tagged as `Animation`, `Autobiography`, `Business`, `Documentary`, `Programming`, or `Psychology`. For each book, the collected data contains the book's title, the author's name, and the reviews received. Moreover, each review contains the rating score, the reviewer's name, the timestamp, the number of times the book has been evaluated, and the number of evaluations that deemed it useful. For simplicity, in this study, we only consider the books that have received more than five reviews. The dataset contains 25,167 books and 1,644,871 reviews. Table 1 lists the properties of the dataset.

Table 1: The properties of the dataset downloaded from the bookstore department of Amazon.com

Tag	No. of products	Avg. no. of reviews
Animation	4,511	86
Autobiography	2,410	92
Business	6,780	52
Documentary	3,331	52
Programming	2,355	34
Psychology	5,780	63

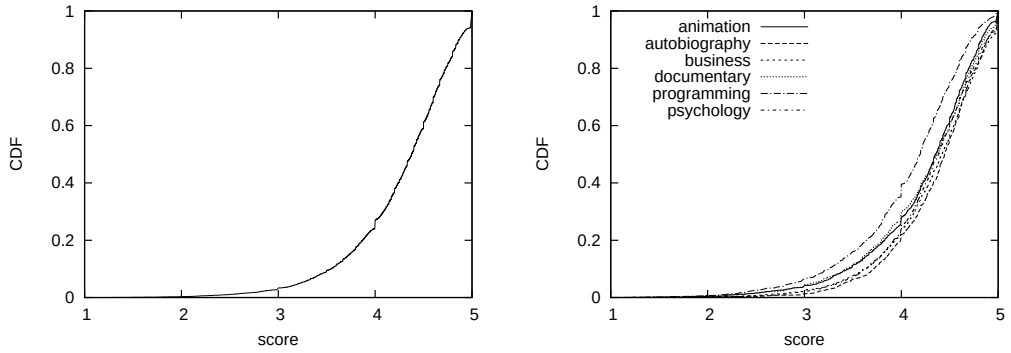
Like the `Amazon` review system, we first calculate the mean of the rating scores for each book in the dataset. Then, we plot the mean score distribution on cumulative distribution function (CDF) curves, as shown in Figure 1. We find that 70% of the books have a mean score higher than 4, and only 5% have a mean score lower than 3. The results confirm the findings of previous studies that the mean score distribution on the `Amazon` website is skewed towards higher scores [6, 28]. Thus, the current Amazon review system cannot be considered representative because it lacks a discriminating capability.

Next, following [33], we set the value of γ to 10, and used Equation 5 to calculate the credibility of each review. We used the *Stanford Parser* [18, 19] to parse every review and compute its features. Then, for each category of books, we collected the features from the reviews that have been reviewed by at least 10 users as the training dataset, and applied the *ϵ-Support Vector Regression* (ϵ-SVR) implemented in LIBSVM [2] to build the SVM models[9], which were used to estimate the credibility values of the reviews which are reviewed by less than 10 users. Table 2 lists the properties of the training dataset and the regression performance (in terms of the mean squared error σ^2). All results are based on 10-fold cross validation.

Figure 2 shows the CDF distribution of the credibility scores of the downloaded reviews. We observe that about 24% of the reviews are not credible (i.e., the credibility value is less than 0.5), and only 12% of the reviews are highly credible (i.e., the credibility value is higher than 0.8). It seems that a substantial number of reviews on the

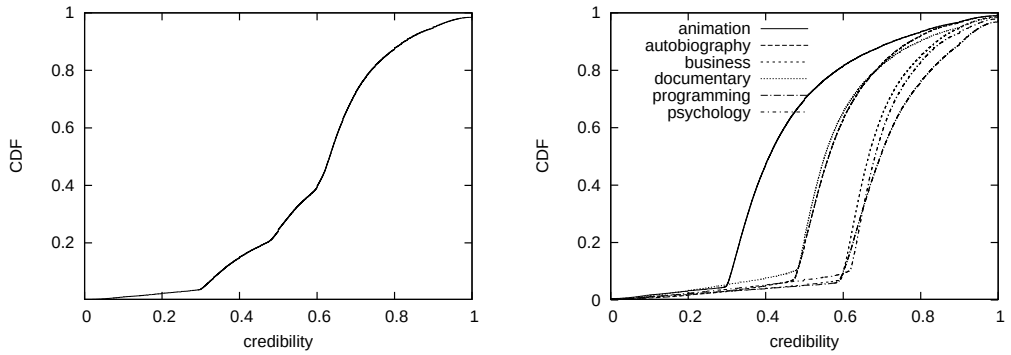
[8]We note that the proposed approach is also applicable to the other product reviews on Amazon.com. However, we do not include the evaluation using the other products on Amazon.com in this study because the number of the items differs greatly among different product categories, and the number of the reviews per item varies a lot even within the same product type. We defer a detailed discussion of this issue to a future work.

[9]To be accurate, the SVM models should be updated periodically, for each category of books, after real-world deployment.



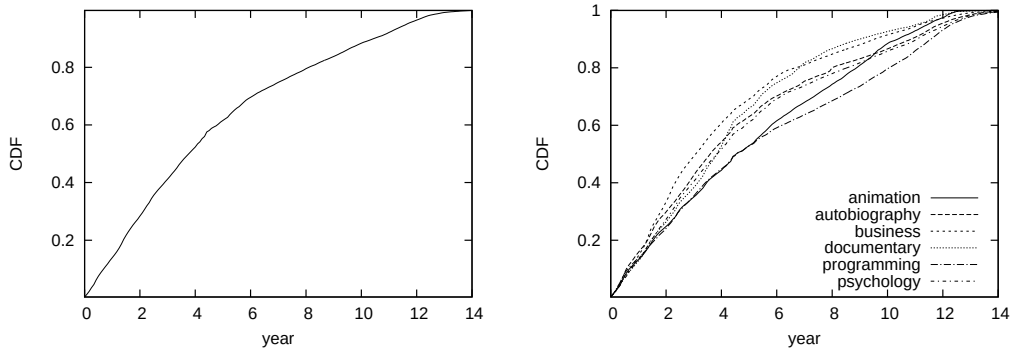
(a) distribution of the whole dataset (b) distribution of each tag of books

Figure 1: The CDF distribution of the mean scores of the downloaded Amazon dataset.



(a) distribution of the whole dataset (b) distribution of each tag of books

Figure 2: The CDF distribution of the review credibility of the downloaded Amazon dataset.



(a) distribution of the whole dataset (b) distribution of each tag of books

Figure 3: The CDF distribution of the ages of reviews in the downloaded Amazon dataset.

Table 2: No. of training data and mean squared error

Tag	No. of training data	σ^2
Animation	53,850	0.0899
Autobiography	44,799	0.0618
Business	84,063	0.0708
Documentary	45,424	0.0775
Programming	17,156	0.0734
Psychology	95,050	0.0741

Amazon website are either unreliable (e.g., due to individual preferences or unintentional bias) or malicious (e.g., due to ballot stuffing, bad mouthing, or intentional bias). Again, the results confirm the findings of previous studies as reported in [7, 15, 21, 28].

Moreover, in Figure 3, we plot the CDF distribution of the ages of the 1,644,871 downloaded reviews (i.e., the time since each review was posted until the data was collected). Interestingly, only 13% of the reviews were posted within the previous year, whereas more than 50% were posted at least four years earlier. The results confirm that the aging of reviews is a significant issue [28, 32]. For instance, over time, a book may become popular (i.e., due to advertising, promotion, or marketing) or outdated (i.e., due

Table 3: Kendall's rank correlation (τ)

Tag	Kendall's rank correlation (τ)
Animation	0.8507614
Autobiography	0.8617765
Business	0.8876215
Documentary	0.8533894
Programming	0.8666127
Psychology	0.8351172

to the release of a newer version of the product). Hence, the aging factor must be carefully managed in order to improve the discriminating capability of the review system.

4.2 Evaluation of the RTBR scheme and the current Amazon review system

Next, we compare the review results of the proposed RTBR approach and those of the Amazon approach (i.e., by taking the mean of all the received rating scores) using the Kendall test [16]. Table 3 shows the Kendall's rank correlation (τ) results. The Kendall's rank correlation (τ) is effective in evaluating the degree of similarity between two rankings given to the same set of objects. From the evaluation results, we observe that the ranking results of the two approaches have a high correspondence, which is encouraging since the goal of the RTBR approach is to adjust the Amazon approach by considering the credibility and time-decay factor, it should not change the order of ranking results of the Amazon approach drastically.

Then, using the dataset downloaded from the Amazon website, Figure 4 shows the comparison results, where each point represents a product with its corresponding ranking (as a percentage) using the RTBR scheme and the Amazon scheme. Each sub-figure is divided into three areas: 1) area I contains *over-estimated* products (i.e., the review results of the RTBR scheme are far lower than those of the Amazon scheme); 2) area II contains consistently estimated products (i.e., the review results derived by the RTBR scheme and the ordinary Amazon scheme are within $\pm 5\%$ of each other); and 3) area III contains *under-estimated* products (i.e., the review results of the RTBR scheme are far higher than those derived by the Amazon scheme). Table 4 shows an example of two programming books[10] that are over-estimated and under-estimated respectively.

The results in Table 4 show that the RTBR scheme can improve the Amazon approach because it considers the creditability and time-decay factors. Specifically, in Table 4, *Book A* is considered over-estimated under the Amazon approach because most of the high-score reviews are either outdated (e.g., all 5-star reviews were made in 2007) or not creditable (e.g., one of the 5-start reviews were regarded not helpful by 34 reviewers). In contrast, *Book B* is regarded under-estimated under the Amazon approach, as its 5-star reviews are more creditable than 4-star reviews (e.g., one of the 4-star reviews was accepted by one of 27 reviewers).

[10]Book A: *The Book of Qt 4: The Art of Building Qt Applications*, ISBN: 1593271476, http://www.amazon.com/product-reviews/1593271476; Book B: *Object-Oriented Programming in C++*, ISBN: 0470843993, http://www.amazon.com/product-reviews/0470843993. (Accessed on September 1, 2011)

Table 5: The distribution of the comparison results for the downloaded Amazon dataset.

Tag	Area I	Area II	Area III
Animation	50.14%	43.92%	5.94%
Autobiography	57.93%	37.26%	4.81%
Business	63.60%	33.21%	3.19%
Documentary	37.98%	49.53%	12.49%
Programming	33.89%	55.41%	10.70%
Psychology	58.30%	34.59%	7.11%

We summarize the distribution of the comparison results for the six categories of books in Table 5, and the results show that only less than a half of the products have consistent review results in both schemes, while the others are dominated by over-estimation and next by under-estimation. Moreover, the results are consistent with our previous findings, which are based on the dataset of the same categories collected in April 2008. To investigate the causes of the inconsistent results, we design two tests, namely a *credibility test* and a *time-decay test*.

1. *Credibility Test*

   This test is designed to determine whether the inconsistency between the RTBR and Amazon schemes is caused by the credibility of reviews. Suppose that $\delta(s_{i,j}, x)$ is a comparison function that returns 1 if $s_{i,j} = x$, and 0 otherwise, as shown in Equation 7. For the i-th product N_i, we calculate the credibility factor $D_c(i, x)$ of each score value x using Equation 8. Then, we apply linear regression to analyze the relationship between x and $D_c(i, x)$, and obtain the slope $L_c(i)$ of the regression line. Based on the value of $L_c(i)$, the *Credibility Test* reports *TRUE* (i.e., the inconsistency is caused by review credibility) if 1) $L_c(i) < 0$ and the corresponding point of N_i is in the Area I, or 2) $L_c(i) > 0$ and the corresponding point of N_i is in area III; and it reports *FALSE* otherwise.

$$\delta(s_{i,j}, x) = \begin{cases} 1 & \text{, if } s_{i,j} = x \\ 0 & \text{, if } s_{i,j} \neq x \end{cases} \tag{7}$$

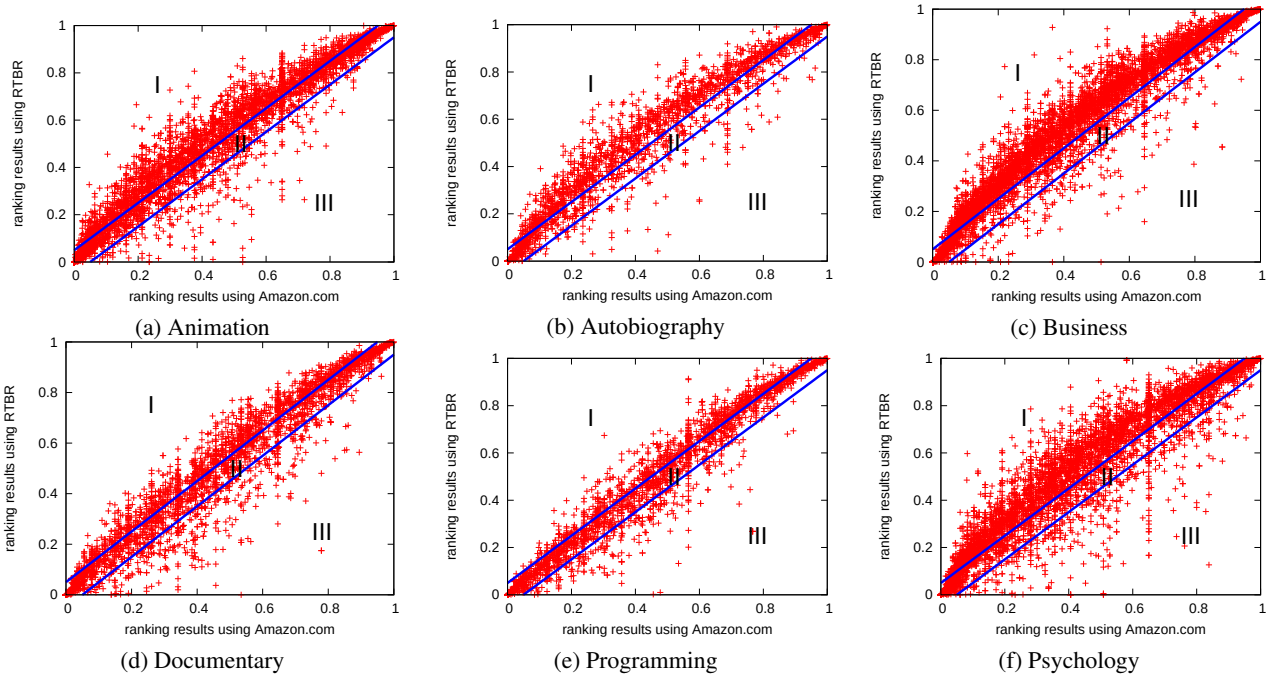
$$D_c(i, x) = \frac{\sum_{j=1}^{r_i} \omega_{i,j} \delta(s_{i,j}, x)}{\sum_{j=1}^{r_i} \omega_{i,j}} - \frac{\sum_{j=1}^{r_i} \delta(s_{i,j}, x)}{r_i} \tag{8}$$

2. *Time-decay Test*

   This test attempts to determine whether the inconsistency between the RTBR and Amazon schemes is caused by the time-decay of the reviews. We denote $t_{i,max}$ and $t_{i,min}$ as the maximum and minimum values of $t_{i,j}$ for $1 \leq j \leq r_i$ respectively. We divide

Table 4: The example of two programming books on Amazon.com that are over-estimated and under-estimated respectively

Book A				Book B			
Score	Data	Review Title	Ratio of people felt helpful	Score	Data	Review Title	Ratio of people felt helpful
5	2007/09/02	"Best book on QT 4"	12 of 16	4	2003/01/17	"Good OOP Book"	8 of 11
5	2007/10/14	"Arrived in good order"	0 of 34	5	2003/03/25	"Extremely well written and ENJOYABLE Book"	4 of 4
5	2007/12/04	"It's an excellent guide for any QT programmer"	6 of 7	5	2003/05/17	"Which C++ Book To Read First?"	39 of 40
3	2008/07/02	"A Mixed Bag"	14 of 14	5	2003/09/16	"C++ enthusiast"	9 of 9
3	2009/03/21	"UI files are incompatible with Qt 4.5"	3 of 3	5	2003/12/13	"Pure C++ Tutorial"	6 of 6
4	2009/04/02	"Very Good Book"	1 of 2	4	2004/02/15	"GOOD BOOK, BUT..."	1 of 27
4	2009/10/15	"Pretty Good"	1 of 1	5	2008/01/11	"Good Start Point for Professionals"	2 of 2
4	2009/11/26	"Full of useful informaton"	0 of 0	5	2008/07/09	"Well written, good examples"	0 of 0
Ranking by Amazon.com: 50% Ranking by RTBR: 68%				Ranking by Amazon.com: 7% Ranking by RTBR: 2%			



(a) Animation (b) Autobiography (c) Business

(d) Documentary (e) Programming (f) Psychology

Figure 4: Comparison of the review results derived by the proposed RTBR scheme and the Amazon scheme on the downloaded dataset. The sample points in areas I, II, and III are considered to be overestimated, consistent (within $\pm 5\%$ error), and underestimated respectively.

the period between $t_{i,min}$ and $t_{i,max}$ into Y equal intervals (for simplicity, Y is fixed at 10 in this study), and assume that $\sigma(t_{i,j}, y)$ is equal to 1 when $t_{i,j}$ falls in the y-th interval, as shown in Equation 9. For the i-th product N_i, we calculate its time-decay factor $D_t(i, y)$ for each time interval y using Equation 10. More specifically, in Equation 10, $\Delta(s_{i,j}, \overline{s}_i)$ is equal to 1 when $s_{i,j} > \overline{s}_i$, and 0 otherwise. Then, we apply linear regression to analyze the relationship between y and $D_t(i, y)$, and obtain the slope $L_t(i)$ of the regression line. Based on the value of $L_t(i)$, the *Time-decay Test* reports *TRUE* (i.e., the inconsistency is due to the time-decay of the reviews) if 1) $L_t(i) > 0$ and the corresponding point of N_i is in area I, or 2) $L_t(i) < 0$ and the corresponding point of N_i is in area III; and it reports *FALSE* otherwise.

$$\sigma(t_{i,j}, y) = \begin{cases} 1 & \text{, if } t_{i,min} + \frac{(y-1)\times(t_{i,max}-t_{i,min})}{Y} \\ & \leq t_{i,j} < t_{i,min} + \frac{y\times(t_{i,max}-t_{i,min})}{Y} \\ 0 & \text{, otherwise} \end{cases} \tag{9}$$

$$D_t(i, y) = \frac{\sum_{j=1}^{r_i} \sigma(t_{i,j}, y)\Delta(s_{i,j}, \overline{s}_i)}{\sum_{j=1}^{r_i} \sigma(t_{i,j}, y)} \tag{10}$$

We examine the items that fall in area I and III using the two test approaches, and summarize the results (i.e., whether they are caused by the credibility or time-decay factors, or a combination of the two) in Table 6. From the

Table 6: The evaluation results of the causes of under-estimations and over-estimations using the designed credibility test and time-decay test.

Subject	Area	Credibility	Time-decay	Union
Animation	I	63.31%	56.85%	91.15%
	III	98.88%	66.67%	99.63%
Autobiography	I	36.97%	67.91%	82.34%
	III	96.55%	65.52%	100.00%
Business	I	41.95%	70.67%	86.10%
	III	97.69%	67.59%	100.00%
Documentary	I	51.11%	68.81%	91.19%
	III	99.04%	52.16%	100.00%
Programming	I	58.34%	76.91%	94.35%
	III	98.41%	55.16%	100.00%
Psychology	I	30.77%	72.31%	83.35%
	III	98.78%	63.50%	100.00%

results, we observe that most of the inconsistency is caused by the credibility of reviews. Moreover, we observe that the credibility issue tends to cause more *under-estimations*, while the time-decay issue causes more *over-estimations*. We also find that, by combining the credibility and time-decay tests, more than 82% of the inconsistency can be classified. Since the RTBR approach considers the credibility and time-decay issues, it is superior to the Amazon approach because it provides more representative review results.

4.3 Discussion

In this subsection, we discuss the implementation issue of the proposed approach, and we demonstrate that the score values can be updated in an *incremental* manner in the proposed approach, thereby reducing greatly the computational complexity in real systems.

More specifically, we let N_i denote the i-th item, $s_{i,j}$ denote the j-th rating score of N_i, r_i denote the number of users that have reviewed N_i, and $\mathcal{S}_i$ denote the score of N_i at time T. In addition, for the j-th review of N_i, $\omega_{i,j}$ denotes the review-credibility factor, and $\phi_{i,j}$ denotes the time-decay factor. The system has to calculate the numerator ($\mathcal{A}_i = \sum_{j=1}^{r_i} \omega_{i,j}\phi_{i,j}s_{i,j}$) and the denominator ($\mathcal{B}_i = \sum_{j=1}^{r_i} \omega_{i,j}\phi_{i,j}$) respectively in order to obtain the value of $\mathcal{S}_i$ (cf. Equation 1), and there are two cases to update the value of $\mathcal{S}_i$:

1. *Case 1: a new review for N_i is input at time T'*

   In this case, we first obtain the review-credibility factor, ω_{i,r_i+1}, and the time-decay factor, ϕ_{i,r_i+1}, using Equations 5 and 6 respectively. Then, the system will update the values of $\mathcal{A}_i$ and $\mathcal{B}_i$ using Equations 11 and 12 (i.e., consider the time decay of the previous values of $\mathcal{A}_i$ and $\mathcal{B}_i$ by multiplying $\lambda^{T'-T}$, and plus the input of the new $r_i + 1$ *th* review), and derive the updated score value $\mathcal{S}'$ using Equation 13.

$$\mathcal{A}'_i = \mathcal{A}_i\lambda^{T'-T} + \omega_{i,r_i+1}\phi_{i,r_i+1}s_{i,r_i+1} \quad (11)$$

$$\mathcal{B}'_i = \mathcal{B}_i\lambda^{T'-T} + \omega_{i,r_i+1}\phi_{i,r_i+1} \quad (12)$$

$$\mathcal{S}'_i = \frac{\mathcal{A}'_i}{\mathcal{B}'_i} \quad (13)$$

2. *Case 2: the j-th review of N_i is changed at time T'*

   In this case, we obtain the the new review-credibility factor of the j-th review, ω'_{i,r_j} by Equation 5, and update the values of $\mathcal{A}_i$ and $\mathcal{B}_i$ using Equations 14 and 15 respectively (i.e., plus the offset caused by the update of the j-th review). Then, we obtain the updated score value $\mathcal{S}'_i$ using Equation 13.

$$\mathcal{A}'_i = \mathcal{A}_i + (\omega'_{i,j} - \omega_{i,j})\phi_{i,j}s_{i,j} \quad (14)$$

$$\mathcal{B}'_i = \mathcal{B}_i + (\omega'_{i,j} - \omega_{i,j})\phi_{i,j} \quad (15)$$

As we can see in the above two cases, the score values of each item can be updated in an incremental manner. Hence, the computational complexity of the proposed approach is moderate, and the proposed approach is feasible to be implemented in real systems.

5 Conclusion

In this paper, we have discussed the review system of `Amazon.com`, one of the most popular online vendors in the world. We argue that the results published by the `Amazon` review system are not representative because they do not consider two essential factors, namely the credibility and time-decay of reviews submitted by the public. To address this issue, we propose the *Review-credibility and Time-decay Based Ranking* (RTBR) scheme. Using a dataset downloaded from the bookstore department of `Amazon.com`, we compare the proposed scheme with the current `Amazon` scheme, and demonstrate that the proposed scheme is superior because it is more credible and it provides timely review results. Moreover, we demonstrate that the proposed scheme can update its parameters in an incremental manner, and thus reduce greatly the computational complexity in real world implementation. The scheme is simple, effective, and applicable to other Web 2.0-based review systems in which the product reviews are time-stamped and they can be evaluated by other users.

Acknowledgement

We are grateful to the editors and anonymous reviewers for their insightful comments. This material is based upon work supported by the Taiwan E-learning and Digital Archives Program (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 99-2631-H-001-020, NSC 100-2631-H-001-013, and NSC 100-2631-S-003-006.

References

[1] IT Facts. http://www.itfacts.biz/.

[2] LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

[3] R. Bhattacharjee and A. Goel. Avoiding ballot stuffing in ebay-like reputation systems. In *ACM SIGCOMM workshop on Economics of peer-to-peer systems*, 2005.

[4] V. Carchiolo, A. Longheu, M. Malgeri, and G. Mangioni. Context-based Global Expertise in Recommendation Systems. *Informatica*, 34(4):409–417, 2010.

[5] P.-Y. Chen, S. Dhanasobhon, and M. D. Smith. All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon.com. *SSRN eLibrary*, 2008.

[6] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 48(3):345–354, August 2006.

[7] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *ACM Electronic Commerce Conference*, 2000.

[8] P. Dondio and S. Barrett. Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project. *Informatica*, 31(2):151–160, 2007.

[9] D. Houser and J. Wooders. Reputation in auctions: Theory, and evidence from ebay. *Journal of Economics & Management Strategy*, 15(2):353–369, June 2006.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *ACM SIGKDD*, 2004.

[11] N. Hu, P. A. Pavlou, and J. Zhang. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *ACM Electronic Commerce Conference*, 2006.

[12] N. Jindal and B. Liu. Analyzing and detecting review spam. In *IEEE International Conference on Data Mining*, 2007.

[13] A. Josang and J. Haller. Dirichlet reputation systems. In *International Conference on Availability, Reliability and Security*, 2007.

[14] A. Josang and R. Ismail. The beta reputation system. In *15th Bled Conference on Electronic Commerce*, 2002.

[15] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, March 2007.

[16] M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Charles Griffin & Company Limited, 5 edition, September 1990.

[17] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *International Conference on Empirical Methods in Natural Language Processing*, 2006.

[18] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *The Annual Meeting on Association for Computational Linguistics*, 2003.

[19] D. Klein and C. D. Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, 2003.

[20] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing*, 7(1):76–80, January/February 2003.

[21] R. A. Malaga. Web-based reputation management systems: Problems and suggested solutions. *Electronic Commerce Research*, 1(4):403–417, October 2001.

[22] L. Mui, M. Mohtashemi, C. Ang, and P. Szolovits. Ratings in distributed systems: A bayesian approach. In *Workshop on Information Technologies and Systems*, 2001.

[23] P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *Advances in Applied Microeconomics*, 11:127–157, 2002.

[24] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[25] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101, June 2006.

[26] J. B. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In *ACM Electronic Commerce Conference*, 1999.

[27] M. Srivatsa, L. Xiong, and L. Liu. Trustguard: countering vulnerabilities in reputation management for decentralized overlay networks. In *International World Wide Web Conference*, 2005.

[28] B.-C. Wang, W.-Y. Zhu, and L.-J. Chen. Improving the amazon review system by exploiting the credibility and time-decay of public reviews. In *International Workshop on Web Personalization, Reputation and Recommender Systems*, 2008.

[29] Y. Wang and J. Vassileva. Trust and reputation model in peer-to-peer networks. In *International Conference on Peer-to-Peer Computing*, 2003.

[30] Y. Weng, C. Hu, X. Zhang, and L. Zhao. BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis. *Informatica*, 34(4):419–328, 2010.

[31] A. Whitby, A. Josang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *The International Joint Conference on Autonomous Agenst Systems*, 2004.

[32] G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms for electronic marketplaces. *Decision Support Systems*, 29(4):371–388, December 2000.

[33] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *ACM International Conference on Information and Knowledge Management*, 2006.

Human Detection Based on Large Feature Sets Using Graphics Processing Units

William Robson Schwartz
Institute of Computing, University of Campinas, Campinas-SP, Brazil, 13083-852
schwartz@ic.unicamp.br
http://www.liv.ic.unicamp.br/~wschwartz

Keywords: human detection, large feature sets, graphics processing unit, partial least squares

Received: December 8, 2010

To obtain a better representation for human detection, edge, color, and texture information have been combined and employed. However, this combination results in an extremely high-dimensional feature space. The large number of feature descriptors results in expensive feature extraction and requires a dimension reduction process. Frameworks based on general purpose graphics processing unit (GPU) programming have been successfully applied in computer vision problems and in this work we model the human detection problem so that multi-core CPUs and multiple GPU devices can be used to speed-up the process. The experimental results show significant reduction on computational time when compared to the serial CPU based approach.

Povzetek: Članek obravnava prepoznavanje ljudi na slikah s pomočjo grafičnega procesorja.

1 Introduction

Human detection is a topic of interest in computer vision since people's locations play an important role in applications such as human-computer interaction and surveillance. However, detecting humans in single images is a challenging task due to both inter- and intra-person occlusion and variations in illumination and individual's appearances and poses.

There are two main approaches to human detection: part-based [15, 8, 6] and subwindow-based [1, 2, 16]. The first class of methods consists of a generative process where parts of the human body are combined according to a prior model. The latter class of methods aim at effectively deciding if a human is located in a window by combining low-level features extracted from subwindows (or blocks).

The work of Dalal and Triggs [2] obtained high detection rates employing histograms of oriented gradients (HOG) as feature descriptors. Subsequent improvements in human detection have been achieved mostly by using combinations of low-level descriptors [7, 9, 16, 18]. A strong set of features provides high discriminatory power.

Edge, color and texture information are among the characteristics able to distinguish between humans and background [13]. These clues can be captured by low-level descriptors: the original HOG descriptors with additional color information, called *color frequency*, and features computed from co-occurrence matrices.

To allow more location and pose flexibility within the detection window and to capture information at different scales, features are extracted at different sizes, using overlapping blocks.

In order to augment the information of edge-based features, we combine the original HOG with features providing texture and color information. Texture is measured using classical co-occurrence matrices [4], which have been used previously for texture classification. To capture color information we use a simple color extension of HOG features, called color frequency. A consequence of augmenting the HOG features with color and texture features is an extremely high-dimensional feature space.

Even though good results can be achieved using the described feature combination (as we will show in the experiments), the computational cost is directly influenced by the large number of features considered; therefore, feature extraction and dimensionality reduction become very expensive processes and need to be optimized.

GPU devices have become a powerful computational hardware for a given price and multiple of such devices may be attached to a single computer. This results in a powerful computational tool when the application can be split in several independent parts suitable to run in parallel. The subwindow-based approach for human detection is suitable for GPU implementation since the detection windows for different regions in the image are independent and therefore can be considered in parallel. Therefore, this work models the human detection problem so that multi-core CPUs and multiple GPU devices can be used to speed-up the process.

This paper is organized as follows. Section 2 describes works related to the proposed one. Section 3 describes the serial approach for the problem of human detection. Then, in Section 4 we present the proposed parallel approach for the problem. Experimental results comparing the serial and parallel approaches are shown in Section 5. Finally, Section 6 concludes with some final remarks.

2 Related Work

Frameworks based on general purpose GPU programming have been applied in computer vision problems and have provided high performance computation in problems such as feature tracking and matching [14], real-time stereo [5, 19], background subtraction [3], and motion detection [20].

To speed-up the detection process, we design the human detector implementation in such a way that we are able to exploit parallel computational power provided by multi-core CPUs and GPU devices. Our design also avoids redundant computation of features shared by different detection windows. As demonstrated in the experimental results, the proposed approach leads to a significant reduction on the computational time.

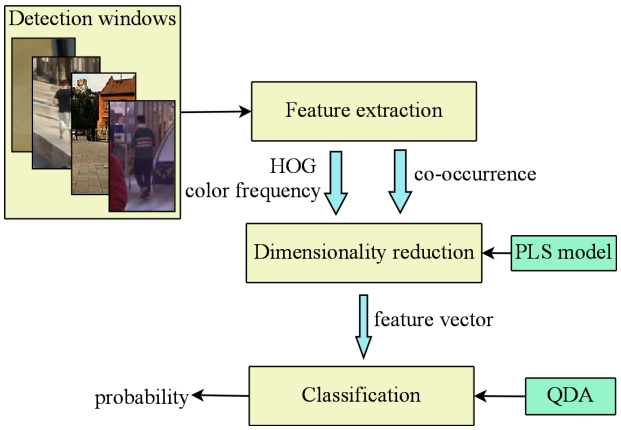


Figure 1: Serial approach. For each detection window in the image, features extracted using feature combination concatenated and analyzed by the partial least squares (PLS) model to reduce dimensionality, resulting in a low dimensional vector. Then, quadratic discriminative analysis (QDA) is used to classify this vector as either human or non-human. Finally, the probability of the detection window be classified as a human is output.

The work developed by Zhang and Nevatia [21] also uses GPU for human detection. However, they do not exploit the computational power of the multiple cores available in nowadays CPUs and the availability of multiple GPU devices attached to the computer, as we propose in this work. In addition, since Zhang and Nevatia implement the work proposed in [2], only features based on HOG are implemented, which leads to poor results (as it will be shown in the experimental results).

3 Serial Approach

In this section we review the serial approach for human detection, based on the work of Schwartz et al. [13]. The diagram shown in Figure 1 illustrates the steps of the serial approach for human detection. We decompose a detection window, d_i, into overlapping blocks and, extracting a set of features for each block, we construct the feature vector $\boldsymbol{v_i}$, describing d_i.

3.1 Feature Extraction

To capture texture, features extracted from co-occurrence matrices are used [4]. These matrices provide information regarding homogeneity and directionality of patches, which are important in human detection once a person tends to wear clothing composed of homogeneous textured regions and there is a significant difference between the regularity of clothing texture and background textures.

Edge information is captured using histogram of oriented gradients, which captures gradient structures that are characteristic of local shape [2]. In HOG, the orientation of the gradient for a pixel is chosen from the color band corresponding to the highest gradient magnitude. Some color information is captured by the number of times that each color band is chosen. A three bin histogram is build to tabulate the number of times that each color band is chosen. This feature is called color frequency [13].

3.2 Dimensionality Reduction

To handle the high dimensionality resulting from the combination of features, a statistical method called partial least squares (PLS) [17] is employed as a linear dimensionality reduction technique. PLS provides dimensionality reduction for even hundreds of thousands of variables, accounting for class labels in the process.

The basic idea of PLS is to construct a set of projection vectors $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \boldsymbol{w}_p\}$ given the standardized data summarized in the matrix $\boldsymbol{X}$ of descriptor variables (features) and the vector $\boldsymbol{y}$ of response variables (class labels). The objective of the procedure is to derive a small, relevant set of latent variable vectors that captures the information inherent in the matrix $\boldsymbol{X}$ of descriptor variables in a compact form [12].

The dimensionality reduction is performed by projecting the feature vector $\boldsymbol{v_i}$ extracted from a detection window d_i onto the latent vectors $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \boldsymbol{w}_p\}$. Vector $\boldsymbol{z_i}$ $(1 \times p)$ is obtained as a result. This vector is used in classification.

3.3 Classification

Once the feature extraction process is performed for all blocks inside a detection window d_i, features are concatenated creating an extremely high-dimensional feature vector $\boldsymbol{v_i}$. $\boldsymbol{v_i}$ is then projected onto set of latent variables W resulting in a low dimensional vector $\boldsymbol{z_i}$. For each vector $\boldsymbol{z_i}$, we use quadratic discriminant analysis to estimate probabilities for the two classes, human and non-human.

PLS tends to produce latent vectors that provide an approximately linear separation of the two classes. This enables us to use simple classifiers, such as QDA. Figure 2

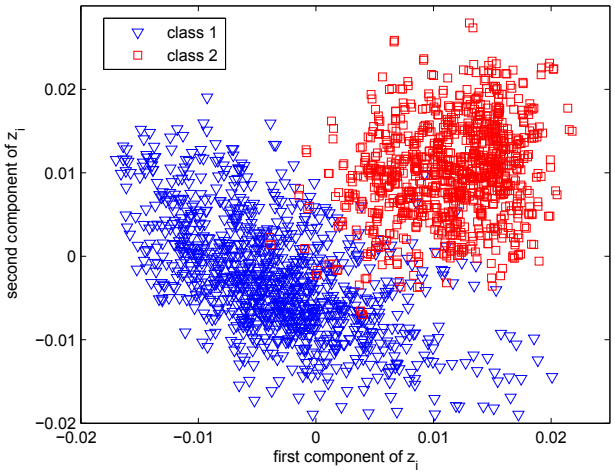


Figure 2: First two components of z_i obtained from projecting feature vectors into latent vectors for human and non-human classes. It is possible to see the clear separation between both classes provided by PLS used as a linear dimensionality reduction technique.

shows the first two components of vectors z_i for the different classes extracted from the training data. We see that the classes do not overlap much, even in a 2-dimensional projection space. One of the advantages of using a simple classifier as QDA is that the computational time to perform the classification task is very low.

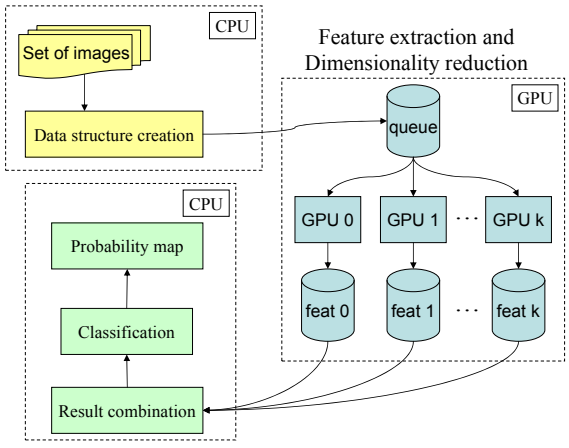


Figure 3: Parallel approach. Processing modules are executed simultaneously in CPU and GPU devices. Data structures used in co-occurrence and HOG methods are created in CPU. Such structures are used during the feature extraction and dimensionality reduction process performed on GPU. Finally, the low dimensional features resulting from multiple GPUs for each detection window are combined and classified.

4 Parallel Approach

Designs like the one illustrated in Figure 1 are not able to take advantage of the full computational power provided by current computers. Therefore, it is of interest to design a detection approach that accomplishes the same task in a parallel fashion.

To exploit parallelism, we propose the approach illustrated in Figure 3. This design distributes the processing among the CPU and GPU devices available in the system to reduce the idle time of the processors. In the following subsections we describe the modules composing this approach.

4.1 Data Structure Creation

Before performing feature extraction, some data structures need to be created for each input image. Integral histograms are created for HOG and matrices for the co-occurrence methods.

Once the time to create the data structures is substantially smaller than the feature extraction and dimensionality reduction (as we will show in the experiments), these two set of operations may be decoupled and executed in parallel increasing the use of the computational resources. The data structure creation is performed on the CPU and the feature extraction and dimensionality reduction on the GPU. To synchronize between these devices, we add a queue so that once the data structures are created for an image, they are stored in the queue until the GPU devices become available to process that particular image.

4.2 Feature Extraction and Dimensionality Reduction

In the serial approach, a detection window is decomposed into overlapping image blocks from which are extracted features to compose a feature vector. Since features for different image blocks can be extracted independently and a linear technique is used for dimensionality reduction, we can exploit the multiprocessors available in GPUs to extract features and reduce the dimensionality of multiple image blocks simultaneously, then, at the end, combine the results to obtain a low dimensional feature vector to describe a detection window.

In the GPU device, each multiprocessor consists of a set of scalar processor cores and employs an architecture called SIMT (single-instruction, multiple-thread). The multiprocessor maps each thread to one scalar processor core, and each scalar thread executes independently with its own instruction address and register state. The implementation in this work uses the parallel computing architecture developed by NVIDIA called Compute Unified Device Architecture (CUDA) [11]. The extensions to C programming language provided by CUDA allow that general-purpose computation be performed in GPUs in a well-defined and structured manner.

To exploit the set of multiprocessors in a GPU, several image blocks are processed simultaneously. This way, the input of each GPU is a range of image blocks that need to be processed and the intermediate results of feature extraction and dimensionality reduction are stored in an array $\boldsymbol{f}$ containing an entry for each detection window in the current image. $\boldsymbol{f}$ is indexed by the detection window index, e.g. $\boldsymbol{f}_i$ represents low dimensional features for the i-th detection window. At the end, the results obtained from all GPUs are combined so that the detection windows can be classified.

Algorithm 5 Steps performed for each GPU.

Input: set of blocks $\{b_i, b_{i+1}, \ldots, b_j\}$.
launch simultaneously GPU processes to extract features and reduce dimensionality of all image blocks $b_k \in \{b_k : k = i, \ldots, j\}$
Output: array $\boldsymbol{f}$ containing low dimensional feature vectors for each detection window in the image.

Algorithm 5 describes the steps for each GPU device used during human detection. According to this algorithm, the feature extraction and dimensionality reduction are scalable with the number of GPU devices available since the image blocks can be divided among the devices and their processing is independent.

Algorithm 5 launches a multithread process to extract feature for each image block. Then, the dimensionality reduction for all detection windows sharing that specific block is performed. Once we use a sliding window to search for humans, one image block may be shared by several detection windows. The dimensionality reduction needs to be performed for each one of these detection windows. We describe this process as follows.

Given a image block b_k, features are extracted using co-occurrence, HOG and color frequency methods and a feature vector v_k is composed. Let $l_k = \{d_{k,0}, d_{k,1}, \ldots, d_{k,l}\}$ be the set of detection windows sharing the image block b_k. We project v_k onto latent vectors (learned using PLS) corresponding to each detection window $d_{k,j}$. This computes the contribution of features extracted from b_k to the detection window $d_{k,j}$. Finally, we add this contribution to $\boldsymbol{f}_{d_{k,i}}$, position corresponding to low dimensional feature vector for detection window $d_{k,j}$. Algorithm 6 shows the steps performed to process each image block assigned by algorithm 5.

Since the last step of Algorithm 6 adds the contribution of each image block to a detection window and multiple image blocks are processed simultaneously, two multithread processes may write at the same position of $\boldsymbol{f}$ at once. To prevent this behavior without incurring unwanted overhead, the image blocks are sorted so that different blocks being processed at the same time do not share a common detection window.

Algorithm 6 Steps to process each image block.

Input: image block b_k, list l_k of detection windows sharing block b_k, projection vectors of PLS model for dimensionality reduction.
$v_k \leftarrow$ co-occurrence, HOG and color frequency features extracted from image block b_k.
for each detection window $d_{k,j} \in l_k$ **do**
load projection vectors for block b_k for detection window $d_{k,j}$
project v_k onto projection vectors
add projection result to $\boldsymbol{f}_{d_{k,j}}$
end for

4.3 Classification

Once the feature extraction and dimensionality reduction is finished for an image, classification is performed. This module is performed on the CPU because results (low dimensional feature vectors from GPUs) need to be combined prior to the classification. Additionally, due to the asynchronism between the CPU and GPU devices, queues store the low dimensional feature vectors outputted from each GPU device. After that, the feature vectors corresponding to an image are added and then the classification is performed resulting in a probability map for each image.

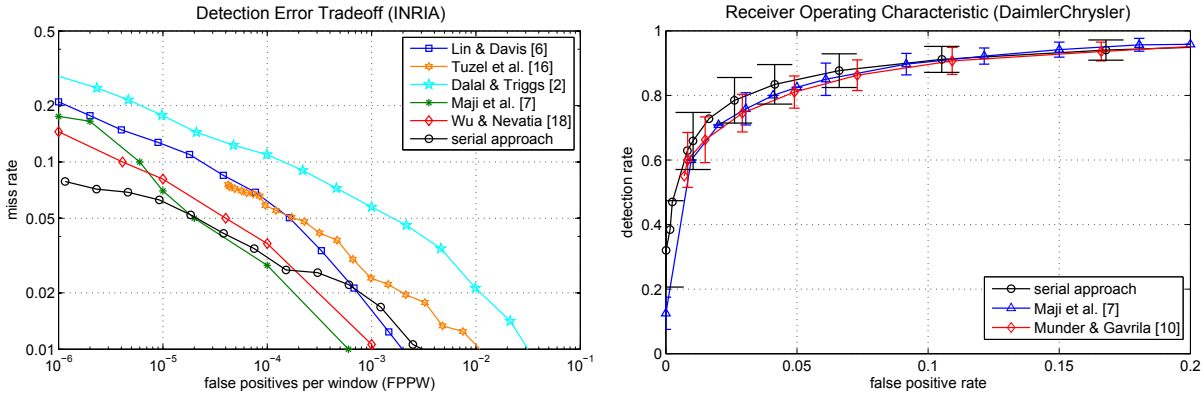
The addition of queues in the process also allows the use of heterogeneous GPU devices, e.g. GPU models presenting different computational power.

5 Experimental Results

Implementation Details For the INRIA Person dataset [2], the setup of the feature extraction proposed in [13] is used, as described as follows. The co-occurrence features use block sizes of 16×16 and 32×32 with strides of 8 and 16 pixels respectively. The color space is converted to HSV and for each color band, 12 features are extracted from co-occurrence matrices created for each one of the four directions. The displacement considered is 1 pixel and each color band is quantized into 16 bins. The described setup results in $63,648$ features.

For HOG feature extraction blocks with sizes ranging from 12×12 to 64×128 are considered, resulting on $2,748$ blocks. For each block, 36 features are extracted, resulting in a total of $98,928$ features. In addition, the same set of blocks is employed to extract features using the color frequency method. This results in three features per block, and the total number of resulting features is $8,244$. Considering the aggregation of the three feature channels, the resulting feature vector extracted from one detection window contains $170,820$ elements.

Performance of the Serial Approach Initially, we describe results achieved by the serial detection method in order to show that efforts to obtain a fast implementation are



(a) INRIA person dataset (b) DaimlerChrysler pedestrian dataset

Figure 4: Comparison of the serial approach used in the optimization with other methods in multiple human detection datasets.

worthwhile. We compared its performance to other methods in the literature using two standard human detection datasets: the INRIA Person dataset [2] and the Daimler-Chrysler Pedestrian Classification Benchmark dataset [10]. For the DaimlerChrysler dataset, the number of elements in the feature vector is reduced due to he smaller size of the samples.

Figure 4(a) compares results obtained by the serial approach used in the GPU optimization with works published previously. Figure 4(b) compares results obtained by the serial approach with results reported in [7, 10]. Curves in Figure 4(a) are presented using detection error tradeoff curves on log-log scales. The x-axis corresponds to false positives per window (FPPW), defined by *FalsePos/(TrueNeg + FalsePos)* and y-axis shows the miss rate, defined by *FalseNeg/(FalseNeg + TruePos)*. While, curves in Figure 4(b) show detection rate instead of miss rate on the y-axis and both axes are shown in linear scales.

For both datasets, the results obtained with the described feature combination method outperform the other methods in regions of low false alarm rates, which are considered important regions for the human detection problem.

In addition, Figure 5 shows results obtained in full images obtained from the Google images and from the testing set of the INRIA Person dataset. To be able to detect people with different sizes, 11 different scales were considered for each image. The high number of detection windows resulting from the multiple-scale detection in each image also justifies the idea of considering a GPU-based implementation to speed-up the detection process.

Speed-up To test the performance of the parallel implementation we conducted experiments using sets of GPU devices NVIDIA GeForce 9800 GX2 and NVIDIA Tesla C870. Four GPU devices model GeForce 9800 GX2 were available in a Intel Core 2 Quad Q9450 2.66GHz with 4GB of RAM memory and two GPU devices model Tesla C870 were available in an AMD Opteron Dual 2218 2.6GHz with 2Gb of RAM memory. In the experiments we performed human detection in 100 images with 320×240 pixels using shift of 4 pixels between consecutive detection windows.

Table 1 shows the computational time obtained by the serial and parallel approaches using GeForce 9800 GX2 GPU devices. Similarly, Table 2 shows the results obtained using Tesla C870 GPU devices. Transfer time refers to the time spent copying data from the CPU to GPU, and vice-versa. Overhead is the computational time required to cache the features to be used for different detection windows in the serial approach. On the serial implementation, the cache is implemented so that features computed for an image block are stored to be used subsequently by different detection windows sharing that block.

According to Tables 1 and 2 we see that the parallelism between CPU and GPU is being exploited since the data structure creation time does not contribute significantly to the total computational cost. In addition, the multiprocessors in the GPU allow time reduction during feature extraction and dimensionality reduction.

The last two rows of tables 1 and 2 show the number of detection windows processed per second and the speed-up obtained when compared to the serial approach. Although the gain in speed is not linearly proportional to the number of GPUs, due to increasing time to data structure creation and data transfer, we see significant speed-up when more GPU devices are added into the system.

6 Conclusions

In this work we described a parallel design exploiting multi-core CPUs and multiple GPU devices for the human detection problem. The results have shown that the computational power of both CPU and GPUs can be exploited to obtain a faster implementation.

Even though the optimization framework described in this paper is focused on human detection, it is general enough to be easily applied to other object detection tasks relying on sliding detection windows.

	CPU	1 GPU	2 GPUs	3 GPUs	4 GPUs
Data structure creation	-	39.86s	41.82s	43.97s	47.42s
Transfer time	-	21.27s	23.96s	26.44s	34.83s
Overhead	243.76s	-	-	-	-
Feature extraction + dim. reduction	785.58s	394.06s	207.97s	137.56s	112.31s
Total time	1029.34s	418.14s	217.15s	149.19s	128.95s
detections/second	183.44	451.57	869.55	1265.65	1464.41
speed-up	1×	2.4×	4.7×	6.8×	7.9×

Table 1: Computational time and speed-up for serial and parallel approaches using NVIDIA GeForce 9800 GX2.

	CPU	1 GPU	2 GPUs
Data structure creation	-	66.06s	93.52s
Transfer time	-	8.20s	26.87s
Overhead	904.02s	-	-
Feature extraction + dim. reduction	1897.48s	228.68s	133.29s
Total time	2801.50s	238.02s	164.05s
detections/second	67.40	793.30	1151.01
speed-up	1×	11.7×	17.1×

Table 2: Computational time and speed-up for serial and parallel approaches using NVIDIA Tesla C870.

(a) (b) (c) (d)

Figure 5: Results obtained in full images containing people of multiple sizes.

Acknowledgments

This research was supported by FAPESP grant 2010/10618-3.

References

[1] J. Begard, N. Allezard, and P. Sayd. Real-Time Human Detection in Urban Scenes: Local Descriptors and Classifiers Selection with AdaBoost-like Algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[3] A. Griesser, S. De Roeck, A. Neubeck, and L. Van Gool. GPU-Based Foreground-Background Segmentation using an Extended Colinearity Criterion. In *Vision, Modeling, and Visualization*, pages 319–326, 2005.

[4] R. Haralick, K. Shanmugam, and I. Dinstein. Texture Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6), 1973.

[5] P. Labatut, R. Keriven, École Nationale, and P. Chaussées. A GPU Implementation of Level Set Multi-View Stereo. In *International Conference on Computational Science Ű Workshop General Purpose Computation on Graphics Hardware*, pages 212–219, 2005.

[6] Z. Lin and L. S. Davis. A Pose-Invariant Descriptor for Human Detection and Segmentation. In *European Conference on Computer Vision*, pages 423–436, 2008.

[7] S. Maji, A. Berg, and J. Malik. Classification using Intersection Kernel Support Vector Machines is Efficient. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2008.

[8] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, 2004.

[9] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou. Discriminative Local Binary Patterns for Human Detection in Personal Album. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2008.

[10] S. Munder and D. Gavrila. An Experimental Study on Pedestrian Classification. *PAMI*, 28(11):1863–1868, 2006.

[11] NVIDIA. *NVIDIA CUDA Programming Guide 2.0*. 2008.

[12] R. Rosipal and N. Kramer. Overview and Recent Advances in Partial Least Squares. *Lecture Notes in Computer Science*, 3940:34–51, 2006.

[13] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human Detection Using Partial Least Squares Analysis. In *IEEE International Conference on Computer Vision*, pages 24–31, 2009.

[14] S. N. Sinha, J. michael Frahm, M. Pollefeys, and Y. Genc. GPU-based Video Feature Tracking and Matching. Technical report, In Workshop on Edge Computing Using New Commodity Architectures, 2006.

[15] D. Tran and D. Forsyth. Configuration Estimates Improve Pedestrian Finding. In *NIPS 2007*, pages 1529–1536. 2008.

[16] O. Tuzel, F. Porikli, and P. Meer. Human Detection via Classification on Riemannian Manifolds. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2007.

[17] H. Wold. Partial least squares. In S. Kotz and N. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. 1985.

[18] B. Wu and R. Nevatia. Optimizing Discrimination-Efficiency Tradeoff in Integrating Heterogeneous Local Features for Object Detection. In *IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2008.

[19] R. Yang and M. Pollefeys. Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware. pages 211–217, 2003.

[20] Q. Yu and G. Medioni. A GPU-Based Implementation of Motion Detection from a Moving Platform. 2008.

[21] L. Zhang and R. Nevatia. Efficient Scan-Window Based Object Detection Using GPGPU. In *Computer Vision and Pattern Recognition Workshops*, 2008.

Examining Whether Highly E-Innovative Firms are More E-Effective

Pedro Soto-Acosta
Department of Management and Finance, University of Murcia, Campus de Espinardo
30100 Espinardo (Murcia), Spain
E-mail: psoto@um.es

Ricardo Colomo-Palacios
Computer Science Department, Universidad Carlos III de Madrid
Av. Universidad, 30, 28911 Leganés (Madrid), Spain
E-mail: ricardo.colomo@uc3m.es

Daniel Perez-Gonzalez
Department of Business administration, University of Cantabria
Avda. de los Castros S/N, 39005 Santander (Cantabria), Spain
E-mail: daniel.perez@unican.es

Keywords: e-Business, resource-based theory, innovation

Received: September 8, 2010

The resource-based view (RBV) ascribes superior firm performance to firm resources and capabilities. In recent years, much debate about the value of e-business has been raised because of the costly investments required. Although studies have found positive relationships between e-business and firm performance, there is a need to further investigate into these topics. Since innovation has become a key factor for increasing the competitiveness of firms and e-business has been proposed as complement to innovation, this paper analyses, based on the RBV perspective, whether companies with high level of Internet resources and with high e-innovation are more effective electronically. The methodology involved a large sample firms and data collected by the European e-Business Market Watch, an established e-business observatory sponsored by the European Commission. Results indicated that differences of e-sales effectiveness of firms with high and low Internet resources were not statistically significant, while on the contrary firms with a high level of e-innovation outperformed on e-sales effectiveness.

Povzetek: Članek preučuje, koliko raznovrstna uporaba interneta izboljša spletno prodajo.

1 Introduction

The relationship between information technology (IT) and business value has been the subject of much research over the past decade. The results of these studies were varied and the term "productivity paradox" was coined to describe such findings. Nonetheless, recent studies have found positive and stronger linkages, and have attributed the productivity paradox to variation in methods and measures [22, 44]. Firms make important investments in the development of costly IT infrastructures to benefit from the real-time connectivity and collaboration capabilities provided by the Internet, and to conduct various types of e-business activities [18, 38, 50, 51]. Therefore it is quite important to understand whether and how such IT and Internet-related infrastructures create business value, so that appropriate guidance can be provided to managers.

Although IT in general and e-business provide distinct value propositions to the firm, it has been argued that the technology itself is available to all firms (including competitors), so it will rarely create superiority. In this sense, evidence suggesting that IT spending rarely correlates to superior performance exists [22, 9, 11, 34, 40, 44]. However, even though competitors may copy an IT infrastructure, relative advantage can be created and sustained in cases where the technology leverages some other critical resources. The literature suggests that a number of such complementary resources, such as size, structure, culture, and so on, that could make it difficult for competitors to copy the total effect of the technology [3, 2, 30]. This complementarity of resources is a corner stone of the resource-based view (RBV) of the firm [4, 28] and has been offered as an explanation of how IT has largely overcome its paradoxical nature and is contributing to business value [6, 7, 15, 34, 43].

Innovation can be defined as the search for, the discovery and development of new technologies, new products and/or services, new processes and new organizational structures [10]. Many researchers [e.g. 23] emphasized the role of IT as an enabler of innovation, suggesting that IT produces innovations in business processes, products

and services that lead to better firm performance [9, 12, 13, 16, 24, 25, 27]. In this sense, there is considerable literature arguing that Internet technologies have enabled substantial transformations in firms with regard to their business models, internal processes, value propositions and services, providing considerable benefits [1, 36, 49, 54, 55, 59]. As a result, research is starting to focus on analysing whether and how the web is and will change innovation within and between companies.

To respond to these challenges, this paper examines, grounded in resource-based view (RBV) of the firm perspective, whether companies with high level of e-innovation are more effective electronically, which is measured as the effectiveness of online sales. Also, it assesses whether the level of Internet resources is related to e-sales effectiveness. The analysis employs data from a large sample of firms from different industries, which have been collected by the European e-Business Market Watch (www.ebusiness-watch.org), an established e-business observatory organization sponsored by the European Commission. The results of this analysis are interesting to researchers, firms' managers of various levels and consultants dealing with e-business and/or innovation.

The paper consists of six sections and is structured as follows: The next section 2 outlines the background of this study. Following that, the data and methodology of this study are discussed in section 3. Then, data analysis and empirical results are presented in section 4. Finally, the paper ends with a discussion of research findings in section 5, and conclusions, limitations and proposed future research directions in section 6.

2 Literature Review

2.1 The RBV within IT and e-business literature

The RBV of the firm has its origins in the management strategy literature and has been used in a variety of management, including management of information systems, to explain and study the sources of sustained competitive advantages [4, 42]. The RBV is based on two underlying arguments: resource heterogeneity and resource immobility. Resources and capabilities possessed by competing firms are heterogeneously distributed and may be a source of competitive advantage when they are valuable, rare, difficult to imitate, and not substitutable by other resources [4, 52]. At the same time, resources and capabilities are a source of sustained competitive advantage, that is, differences may be long lasting (resource immobility) when protected by barriers to imitation [33] or isolating mechanisms such as time-compression diseconomies, historical uniqueness, embeddedness and causal ambiguity [4, 39]. Consequently, the RBV suggests that the effects of individual, firm-specific resources and capabilities on performance can be significant [33].

The RBV provides a solid foundation to differentiate between IT resources and IT capabilities and to study their separate influences on performance [43]. Based on this analysis, Bharadwaj [6] suggested that if firms can combine IT related resources to create unique IT capabilities, they can improve their performance. IS researchers have followed this consideration of IT capability because competition may easily result in the duplication of investment in IT resources, and companies can purchase the same hardware and software to remove competitive advantage [43]. In this respect, research offers a useful distinction between IT resources and IT capabilities. The former is asset-based, while the latter comprises a mixture of assets formed around the productive use of IT, being capabilities are rooted in processes and business activities [44].

In general, IT resources are not difficult to imitate; physical technology is by itself typically imitable. If one firm can purchase these physical technologies and thereby implement some strategies, then other firms should also be able to purchase these technologies, and thus such tools should not be a source of competitive advantage [4]. However, firms may obtain competitive advantages from exploiting their physical technology in a better (and/or different) way than other firms, even though competing firms do not vary in terms of the physical technology they possess. IT resources are necessary, but not a sufficient condition, for competitive advantages [15]. IT resources rarely contribute directly to competitive advantage. Instead, they form part of a complex chain of assets (IT capabilities) that may lead to better performance. Thus, some researchers have described this in terms of IT capabilities and argue that IT capabilities can create uniqueness and provide organizations a competitive advantage [6, 7, 34, 43].

This research framework is very useful for our study, because it enables on the one hand to distinguish between Internet resources (an IT resource) and, on the other hand, the results from the e-innovation capability (a mixture of resources, including IT resources) and, then, to examine the effect of each one on e-effectiveness. Internet resources are not difficult to imitate. Internet technology is by itself imitable. If one firm can purchase certain Internet technologies and thereby implement some strategies, then other firms should also be able to purchase these technologies and implement similar strategies. These arguments suggest that Internet resources may not have a significant impact on e-effectiveness.

2.2 E-business and Innovation

There is considerable literature analyzing the innovative potential of the Internet/e-business. This existing literature concludes that e-business enables and drives significant innovative transformations regarding business models, value propositions, products and services of firms and internal business processes, which can offer substantial benefits [1, 54, 55, 49, 48, 59]. Amit and Zott [2], based on one hand on a broad theoretical foundation concerning virtual markets, value chain analysis, Schumpeterian innovation, resource-based view of the firm, strategic networks and transaction cost economics,

and on the other on extensive cases study, proposed four dimension of innovation and value creation in e-business: transaction efficiency, novelty, complementarities (between various products and services, on-line and off-line assets, activities) and customers lock-in. Wu and Hisa [54, 55] categorise the innovations caused by e-commerce based on the extent of change in product's core components (defined as 'the distinct portions of the product that embody the core design concept and perform a well-defined function') and on the extent of change in the business model (defined as 'the way in which the components are integrated and linked into a coherent whole') into four groups: incremental innovation (no significant changes in core components and business models), modular innovation (considerable changes in core components but not in business model), architectural innovation (considerable changes in business model but not in core components) and radical innovation (considerable changes in both core components and business model). Tavlaki and Loukis [48] propose a methodology for designing new 'digital business models', which consists of six stages: design of value proposition, design of production architecture (value chain), definition of value chain actors, analysis of competition, design of economic model and elaboration of relations among actors. Another research stream focuses on analysing how the web supports 'distributed' collaborative innovation creation both within and among firms. Timmers [49] argues that Internet gives rise to new business models, and describes the most important of them: e-shop, e-procurement, e-auction, e-mall, third party marketplace, virtual community, value chain service provider, value chain integrator, collaboration platform, information brokerage and trust services. Zwass [59] argues that the WWW/Internet compound enables significant innovations in the way organizations arrange their business processes, address their marketplaces and partner with other organizations; also, he proposes a large number of innovation opportunities grouped in eleven categories associated with marketplace, universal supply-chain linkage, network of relationships, collaboration, use of forum, interactive media, goods and services delivery, anytime-anywhere connectivity, development platforms, universal telecommunications networks and computing utility.

The RBV research framework is also useful for our study, because it enables to suggest that the results from the e-innovation capability (a mixture of resources, including Internet resources) are firm-specific and, hence, may have a positive impact on e-effectiveness. That is, merely having Internet resources may not generate value per se, but if these resources are used in combination with other resources to build IT capabilities such as the e-innovation capability, the output from this type of capabilities is, in accordance with the RBV, business value and effectiveness improvements.

2.3 Organizational Impact of e-Business and e-innovation: e-effectivess

The evaluation of the organisational performance impact of ITs is also an important issue within the area management information systems. In this sense, firm performance has been principally measured by subjective measures [e.g., 17, 32, 44, 45, 57] or by using financial measures [e.g., 5, 35, 58]. The first normally uses senior executives as the key informants on the subjective measures of firm performance. These studies has produced considerable evidence that e-business has a positive impact on various non-financial and financial measures of organizational performance. However, none of these studies has dealt with e-innovation and its impact on performance. Given the fact that e-business investments may provide benefits after a certain period but increase operating costs in the short term, the locus of impact, the business process, should be the primary level of analysis. As a result, some researchers have given up on trying to correlate financial results with IT investments and suggest focusing on the actual processes that IT is supposed to enhance [37]. These arguments lead to the conclusion that a process approach should be used to explain the generation of IT value from a resource-based perspective, and this is the approach adopted in this study. The present research uses the effectiveness of online as a measure of firm performance. The business value of this process is discussed here.

Selling online can potentially provide distinct value propositions to the firm. These come from its positive impact on the volume of sales, the number of customers and the quality of customer service. The Internet enables high reach and richness of information [19] and connects firms to consumers or potential consumers in geographic areas that would be costly to reach before the Internet [46]. All this can help increasing sales and number of customers. Moreover, virtual communities enable frequent interactions with customers on a wide range of topics and thereby create a loyalty and enhance transaction frequency [2]. At the same time, e-business allows innovation in the way firms do business (new business models) and the introduction of new products and services, which may again influence sales and number of customers. In addition, selling online can provide value through the automation of the sales processes, which reduces overall load on staff supporting the customer and allows staff to focus on more complex tasks or on exceptions instead of routine tasks.

3 Methodology

3.1 Data

The data source for the present study is the European e-Business Market Watch (www.ebusiness-watch.org), an initiative launched by the European Commission for monitoring the adoption of IT and e-business activity in Europe. The field work of the survey was conducted by Ipsos Eco Consulting on behalf of the e-business Watch and was carried out using computer-aided telephone

interview (CATI) technology. Telephone interviews with decision-makers in firms were conducted. The decision-maker targeted by the survey was normally the person responsible for IT within each firm, typically the IT manager. Alternatively, particularly in small firms not having a separate IT unit, the managing director or owner was interviewed.

Number of employees	%	N
1-9	38.4	338
10-49	25.8	261
50-249	26.8	271
More than 249	8,9	90
Respondent title	%	N
Owner/proprietor	12.1	122
Managing director	19.6	198
Strategy development	1.9	19
Head of IT/DP	22	222
Other IT senior member	32.4	327
Others	12.1	122

Table 1: Sample Characteristics.

The population considered in this study was the set of all firms which are active at the national territory of Spain and which have their primary business activity in one of ten highly important sectors considered. The sample drawn was a random sample of firms from the respective sector population with the objective of fulfilling strata with respect to business size. A share of 10% of large companies (250+ employees), 30% of medium sized enterprises (50-249 employees) and 25% of small enterprises (10-49 employees) was intended. The final number of firms totalled 1,010. As shown in Table 1, 91.1% of firms were small or medium-sized, and each sector considered had a share of around 10% of the total sample.

With regard to respondents' positions, 54.4% were IS managers, nearly 20% were managing directors, and 12.1% were owners. The dataset was examined for potential bias in terms of the respondents' positions. Since respondents included both IT managers and non-IT managers, one could argue that IT managers may overestimate e-business value. To test this possible bias, the sample was divided into two groups: responses from IS managers (heads of IT/DP and other IT senior managers) versus responses from non-IS managers (owners, managing directors and others). One-way ANOVA was used to compare the means of factor scores between the two groups. No significant differences were found, suggesting that the role of the respondents did not cause any survey biases.

3.2 Measures of variables

Measurement items were introduced on the basis of a careful literature review. Confirmatory factor analysis (CFA) was used to test the constructs. Based on the CFA assessment, the constructs were further refined and then fitted again. Constructs and associated indicators, as well as prior research support, are listed in the Appendix and discussed below.

Internet resources: This construct represents the adoption of physical Internet technologies. In this sense, respondents were required to assess the presence various Internet technologies. These indicators were obtained from the literature on e-business adoption [31, 44, 57, 58].

E-innovation: This construct assessed whether firm made innovations in product/services and processes directly related to or enabled by Internet-based technology. Indicators were extracted from the literature on e-innovation [1, 23, 29].

E-sales effectiveness: As discussed in section 2, the present research measures the effectiveness of online sales by its impact on the volume of sales, the number of customers, the quality of customer service and the costs of logistics and inventory) for measuring e-business value. It was measured by 5 items following previous literature [44, 54, 55, 57].

3.3 Instrument validation

CFA using AMOS was conducted to assess empirically the above constructs theorized. Multiple tests on construct validity and reliability were performed. Model fit was evaluated using the maximum likelihood (ML) method. The measurement properties are reported below (Table 2):

Construct reliability. All constructs had a composite reliability over the cut-off of 0.70 [47], and also the average variance extracted for all exceeded the preferred level of 0.5 [14].

Content and construct validity. This validity was verified by checking the meanings of indicators and by a careful literature review. Construct validity has two components: convergent and discriminant validity. After dropping insignificant items, all estimated standard loadings were significant, suggesting good convergent validity. To assess the discriminant validity, the Fornell and Larcker's [20] criterion, that average variance extracted for each construct should be greater than the squared correlation between constructs, was used. All constructs met this criterion.

The insignificant p-value ($p = 0.187$) for the chi-square statistics implied good absolute fit. The root mean square error of approximation (RMSEA) was below the cut-off value 0.08 suggested by Browne and Cudeck [8]. Five incremental fit indices were all above the preferred level of 0.9 [21].

In conclusion, the overall fit statistics, validity, and reliability measures allow the confirmation of the proposed constructs.

Factor	*Indicat.*	*Loadings*	*CV (t-value)*	*Composite Reliability*
IR	IR1	0.506	--	SCR = 0.909 AVE = 0.716
	IR2	0.722	11.52 [b]	
	IR3	0.560	10.79 [b]	
	IR4	0.576	10.96 [b]	
EI	EI1	0.700	--	SCR = 0.960 AVE = 0.923
	EI2	0.860	4.855 [a]	
ESE	ESE1	0.655	--	SCR = 0.830 AVE = 0.621
	ESE2	0.827	5.157[b]	
	ESE3	0.683	5.311[b]	

Note. $p<0.05$[a]; $p<0.01$[b]
IR: Internet resources; EI: e-Innovation
ESE: e-Sales effectiveness
Insignificant factors are dropped (IR5 and ESE4)
CV: Convergent validity; SCR: Scale composite reliability
AVE: Average variance extracted; (--): Fixed items in the scale

Table 2: Measurement Model.

4 Empirical Results

In order to test whether E-sales effectiveness is influenced by the level of Internet resources and the level of e-innovation within firms, statistical techniques of group differences were employed. More specifically, the T-test was applied after having checked parametric assumptions as well as homogeneity of group variances (Levene's test of significance >0.05). The sample was divided according to the mean of then Internet resources and the mean of e-innovation constructs, respectively. Internet resources was introduced as a two-level categorical variable, coding whether the firm had Internet resources above the mean (low level of Internet resources firms) or below it (high level of Internet resources firms). E-innovation was introduced as a two-level categorical variable, coding whether the firm had introduced e-innovations above the mean (low level of e-innovation firms) or above it (high level of e-innovation firms). Internet resources were those with Internet resources below the mean. Similarly, firms with high e-innovation were firms with e-innovation above the mean, while firms with low e-innovation were those with e-innovation below the mean.

An examination of the underlying distribution of the variables and the Levene's test of significance ($p> 0.05$) suggested a parametric test would be more appropriate (see table 3). Results showed that the association between Internet resources and e-sales effectiveness was not statistically significant ($p= 0.934$), while e-sales effectiveness was influenced by e-innovation ($p= 0.001$).

Internet Resources (IR)	Mean (ESE)	Levene (Sig.)	T-test (Sig.)
High level IR firms	11.71	0.845	0.934
Low level IR firms	11.66		
E-innovation (EI)	Mean (ESE)	Levene (Sig.)	T-test (Sig.)
High level EI firms	12.01	0.428	0.001
Low level EI firms	9.55		

Table 3: Internet resources, e-innovation and e-sales effectiveness.

5 Discussion

Previous literature concludes that e-business enables and drives significant innovative transformations regarding business models, value propositions, products and services of firms and internal business processes, which can offer substantial benefits This paper examines, grounded in the resource-based view (RBV) firms, whether companies with high level of Internet resources and with high e-innovation are more effective electronically. Moreover, it is intended to offer results more widely applicable than studies of Internet leaders or IT industry companies. In this sense, this study attempts to offer an explanation to why there are cases where firms engage in e-business without deriving any benefits.

The results showed that firms with a high level of Internet resources did not outperformed on e-sales effectiveness. This finding indicates that, since competitors may easily duplicate investments in Internet resources by purchasing the same hardware and software, Internet resources per se do not provide better performance. This can be explained through the RBV, because IT is not considered a resource that is difficult to imitate, since IT is widely available and at declining prices. This result supports the findings of recent research [7] that did not find evidence of a positive link between IT quality and firm performance. Similarly, Powell and Dent-Micallef [40] showed that IT by itself cannot be a source of competitive advantage. Thus, our results extend the conclusion of previous research that technology by itself will rarely create business value.

Moreover, results demonstrated that firms with a high level of e-innovation outperformed on e-sales effectiveness. This finding supports existing empirical research using the RBV domain [6, 41, 43], which found that firms create competitive advantages though intermediary effects, such as IT being embedded in products and services and streamlined business processes, which in turn affect higher levels of firm performance. Findings also support extant literature which concludes that e-business enables and drives significant innovative transformations regarding business models, value propositions, products and services of firms and internal business processes, which can offer substantial benefits [2, 54, 55, 48, 49, 59].

6 Conclusion

In recent years, much debate about the value of IT and e-business has been created, due to the gap between e-business investment and the lack of empirical evidence on e-business value. Thus, today IS researchers face pressure to answer the question of whether and how e-business affects firm performance. Since innovation has become a key factor for increasing the competitiveness of firms and e-business has been proposed as complement to innovation. To respond to these challenges, this paper examines, grounded in resource-based view (RBV) of the firm perspective, whether companies with high level of e-innovation are more effective electronically, which is measured as the effectiveness of online sales. Also, it assesses whether the level of Internet resources is related to e-sales effectiveness. Results indicated that differences of e-sales effectiveness of firms with high and low Internet resources were not statistically significant, while on the contrary firms with a high level of e-innovation outperformed on e-sales effectiveness.

The study provides an important implication for managers. E-business resources are easy to duplicate, and, hence, per se do not provide competitive advantages. Although Internet resources are argued to be valuable, they will rarely lead to superior performance. However, when Internet resources are used appropriately, in combination with other resources, they are expected to facilitate product/service innovation and process innovation. That is, merely having Internet resources may not generate value per se, but if these resources are used in combination with other resources to build IT capabilities such as the e-innovation capability, the output from this type of capabilities, in accordance with the RBV, is business value and effectiveness improvements.

While this study presents interesting findings, it has some limitations which can be addressed in future research. First, the sample used was from Spain. It may be possible that the findings could be extrapolated to other countries, since economic and technological development in Spain is similar to other OECD Member countries. However, in future research, a sampling frame that combines firms from different countries could be used in order to provide a more international perspective on the subject. Second, the e-business value measures are subjective in the sense that they were based on Likert-scale responses provided by managers. Thus, it could also be interesting to include objective performance data for measuring e-business value. Third, the key informant method was used for data collection. This method, while having its advantages, also suffers from the limitation that the data reflects the opinions of one person. Future studies could consider research designs that allow data collection from multiple respondents within firms.

Acknowledgement

We would like to thank e-business W@tch for the support provided.

References

[1] Adamides, E.D. & Karacapilidis, N. (2006). Information technology support for the knowledge and social processes of innovation management. Technovation, Vol. 26, No. 1, pp. 50-59.

[2] Amit, R. and Zott, C. (2001). Value creation in e-business. Strategic Management Journal, 22, pp. 493-520.

[3] Arvanitis, S., (2005). Computerization, Workplace Organization, Skilled Labour and Firm Productivity: Evidence for the Swiss Business Sector. Economics of Innovation and New Technology, 14(4), pp. 225-249.

[4] Barney, J.B. (1991). Firm Resources and Sustained Competitive Advantage. Journal of Management, Vol. 7, pp.99-120.

[5] Barua, A., Konana, P., Whinston, A.B., and Yin, F. (2004). An Empirical Investigation of Net-Enabled Business Value. MIS Quarterly, Vol. 28, No. 4, pp. 585-620.

[6] Bharadwaj, A.S. (2000). A resource-based perspective on information technology capability and firm performance: an empirical investigation. MIS Quarterly, Vol. 24, No. 1, pp. 169-196

[7] Bhatt, G.D. & Grover, V. (2005). Types of information technology capabilities and their role in competitive advantage: an empirical study. Journal of Management Information Systems, Vol. 22, No. 2, pp. 253-277.

[8] Browne, M. W. and Cudeck, R., 1993. Alternative ways of assessing model fit. In K. A. Bollen and J.S. Long, (Eds.), Testing Structural Equation Models (pp. 136-162). Beverly Hills: Sage.

[9] Brynjolfsson, E. and Hitt L. M. (2000). Beyond computation: Information technology. organizational transformation and business performance. Journal of Economic Perspectives, Vol. 14, No. 4, pp. 23–48.

[10] Carneiro, A. (2000). How does knowledge management influence innovation and competiteveness?. Journal of Knowledge Management, Vol. 4, No. 2, pp. 87-98.

[11] Carr, N. (2003). IT doesn't matter. Harvard Business Review, May 2003, pp. 41-49.

[12] Champy J. (2002). Seven steps to X-engineering. Executive Excellence, 19(6), pp. 15-17.

[13] Champy, J. (2002). X-Engineering the corporation: Reinventing your business in the digital age. Warner Books, New York, NY.

[14] Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. Journal of Marketing Research 16(1), pp. 64-73.

[15] Clemons, E. K., Row, M.C. (1991). Sustaining IT advantage: the role of structural differences. MIS Quarterly, Vol. 15, No. 3, pp. 275-292.

[16] Colomo-Palacios R., García-Crespo A., Soto-Acosta P., Ruano-Mayoral M., Jimenez-Lopez D. (2010). A case analysis of semantic technologies for R&D intermediation information management.

International Journal of Information Management, 30(5), pp. 465-469.

[17] Devaraj, S., Krajewski, L. and Wei, J.C. (2007). Impact of eBusiness technologies on operational performance: The role of production information integration in the supply chain. Journal of Operations Management, Vol. 25 No. 6, pp. 1199-1216.

[18] European e-Business Market Watch (2008). The European e-Business Report 2008. Edited by Selhofer, H., Lilischkis, S., Woerndl, M., Alkas, H. and O'Donnell, P. Office for Official Publications of the European Communities, Luxemburg.

[19] Evans, P.B., Wruster, T.S. (1999). Blown to bits: how the new economics of information transforms strategy. Harvard Business School Press: Boston, MA.

[20] Fornell, C. and Larcker, F.D. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. Journal of Marketing Research, 18(1), pp. 39-50.

[21] Gefen, D., Straub, D.W. and Boudreau, M.C. (2000). Structural Equation Modeling and Regression: Guidelines for Research Practice. Communications of the AIS, 4(7), pp. 1-78.

[22] Gonzalvez-Gallego N., Soto-Acosta P., Molina-Castillo F.J., Trigo A., & Varajao J. (2010). El papel de las TIC en el rendimiento de las cadenas de suministro: el caso de las grandes empresas de España y Portugal. Universia Business Review, 28(4), pp. 102-114.

[23] Hamel, G. (2002). Leading the Revolution, Plume, New York.

[24] Hammer M. (1990). Re-engineering work: Don't automate, obliterate. Harvard Business Review, 68(4), pp. 104-112.

[25] Hammer, M., Champy, J. (1993). Re-engineering the corporation: A manifesto for business revolution. Harper Press, New York, USA.

[26] Hempell, T. (2003). Do Computers Call for Training? Firm-level Evidence on Complementarities between ICT and Human Capital Investments. ZEW Discussion Paper No. 03-20, Mannheim.

[27] Hernández-López A., Colomo-Palacios R., García-Crespo A., & Soto-Acosta P (2010). Team software process in GSD teams: a study of new work practices and models. International Journal of Human Capital and Information Technology Professionals, 1(3), pp. 32-53.

[28] Hoopes, D. G., Madsen, T. L. and Walker, G. (Eds.) (2003). Guest Editors' Introduction to the Special Issue: Why is there a Resource-Based View? Toward a Theory of Competitive Heterogeneity. Strategic Management Journal, 24(10), pp. 889-902.

[29] Kessler, E.H. (2003). Leveraging e-R&D processes: a knowledge-based view. Technovation, Vol. 23, pp. 905-915.

[30] Kettinger, W.J., Grover, V., Guha, S. & Segars, A.H. (1994). Strategic information systems revisited: a study insustainability and performance. MIS Quarterly, Vol. 18, pp. 31-58.

[31] Kowtha, N.R. & Choon, T.W.I. (2001). Determinants of website development: a study of electronic commerce in Singapore. Information & management, Vol. 39, No. 3, pp. 227-242

[32] Lederer, A.L., Mirchandani, D.A. & Sims, K. (2001). The search for strategic advantage from the world wide web. International Journal of Electronic Commerce, Vol. 5, No. 4, pp. 117-133.

[33] Mahoney, J.T., Pandian, J.R. (1992). The resource-based view of the firm within the conversation of strategic management. Strategic Management Journal, 13(5), 363-380.

[34] Mata, F. J., Fuerst, W. L. & Barney, J. B. (1995). Information technology and sustained competitive advantage: a resource-based analysis. MIS Quarterly, Vol. 19, No. 4, pp. 487-505.

[35] Meroño-Cerdan, A.L. & Soto-Acosta, P. (2007). External web content and its influence on organizational performance. European Journal of Information Systems, Vol. 16, No. 1, pp. 66-80.

[36] Meroño-Cerdan, A.L., Soto-Acosta. P. & Lopez-Nicolas, C. (2008). How do collaborative technologies affect innovation in SMEs?. International Journal of e-Collaboration, Vol. 4, No. 4, pp. 33-50.

[37] Mukhopadhyay, T., Kekre, S. & Kalathur, S. (1995). Business value of information technology: a study of electronic data interchange. MIS Quarterly, Vol. 19, No. 2, pp. 137-156.

[38] Organization for Economic Cooperation and Development (OECD) (2009). Communications Outlook 2009. Paris, France.

[39] Peteraf, M. A. (1993). The cornerstones of competitive advantage: a resource-based view. Strategic Management Journal, 14, pp. 179-191.

[40] Powell, T. C., & Dent-micallef, A. (1997). Information technology as competitive advantage: the role of human, business, and technology resources. Strategic Management Journal, Vol. 18, No. 5, pp. 375-405.

[41] Ravichandran, T. and Lertwongsatien, C. (2005). Effect of Information Systems Resources and Capabilities on Firm Performance: A Resource-Based Perspective. Journal of Management Information Systems, Vol. 21, No. 4, pp. 237-276.

[42] Rumelt, R.P., Schendel, D., & Teece, D.J. (1991). Strategic management and economics. Strategic Management Journal, 12, pp. 5-29.

[43] Santhanam, R. & Hartono, E. (2003). Issues in linking information technology capability to firm performance. MIS Quarterly, Vol. 27, No. 1, pp. 125-153.

[44] Soto-Acosta. P. & Meroño-Cerdan, A.L. (2008). Analyzing e-Business value creation from a resource-based perspective. International Journal of Information Management, Vol. 28, No. (1), pp. 49-60.

[45] Soto-Acosta P., Martinez-Conesa I., Colomo-Palacios, R. (2010). An empirical analysis of the

relationship between IT training sources and IT value. Information Systems Management, 27(3), pp. 274-283.

[46] Steinfield, C., Mahler, A. & Bauer, J. (1999). Electronic commerce and the local merchant: opportunities for synergy between physical and Web presence. Electronic Markets, Vol. 9, pp. 51-57.

[47] Straub, D.W. (1989). Validating Instruments in MIS Research. MIS Quarterly, 13(2), pp. 147-169.

[48] Tavlaki, E. and Loukis, E. (2005). Business Model: A prerequisite for success in the network economy. In Proceedings of 18th Bled eConference - eIntegration in Action proceedings 2005, June 6-8, Bled, Slovenia.

[49] Timmers P. (1998), "Business Models for Electronic Markets", Electronic Markets, Vol. 8 No. 2, pp. 3-8.

[50] Trigo A, Varajão J., Soto-Acosta P, Barroso J, Molina-Castillo FJ, Gonzalvez-Gallego N (2010). IT Professionals: An Iberian Snapshot, International Journal of Human Capital and Information Technology Professionals, 1(1), pp. 61-75.

[51] Turban, E., Lee, J., King, D., McKay, J. and Marshall, P. (2008). Electronic Commerce – A Managerial Perspective 2008 – Fifth Edition, Pearson Prentice Hall, New Jersey, USA.

[52] Wernerfelt, B. (1984). A resource-based view of the firm. Strategic Management Journal, 5, pp. 171-180.

[53] Wu, F., Mahajan, V., & Balasubamanian, S. (2003). An analysis of e-business adoption and its impacts on business performance. Journal of the Academy of Marketing Science, Vol. 31, No. 4, pp. 425-447.

[54] Wu, J. H. and Hisa, T. L. (2004). Analysis of E-commerce innovation and impact: a hypercube model. Electronic Commerce Research and Applications, 3, pp. 389 – 404.

[55] Wu, J. H. and Hisa, T. L. (2008). Developing E-Business Dynamic Capabilities: An Analysis of E-Commerce Innovation from I-, M- to U-Commerce. Journal of Organizational Computing and Electronic Commerce, 18, pp. 95 – 111.

[56] Zhu, K. & Kraemer, K. (2002). E-commerce metrics for net-enhanced organizations: assessing the value of e-commerce to firm performance in the manufacturing sector. Information Systems Research, Vol. 13, No. 3, pp. 275-295.

[57] Zhu, K. & Kraemer, K. (2005). Post-adoption variations in usage and value of e-business by organizations: cross-country evidence from the retail industry. Information Systems Research, Vol. 16, No. 1, pp.61-84.

[58] Zhu, K., Kraemer, K. & Xu, S. (2003). Electronic business adoption by European firms: a cross-country assessment of the facilitators and inhibitors. European Journal of Information Systems, Vol. 12, No. 4, pp. 251-268.

[59] Zwass, V. (2003). Electronic Commerce and Organizational Innovation: Aspects and Opportunities. International Journal of Electronic Commerce, 7(3), pp. 7-37.

Appendix. Measures

Constructs & Indicators		*Description*	*Literature support*
IR	IR1	Does your company have a website? (Y/N)	[31, 44, 57, 58]
	IR2	Does your company use an Intranet? (Y/N)	[31, 44, 57, 58]
	IR3	Does your company use an Extranet? (Y/N)	[31, 44, 57, 58]
	IR4	Does your company use a LAN? (Y/N)	[31, 44, 57, 58]
	IR5	Does your company use a WAN? (Y/N)	[31, 44, 57, 58]
EI	EI1	Have any of your product/service innovations over the past 12 months been directly enabled by Internet-based technology? (Y/N)	[1, 23, 29]
	EI2	Have any of your process innovations over the past 12 months been directly related to or enabled by Internet-based technology? (Y/N)	[1, 23, 29]
ESE	ESE1	What effect has selling online on your sales? (1-5)	[44, 54, 55, 57]
	ESE2	What effect has selling online on the num. of customers? (1-5)	[44, 54, 55, 57]
	ESE3	What effect has selling online on the quality of your customer service? (1-5)	[44, 54, 55, 57]

Note. IR: Internet resources; EI: e-Innovation;
ESE: e-Sales effectiveness
(1-5), five-point Likert-type scale; (Y/N), dummy variable

Message-Optimal Algorithm for Detection and Resolution of Generalized Deadlocks in Distributed Systems

Selvaraj Srinivasan and R. Rajaram
Department of Information Technology
Thiagarajar College of Engineering
Madurai, 625015, India
E-mail: ssnit@tce.edu, rrajaram@tce.edu

Keywords: distributed deadlock, generalized model, deadlock detection, wait-for graph, deadlock resolution

Received: October 15, 2010

In this paper, we present a new algorithm to detect and resolve generalized deadlocks in distributed systems. The algorithm constructs a distributed spanning tree by diffusing probes along the edges of the Wait-For Graph (WFG) and collects a reply that carries the dependency information of processes to determine a deadlock. Unlike the previous algorithms, it performs reduction whenever it receives a reply from an active process. Moreover it isolates termination detection from deadlock detection, and terminates the execution once it detects a deadlock. It has a worst-case time complexity of d+2 and message complexity of e+2n; where n is the number of nodes, e is the number of edges and d is the diameter of the WFG. Correctness proof and performance analysis for the algorithm are also provided. Furthermore, it minimizes the message length and message overhead associated with deadlock resolution as compared with the existing algorithms.

Povzetek: Članek preučuje problem smrtnega objema v porazdeljenih sistemih.

1 Introduction

In a distributed computing environment, if a process needs a resource on the remote site for its computation, it sends a request message to the desired site. If the resource is available, it will be granted to the requesting process immediately; otherwise, the requesting process waits indefinitely until its request is granted. This will lead to a deadlock in distributed systems where a set of processes wait indefinitely for each other to satisfy their requests. Since deadlock reduces the resource availability and throughput, it should be detected and resolved promptly. However, deadlock is very difficult to detect as well as resolve in distributed systems due to the presence of multiple sites. In general, the interdependency among the distributed processes is modeled by a directed graph known as the Wait-For Graph (WFG) [1,2]; where each node represent a process and an edge from a node 'i' to node 'j' indicates that process 'i' is requested a resource from process 'j' and process 'j' is not granted a resource to process 'i'. A deadlock is defined differently depending upon the underlying resource request model such as Single-Resource model, AND model, OR model and P out-of Q model [1,7]. In the Single Resource and AND model, a process needs all requested resources to continue its execution. Hence, the presence of cycle in the WFG implies a deadlock. In the OR model, a process proceeds the execution only if any of the requested resource is granted. Therefore, the presence of knot is necessary to determine OR deadlock. In the P out-of Q model, a process makes requests for Q resources and remains blocked until it is granted any P resource. Since AND and OR model are the special case of P out-of Q model, it is also referred as generalized request model. A generalized deadlock corresponds to a deadlock in the generalized request model. The generalized request model is quite common in many domains such as resource management in distributed operating systems, communicating sequential processes and quorum consensus algorithms in distributed databases [11,12,16]. A cycle in the WFG is necessary but not sufficient condition whereas a knot is sufficient but not necessary condition for a generalized deadlock. Since detection of generalized deadlock requires the detection of a complex topology in the WFG, only few generalized deadlock detection and resolution algorithms [4,5,7,8,10,12,15,16] have been proposed in the literature. Most of them have used the diffusion computing technique [1] in which a distributed computation is initiated by a single node and joined by other nodes only after receiving a message. The generalized deadlock detection algorithms are grouped into two categories namely centralized and distributed algorithms based on the existence of the WFG. In the centralized algorithms, the initiator maintains the entire information to determine a deadlock whereas in the distributed algorithm the information is spread across multiple sites.

1.1 Literature survey

In general, the distributed algorithms [4,5,7,10,12] have used 'record and reduce' principle to detect the generalized deadlocks. According to the technique, the algorithm records the consistent snapshot of distributed

WFG and performs reduction later to determine a deadlock. The algorithm proposed by Bracha and Toueg [4] consists of two phases. In the first phase, the initiator records a snapshot by propagating the probes along the edges of the WFG. In the second phase, the algorithm simulates the granting of resources to determine a deadlock. The second phase is nested within the first phase. It exchanges 4e messages in 4d time units, where e is the number of edges and d is the diameter of the WFG. By following the approach in [4], Wang.et.al [5] developed another algorithm in which an explicit termination technique is used to detect the end of the first phase. The second phase begins only after the first phase is finished. Although it reduces the time complexity of [4] into 3d+1, it needs 6e messages to detect a deadlock. Moreover, both [4] and [5] have failed to resolve deadlocks. The algorithm in [10] records as well as reduces the WFG simultaneously to determine a deadlock. It records a consistent snapshot of distributed WFG in the outward sweep and reduces in the inward sweep in a single phase. It uses 4e-2n+4l messages in 2d time units to find out whether the initiator is deadlocked, where n is the number of nodes and l is the number of leaf nodes in the WFG. However it deals with the complications introduced because the reduction of node in the inward sweep can begin much before the state of all WFG edges incident at that node have been recorded in the outward sweep. And it needs O(e) messages to resolve deadlocks. The algorithm in [11,12] uses lazy evaluation technique by which the reduction of a process in a snapshot is delayed until the initiator detects the termination. Although it minimizes the message complexity into 2e as with the previous algorithms, it uses variable sized messages with the length of O(e) . Since the initiator knows the resource requirement of all deadlocked processes, it minimizes the message overhead associated with deadlock resolution unlike in [10]. The algorithm in [19] achieves the time and message complexity of [12] with fixed sized messages. Unlike [12], it performs the reduction before the initiator terminates the execution of the algorithm. In general, distributed algorithms require two or more rounds of message transfer along the entire edges of the WFG. Hence, they need at least 2d time units to detect deadlock in the worst case.

In the centralized algorithms [8,15,16], the initiator maintains the Local Wait-For Graph (LWFG) to detect a deadlock. The initiator of the algorithm in [8] collects a reply from each process in its reachable set exactly once. Based on the information in replies, it incrementally constructs the WFG locally to determine a deadlock. It needs only 2n messages with the length of O(n) to find out deadlock. However, it has a time complexity of 2d like the distributed algorithms. In addition, it may abort nodes that are not deadlocked and needs O(n) messages to resolve deadlocks. The algorithm in [15] constructs the LWFG by using the ancestor-descendent relationship between the processes in replies. It uses less than 2e messages in 2d time units to detect a deadlock. It reduces the message length into O(e-n+m) where m indicates the number of nodes that are not associated with any non-tree edges in the spanning tree induced by the algorithm. However, it needs additional technique to assign a unique path string to each node in the WFG and to interpret the path strings for constructing LWFG at the initiator. In contrast to [8], it resolves all deadlocks reachable from the initiator. The initiator of the algorithm in [16] collects the dependency information of all nodes to determine the deadlocked processes. It has almost half the time complexity of previous algorithms to detect a deadlock. It needs less than 2e messages in d+2 time units for detecting all deadlocked processes. However it needs additional technique to optimize the message length at each node and requires messages with the length of O(d) in the worst case. Hashemzadeh proposed an algorithm [17,18] based on history based edge chasing technique in which the initiator declares a deadlock once it finds its existence in the message. However, it significantly minimizes the message overhead associated with the executions of concurrent instances and deadlock resolution. We do not consider the algorithms based on edge chasing techniques [17,18] in this paper.

We propose a new centralized algorithm to detect and resolve distributed deadlock in generalized model. Our algorithm improves the message complexity and message size of previous algorithms. The initiator of the algorithm constructs the distributed spanning tree (DST) by propagating probes (CALL messages) along the edges of the WFG. Once a process receives the probe, it sends a reply that carries its dependency information directly to the initiator. However, the initiator performs reduction immediately after receiving a reply from an active process and receives a reply at most twice from each node in its reachable set unlike the earlier algorithms. At the end of termination, it declares all the nodes whose resource requirements are not met as deadlocked. We have formally proved the correctness of the proposed algorithm. It has a worst-case time complexity of d+2 time units and message complexity of less than e+2n, where d is the diameter, n is the number of nodes and e is the number of edges in the WFG. Further, it has a data traffic complexity of O(n) in the worst case. Since it selects a victim without using additional messages, it considerably simplifies deadlock resolution.

Although the proposed algorithm have some similarities with [16], it differs from Lee's algorithm [16] and pervious centralized algorithms [8,15] in the following aspects:

1. The algorithm performs reduction whenever it receives a reply form an active node whereas Lee's algorithm [15,16] performs reduction only after it detects the termination.
2. The algorithm in [15,16] uses an explicit mechanism to reduce the message length, whereas the proposed algorithm does not require any additional techniques.
3. The initiator of the proposed algorithm builds a directed spanning tree of the WFG, whereas the initiator of Chen [8] algorithm does not consider any structural property of the WFG.

4. Unlike in [16], the initiator of the proposed algorithm receives a reply from all nodes in its reachable set at most twice.

The rest of this paper is divided into five main sections. In Section 2, we describe the key definitions and assumptions about the system model and the problem definition. In Section 3, we present the algorithm along with an example. In Section 4, we prove the correctness of the algorithm. In Section 5, we analyze the performance of the algorithm and compare it with that of previous algorithms. Finally, we conclude the paper in section 6.

2 System Model

Although we follow the computation model in [4, 8, 10, 12, 16], we describe it for completeness of this paper. The system has 'n' processes and each one has a unique identity. There is a logical channel between any two processes. The processes can communicate only by message passing. The message delays are arbitrary but finite. The messages are delivered to the destination in the same order as the sender sends them. The messages are neither lost nor duplicated and the entire system is fault-free. The messages are grouped into namely computation and control messages in the system. The computation messages including REQUEST, REPLY, CANCEL and ACK are generated as a result of application's execution. However, the control messages including CALL, REPORT and WEIGHT, which will be discussed in the section 3, are generated by the execution of the deadlock detection algorithm. Both computation and control messages are time stamped based on the requesting process's logical lock [3]. Thus the time stamp of ACK or REPLY should be matched exactly with the corresponding REQUEST message.

A process state is active or blocked at any instant. When a process 'i' makes a generalized request and blocks, the unblocking condition of its request is denoted as a function F_i. For example, $F_i=A\wedge(B\vee C)$ denotes that process 'i' requires a resource from process A and a resource from either process B or C. Function F_i is evaluated in the following manner: substitutes true for a node id in F_i if it has received a REPLY, indicating granting of that request from that node; otherwise, substitutes false for it. Then, evaluate the function F_i. A process unblocks when a sufficient number and combination of its requests to make F_i true are granted.

An active process can send both computation and control messages whereas the blocked process can send either control messages or ACK. When process 'i' blocks on p_i out-of q_i requests, it sends REQUEST message to q_i processes in OUT_i. Therefore, OUT_i gives the domain of F_i. Upon receiving a REQUEST message from 'i', process 'j' records <i, t_block_i> in IN_j and sends an ACK message to the sender of the message. If process 'j' is active, it sends a REPLY message to process 'i' and subsequently removes <i, t_block_i> from IN_j. Once a node is unblocked, it withdraws the remaining requests it had sent earlier but not yet granted.

Each process 'i' maintains the following variables to keep track of its state in the WFG. The initial value of each variable is given within parenthesis.

$parent_i$: the process identifier from which 'i' has received the first probe (NULL)

IN_i : the set of processes which are directly blocked on 'i' (φ)

OUT_i : the set of processes for which 'i' is waiting (φ)

F_i : the condition for unblocking a process 'i'

We denote the set of processes in IN_i as predecessors and the set of processes in OUT_i as successors of process 'i'. A blocked process cannot withdraw any one of its requests spontaneously. And it could not abort any requests abnormally. These two assumptions are essential to ensure that the algorithm records a consistent snapshot. We use the terms process and node interchangeably throughout this paper.

2.1 Problem statement

A generalized deadlock exists in the system if the requesting conditions of one or more processes can never be satisfied. The formal description of deadlock is provided as in [16].

Definition 1: Let evaluate(F_i) be a recursive operation evaluated based on the following:

1. evaluate (i) = evaluate (F_i),
2. evaluate (F_i) = true, for active node 'i',
3. evaluate ($P\vee Q$) = evaluate (P) $\vee$ evaluate (Q)
4. evaluate ($P\wedge Q$) = evaluate (P) $\wedge$ evaluate (Q)

where P and Q are nonempty AND/OR expressions of node identifiers. A generalized deadlock exists in the system if and only if the following topology exists in the WFG.

Definition 2: A generalized deadlock is a sub graph (D,K) of WFG (V,E) where

$\forall i\in D(\neq\varphi)$, evaluate ($F_i$) = false,

No message for computation is under transmission between any nodes in D

Therefore, all processes in D are blocked forever and the resource requirement of processes that do not belong to D can be satisfied at any instant. It should be necessary to abort a node in D to resolve a deadlock.

A distributed deadlock detection algorithm should satisfy the following two correctness conditions:

Liveness: If a deadlock exists, the algorithm will detect it within finite time.

Safety: The algorithm does not report any false deadlock.

3 The Proposed Algorithm

In this section, we present the basic idea behind the execution of single instance our deadlock detection algorithm. Then, we present the algorithm formally and provide an example.

3.1 The description

We assume that the initiator 'i' initiates the deadlock detection algorithm. It includes the unblocking function (F_i) in the set UC_{init} and sends CALL message to each one of its successor 'j' in out_i. If node 'j' receives the first CALL message, it becomes the child of the sender and sends REPORT message that carries the unblocking function (F_j) to the initiator directly. Further, it propagates the CALL message to its own successors. However, if a node that has already joined the execution of current instance receives the second and subsequent CALL message, it does send a reply immediately. Those nodes send a WEIGHT message to the initiator only after receiving CALL messages from all its predecessors. Hence, it minimizes the message overhead to detect a deadlock.

Whenever the initiator receives a REPORT message from a blocked node 'i', it includes a tuple (i, F_i, num_pred_i) in the set UC_{init}. At the same time, if it receives a REPORT message from an active node 'i', it includes 'i' in the set A_{init} and attempts to evaluate all unblocking functions in the set UC_{init}. It performs the evaluation in the following manner: Select a tuple (i, F_i, num_pred_i) from the set UC_{init} and check if the node identifiers in the set A_{init} are sufficient to make F_i as true. If it happens, it includes 'i' in the set A_{init} and removes the corresponding tuple from the set UC_{init}. This process is repeated continuously until there is no more unblocking function in the set UC_{init} can be simplified as true. If the algorithm unblocks all nodes in the set UC_{init} during evaluation, it terminates the execution immediately; otherwise, it continues the execution until it detects the termination based on weight distribution technique. Once the algorithm terminates the execution, it declares all the nodes that have not been reduced in the set UC_{init} as deadlocked.

The algorithm detects the termination based on the weight distribution method like in [10,17]. According to the method, the initiator distributes a weight of one to its successors through CALL messages. When a node receives the first CALL message, it distributes the weight in the message among its successors. However, it accumulates the weight in all subsequent CALL messages until it receives the CALL messages from all its predecessors. It then returns the weight to the initiator through a WEIGHT message. It is accomplished as follows. Each node 'i' has a variable 'num_pred_i' that counts its predecessors. Whenever it receives a CALL message, it decreases the count by one. Hence, the num_pred_i at a node becomes zero signifying that it receives the CALL message from all its predecessors. Whenever the initiator receives a WEIGHT message, it sum ups the weights. The algorithm terminates when the weight at the initiator becomes one.

In a dynamic environment, the algorithm may report a false deadlock due to the presence of phantom edges. Let us consider a phantom edge from node 'i' to node 'j'. This implies that when node 'j' receives a CALL message from node 'i', node 'j' has sent a REPLY to node 'i'. In the proposed algorithm, it is resolved by as follows. Whenever node 'j' receives the CALL message from node 'i', it checks whether node 'i' is in IN_j. If $i \notin IN_j$, it sends an ALERT message to the initiator. Upon receiving the ALERT message, the initiator evaluates f_i by substituting j as true. If node 'i' unblocks during the evaluation, it is included in the set A_{init} and initiates the evaluation of other nodes in the set UC_{init}.

3.2 Formal specification

A formal description of the proposed algorithm executed at node 'i' is presented below. The initial value is given inside the parenthesis.

Data Structure of a node 'i'

$parent_i$: node id (NULL); /* a node from which a CALL has been first received */

$weight_i$: float (0); /* the weight value of 'i' */

in_i : set of nodes (IN_i); /*the set of predecessors of 'i' */

out_i : set of nodes (OUT_i); /*the set of successors of 'i' */

f_i : AND-OR Expression (F_i); /*the condition for 'i' to be active */

num_pred_i: integer ($|IN_i|$); /* the number of predecessors of 'i' */

Additional Data Structures at initiator

UC_{init} → a set of unblocking functions which contains tuples of the form (i, f_i, num_pred_i) where f_i denotes the unblocking condition of a node 'i' (φ).

A_{init} → a set of active nodes (φ)

$weight_{init}$ → the accumulated weight value (0)

$victim_{init}$ → the node identifier to be aborted to resolve the deadlock (φ).

Message Formats

CALL(initiator, sender, weight): Sent by node 'sender' carrying the identifier of the initiator and the weight value for the receiver of this message.

REPORT(sender, f_{sender}, num_pred_{sender}): Sent by node 'sender' as a response to first CALL message carrying the unblocking condition and its number of predecessors.

WEIGHT(sender, $weight_{sender}$):Sent by node 'sender' after receiving CALL messages from all its predecessors.

ALERT (sender, $weight_{sender}$): Sent by node 'sender' after receiving CALL messages through a phantom edge.

I. When a node 'i' initiates the algorithm

$initiator_i := i$;

$parent_i := i$;

$UC_{init} := UC_{init} \cup \{(i, f_i, num\_pred_i)\}$;

send CALL(initiator, i, $1 / |out_i|$) to each $j \in out_i$

II. Upon receipt of CALL(initiator, j, $weight_j$) from j

begin

num_pred$_i$ --;
if (parent$_i$ = NULL ∧ j∈in$_i$) **then**
/* Step II.1 */
parent$_i$:=j;
initiator$_i$:= initiator;
send REPORT(i, f$_i$, num_pred$_i$) to initiator$_i$
if (|out$_i$| >0) **then** /* Step II.1.1 */
send CALL(initiator$_i$,i,weight$_j$/|out$_i$|) to each k∈out$_i$
else if (parent$_i$ ≠ NULL ∧ j∈in$_i$) **then** /* Step II.2 */
if (i=initiator) **then** /* Step II.2.1 */
weight$_{init}$= weight$_{init}$ + weight$_j$;
else if (num_pred$_i$ = 0) **then** /* Step II.2.2*/
send WEIGHT(i, weight$_i$) to initiator$_i$
else
weight$_i$:= weight$_i$ + weight$_j$;
else if (j∉in$_i$) **then** /* Step II.3*/
send ALERT(i, weight$_j$) to initiator$_i$
end

III. Upon receipt of REPORT(i, f$_i$,num_pred$_i$) from i :
begin
if (f$_i$ = φ) **then**
A$_{init}$:= A$_{init}$ ∪ {i};
evaluation();
else
UC$_{init}$:= UC$_{init}$ ∪{(i, f$_i$, num_pred$_i$)};
end

IV. Upon receipt of WEIGHT(i, weight$_i$) from i :
begin
weight$_{init}$:= weight$_{init}$ + weight$_i$;
if (weight$_{init}$ =1 ∧ UC$_{init}$!= φ**)** **then**
resolution(); // Declare a Deadlock
end

V. procedure evaluation()
begin
for each i ∈ UC$_{init}$ **do**
begin
if (evalvate(i.f$_i$) = true) **then**
A$_{init}$:=A$_{init}$ ∪ {i};
UC$_{init}$:= UC$_{init}$ – { i ,f$_i$, num_pred$_i$}
if (UC$_{init}$ = φ) **then**
No deadlock; exit;
else
evaluation();
end for
end

VI. procedure resolution()
begin
count:=0;
repeat
for each i ∈ UC$_{init}$ **do**
begin
if (i.num_pred ≥ count) **then**
count:= i.num_pred;
victim$_{init}$:= i.id;
end for
send ABORT to victim$_{init}$;
UC$_{init}$:=UC$_{init}$–{(victim$_{init}$,f$_{victim}$,num_pred$_{victim}$)};
A$_{init}$:=A$_{init}$ ∪ {victim$_{init}$};
evaluation() ;
until (UC$_{init}$ = φ **)**
end

VII. Upon receipt of ALERT(i, j, weight$_i$) from i :
begin
for each k ∈ UC$_{init}$ **do**
if (k.id = i**) then**
k. f$_i$:= k. f$_i$ | $_{j = true}$;
// Substitutes j by true
if (evaluate (k. f$_i$) = true) **then**
UC$_{init}$:= UC$_{init}$ - {k};
A$_{init}$: = A$_{init}$ ∪ {k.id};
evaluation();
weight$_{init}$:= weight$_{init}$ + weight$_i$;
if (weight$_{init}$ =1∧ UC$_{init}$!= φ**) then**
resolution(); // Declare a Deadlock
end

3.3 Example execution

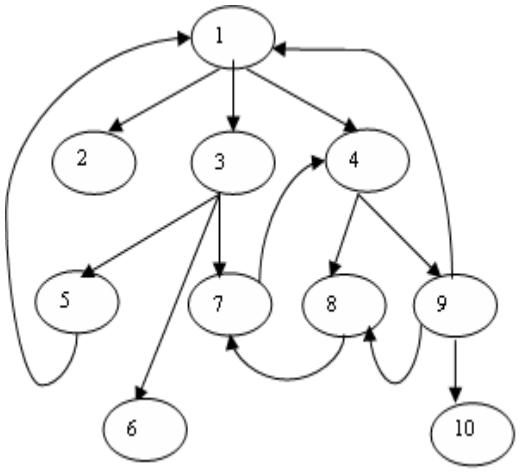


Figure 1: The Wait-For Graph

We illustrate the idea behind our algorithm with the help of an example. Figure 1 shows a distributed WFG that spans 10 nodes labelled from 1 to 10. All the nodes except 2, 6 and 10 are blocked initially. The unblocking conditions of these nodes are as follows: $F_1=(2\wedge3)\vee4$, $F_3=(5\wedge6)\vee7$, $F_4=8\wedge9$, $F_5=1$, $F_7=4$, $F_8=7$ and $F_9=(8\wedge10)\vee1$)

Let us consider node 1 initiates the algorithm and the messages are propagated in such a manner to induce a Breath First Search(BFS) Spanning Tree of the WFG. Figure 2 shows the Directed Spanning Tree, where tree

and nontree edges are indicated by solid and dashed lines respectively.

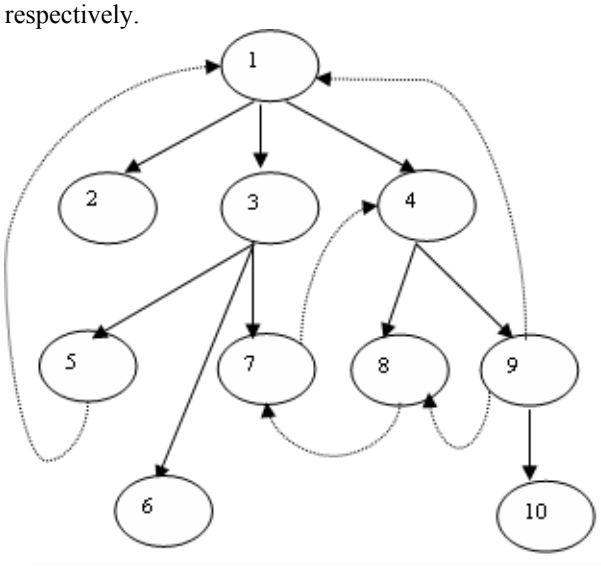


Figure 2: The Distributed Spanning Tree

1. When node 1 initiates the algorithm, it sends CALL(1,1,1/3) to nodes 2,3 and 4 respectively.
2. When node 2 receives the CALL from 1, it sends REPORT(2,ϕ,1) and WEIGHT(2,1/3) to 1.
3. When node 3 receives the CALL from 1, it sends REPORT(3,(5∧6)∨7,2) to 1 and CALL (1,3,1/9) to nodes 5,6 and 7 respectively.
4. When node 4 receives the CALL from 1, it sends REPORT(4,8∧9,2) to 1 and CALL(1.4,1/6) to nodes 8 and 9 respectively.
5. When node 5 receives the CALL from 3, it sends REPORT(5,1,1) and CALL (1,5, 1/9) to 1
6. When node 6 receives the CALL from 3, it sends REPORT(6,ϕ,1) and WEIGHT(6,1/9) to 1.
7. When node 7 receives the CALL from 1, it sends REPORT(7, 4,2) to 1 and CALL(1,7,1/9) to 4.
8. When node 8 receives the CALL from 4, it sends REPORT(8, 7,2) to 1 and CALL(1,8,1/6) to 7.
9. When node 9 receives the CALL from 4, it sends REPORT(9, (8∧10)∨1,1) to 1 and CALL(1,9,1/18) to nodes 1,8 an 10 respectively.
10. When node 1 receives the CALL from 5 through a back edge, it updates $weight_{init}$
11. When node 4 receives the CALL from 7, it sends WEIGHT (4, 1/9) to 1.
12. When node 7 receives the CALL from 8, it sends WEIGHT (7, 1/6) to 1.
13. When the initiator 1 receives the CALL from 9 through a back edge, it updates $weight_{init}$.
14. When node 8 receives the CALL from 9, it sends WEIGHT (8, 1/18) to 1.
15. When node 10 receives the CALL from 9, it sends REPORT(10,ϕ,1) and WEIGHT(10,1/18) to 1.

Figure 3 shows the flow of control messages across the WFG.

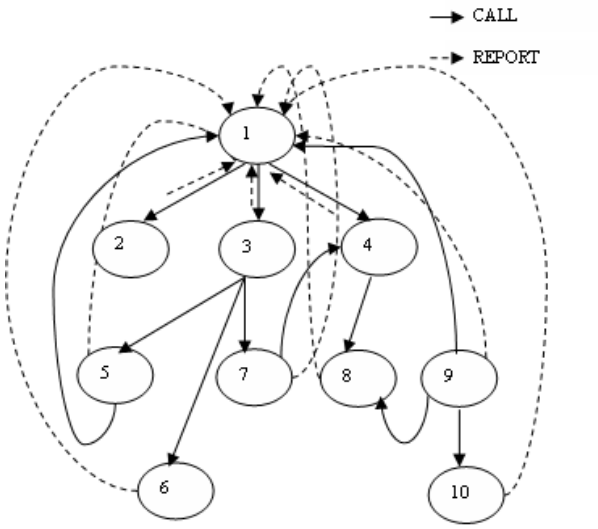


Figure 3: The Message Flow

Whenever the initiator receives the REPORT from nodes 2, 6 and 10, it simplifies the unblocking functions in the set UC_{init}. Finally, it declares the nodes 1,3,4,5,7,8 and 9 as deadlocked nodes.

3.4 Properties of the Algorithm

In this section, we prove the correctness of our algorithm by using several observations (observations 1 - 9) and lemmas (Lemmas 1-4) about the properties of the algorithm.

Observation 1: When the initiator diffuses the CALL messages, it is eventually received by all nodes in its reachable set.

Observation 2: The diffusion of CALL message induces a distributed spanning tree of the WFG.

Observation 3: Whenever a node receives the first CALL message, it propagates the message to each one of its successor.

Lemma 1: If node 'i' receives the CALL message, the unblocking function F_i is sent to the initiator.

Proof: From observation 1, each node that is reachable from the initiator receives the CALL message. Upon receiving the first CALL message, node 'i' sends its unblocking function F_i to the initiator through REPORT message after the execution of Step II.1. Thus, the lemma is proved.

Observation 4: Once a num_pred_i has been recorded in node 'i', it does not change during the execution of the algorithm.

Observation 5: Whenever a node 'i' receives the CALL message through any one of its incoming edge, it decrements num_pred_i by one.

Lemma 2: If a node 'i' sends a WEIGHT message to the initiator then it must have received CALL messages from all its predecessors.

Proof: By observation 5, upon receiving the CALL message from each node $j \in in_i$, node 'i' decreases num_pred_i by one. At the time node 'i' receives the CALL message, if num_pred_i is decremented to 0 then its weight is sent to the initiator through a WEIGHT message by Step II.2.2. Hence, the lemma holds.

Observation 6: Whenever an initiator receives the REPORT message from an active node, it evaluates the unblocking functions in the set UC_{init}.

Definition 3: The initiator reduces a node 'i' iff it has sufficient active nodes in the set A_{init} to simplify F_i as true.

Observation 7: Node 'i' can belong to the set A_{init} only if any one of the following holds.

The initiator receives a REPORT message from node 'i' that contains F_i as true

At the time the initiator evaluates F_i as true during reduction, node 'i' is added to the set A_{init}.

Definition 5: If the initiator is reduced during the evaluation, the algorithm stops the execution.

Observation 8: The weight in a CALL and WEIGHT message is always in transit until they reach the initiator and added to $weight_{init}$.

Definition 3: The algorithm is said to be terminated when $weight_{init} = 1$

Observation 9: When the algorithm terminates the execution, all nodes that is reachable from the initiator either in the set UC_{init} or A_{init}.

Lemma 3: If a deadlock exists in the system, the algorithm will detect it in finite time.

Proof : Assume that a deadlock D exists in the system. The initiator declares a deadlock only if a set $UC_{init} \neq \phi$ after the execution of Step V. Thus, it is sufficient to prove that the initiator has all information about the nodes and their associated edges in D. Let us consider node 'i' in D. It implies that node 'i' has sent F_i to the initiator through a REPORT message by lemma 1. Since node 'i' is blocked forever, the initiator evaluates F_i as false during the execution of Step V at the end of termination. Thus, all nodes in D exist in the set UC_{init}. Let us now assume a edge e = (i,j) and $e \in D$. By lemma 1, node 'i' sends this information to the initiator only after sending a CALL message to node 'j'. Before sending the CALL message, node 'i' must send a REQUEST message to node 'j'. If node 'i' has received a REPLY message from node 'j', an edge e has not been included in F_i. Since both nodes 'i' and 'j' are in D, edge e can not reduced during the execution of Step V in the algorithm. Therefore, UC_{init} contains a deadlock D after the algorithm has terminated. Consequently, the initiator sends an ABORT message to a victim until to resolve a deadlock by Step VI. Thus, the lemma holds.

Lemma 4: If a deadlock is declared, the deadlock exists in the system

Proof: The proposed algorithm reports a deadlock only when $UC_{init}=\phi$. Assume a contrary that the algorithm does not detect a deadlock D. So, it is sufficient to prove that $UC_{init} = \phi$ after the execution of Step V. This reflects the fact that a deadlock D exists in the system and the nodes of D in the set UC_{init} are reduced during the execution of Step V. Let node 'i' be one of such nodes that unblocks first in the set UC_{init}. According to the definition of deadlock, node 'i' is removed from the set UC_{init} only if the unblocking function F_i is simplified as true. It can be possible only when node 'i' had received at least one REPLY from its successors in D at some time T_i. By observation 7, it should happen only before it sends F_i to the initiator. Let node 'j' be one such successor in D that unblocks node 'i'. And node 'j' sends a REPLY to node 'i' only if it has received the REPLY from some of its successors in D before T_i. That is in contradiction with the assumption that node 'i' is the first node that unblocks in the set UC_{init}. Thus is proved.

Theorem 1: The initiator of the algorithm terminates the execution in finite time.

Proof: By step I, the initiator distributes the weight of one to all nodes that are reachable from it through CALL messages. The messages are neither lost nor duplicated according to our network assumptions. From observation 5, for each node 'i' that is reachable from the initiator sends a WEIGHT message that carries the weight value to the initiator after the execution of step II.4. The initiator executes the Step II.2.1 or IV upon receiving the CALL or WEIGHT messages and stops the execution once its weight becomes one. Since the messages transmission takes finite time, the initiator terminates the execution in finite time.

Theorem 2: The algorithm records a consistent snapshot at the initiator.

Proof: Let S be the last snapshot computed by the algorithm, and it contains a edge (p,q). This implies that this dependency relation was included in F_p which has sent by p to the initiator through a REPORT message. Before sending a REPORT message, node 'p' sends a CALL message to all its successors, including node 'q' during the execution of Step I or II. This is so because node 'p' had sent a REQUEST message to node 'q' and $p \in in_q$. This reflects the fact that the edge from p to q indeed exists in the WFG at the time of execution. Hence, the theorem holds.

Theorem 3: The algorithm detects a deadlock if and only if it exists in the system

Proof: Follows from Lemmas 3 and 4.

3.5 Concurrent executions

Since several nodes may block simultaneously, each one of them invokes the deadlock detection algorithm independently. If this happens, a node can be involved in the execution of more than one instance and several initiators may report the same deadlock. In such situations, each instance may select different victims even though a single victim is sufficient to resolve a deadlock. Nevertheless, few instances of the algorithm might be engaged in false deadlock resolution. The various issues associated with the concurrent execution of the algorithm are addressed in [8,11,12,16]. Since the method in [8] needs more messages and prone to useless aborts, we follow the priority based technique in [11,12,16] to handle concurrent executions. According to the method, the algorithm assigns a unique priority to each instance based on its identifier, which comprises the initiator's identifier and the block time / sequence numbers. Since the control messages of every instance carries this label, each instance can be distinguished from others.. When a node involves in the execution of multiple instances, it will support the execution of only high priority instance and suspends the execution of low priority instances.

3.6 Deadlock resolution

The initiator selects a victim that unblocks as many as deadlocked nodes in the set UC_{init} to resolve a deadlock. Then, it sends an ABORT message to the victim directly. It includes the victim into A_{init} and removes the corresponding tuple from the set UC_{init}. It then evaluates the unblocking functions of all nodes in the set UC_{init} and removes the nodes whose unblocking function is simplified as true. If a victim is insufficient to make UC_{init} as empty, it selects another victim. This process continues until UC_{init} is empty. Upon receiving the ABORT message, a node aborts its execution and releases all resources it had acquired earlier. An aborted process restarts its execution as in [11,13]. Thus the proposed algorithm simplifies the deadlock resolution by minimizing the messages and the nodes to be aborted.

4 Performance Analysis

We discuss the performance of the proposed algorithm with respect to time, message and data traffic complexities. The message complexity is the total number of messages exchanged by the algorithm. The time complexity of the algorithm is the time required by the initiator to detect a deadlock. The data traffic complexity defines the total length of data transmitted by the algorithm. The measurements are based on the assumption that the message transmission between any two nodes takes one time unit. We assume that n is the number of nodes, e is the number of edges and d is the diameter of the WFG.

Theorem 4: The algorithm terminates the execution in d+2 time units.

Proof: Whenever a node initiates the algorithm, it sends CALL message to its successors which in turn propagates the message to its own successors. Therefore, the CALL message must travel to the farthest node reachable from the initiator. Let $d_{max} \leq d$ be the maximum diameter of the WFG. Then, the latest time the leaf node of spanning tree receives the CALL message is $d_{max}+1$. Since the leaf node sends a WEIGHT message to the initiator directly, the algorithm will receive all replies at most $d_{max}+2$ time units. Thus the time complexity of the proposed algorithm is d+2 in worst-case.

Theorem 5: The algorithm detects a deadlock using e+2n messages.

Proof: To compute the message complexity, we consider separately each message type.

CALL messages are sent once over any edge of the WFG. Thus, at most e messages are sent totally.

REPORT messages are sent to the imitator over a communication channel directly. Since there is no more than 'n' node, the total number of REPORT messages is bounded by n.

WEIGHT messages are sent to the initiator once by the leaf nodes of spanning tree and thus no more than n-1 of such messages can be sent.

From above, we can conclude that the message complexity at worst case is O(e+2n) messages.

Let us consider the message length of proposed algorithm. Since CALL and WEIGHT messages are fixed sized, we now analyse the length of REPORT message. A REPORT message delivers the unblocking function of a node to the initiator. In the generalized model, the unblocking function of a node 'i' is a AND-OR expression that involves $|out_i|$ node identifiers. In the best case, the unblocking condition can be true and F_i is ϕ. In the worst case, F_i comprises the set of $|out_i|$ node identifiers.

For computational complexity at the initiator, we need to determine computational complexity of two procedures namely evaluation and resolution. The evaluation procedure is executed whenever an initiator receives the REPORT message from an active node. In the worst case, when all nodes are deadlocked, the unblocking functions of all n nodes are in the set UC_{init} and $A_{init}=\phi$. At the time, the algorithm declares the deadlock without evaluating the unblocking conditions and invokes the procedure resolution. In the procedure resolution, a victim is selected, inserted into the set A_{init} and removed from the set UC_{init}. Therefore, the number of processes in the set UC_{init} is reduced at least by one at each execution of resolution. Hence, in the worst case the computational complexity at the initiator is O(n) steps. In contrast, the algorithm in [16] requires $O(n^2)$ steps and the algorithm in [12] needs $O(t^2)$ steps, where t is the number of nodes in the induced spanning tree by those algorithms. However, in the best case ($UC_{init}=\phi$), the local complexity of this algorithm is O(1)

Algorithms	Delay	Number Of Messages	Message Size	Resolution
Barcha et al [4]	4*d*	4*e*	O(1)	no Scheme
Wang et.al [5]	3*d*+1	6*e*	O(1)	no Scheme
Kshemkalyani et.al [10]	2*d*	4*e*-2*n*+2*l*	O(1)	e messages
Kshemkalyani et.al [12]	2*d*	2*e*	O(*e*)	1 message
Brzezinski et.al [7]	4*n*	½ n^2	O(*n*)	no Scheme
Chen et .al [8]	2*d*	2*n*	O(*n*)	3n messages
Soojung Lee [16]	*d*+2	<2*e*	O(*d*)	1 message
Our algorithm	*d*+2	*e*+2*n*	O(*n*)	1 message

Table 1: Performance Comparison

Table 1 compares performance of different generalized deadlock detection algorithms. The message length of O(n) indicates that it consists of all node identifiers in the algorithms [7,8]. And the message used in the algorithms [12,15,16] and the proposed algorithm carries the unblocking functions to the initiator. However, the message length of these algorithms is differed due to the following reason. In the algorithm [16] the unblocking functions of nodes are merged as well as distributed during propagation of probes outward from the initiator whereas in [17], the unblocking function of each node is merged during the propagation of replies backwards to the initiator. As a result, the number of unblocking function in a reply grows as the message goes up in the spanning tree induced by the algorithm [12]. Similarly, if a node has exactly one successor, the number of unblocking conditions in a reply message is at most n-1 in the worst case in [16]. In contrast to [12, 16], the proposed algorithm sends an unblocking function of a node to the initiator disrespect the presence of deadlock and the number of successors of nodes in the WFG. In this conjuncture, the message length of proposed algorithm is a constant.

5 Conclusion

We presented a new algorithm to detect and resolve generalized deadlocks in distributed systems. The initiator of the algorithm collects the unblocking functions of all nodes in its reachable set exactly once. Then it arbitrarily simplifies the unblocking conditions depends on the reply from an active to determine deadlock. We proved the correctness of the algorithm. It has a time complexity of d+2 time units and worst case message complexity of e+2n messages hops delay to detect a deadlock. In addition, it finds out all nodes that are in deadlock with the initiator only if the initiator is deadlocked unlike the earlier algorithms. The performance of the proposed algorithm is better or comparable with the existing algorithms in terms of time, message and data traffic complexities. Furthermore, it simplifies the deadlock resolution by minimizing the additional round of messages. The proposed algorithm is applicable to detect deadlocks in different domains of distributed systems design such as resource management in distributed operating systems, store and forward communication networks, communicating processes and replicated databases.

References

[1] E.Knapp. (1987), Deadlock Detection in Distributed Database Systems, *ACM Computing Surveys*, Vol.19, No. 4 pp.303-327.

[2] M,Singhal.(1989), Deadlock detection in distributed systems. *IEEE Computer*, Vol.22, pp. 37–48.

[3] L.Lamport. (1978), Time, Clocks, and the ordering of events in a distributed systems, *ACM Communications*, vol 21, pp. 558-565.

[4] G.Bracha, and S.Toueg. (1987), A distributed algorithm for generalized deadlock detection, *Distributed Computing*, Vol.2, pp.127–138.

[5] J.Wang, S.Huang, and N.Chen.(1990), A distributed algorithm for detecting generalized deadlocks, Tech. Rep., Dept. of Computer Science, National Tsing-Hua Univ.

[6] W.K.Ng, and C.V.Ravishankar.(1994), On-Line Detection and Resolution of Communication Deadlocks, *Proc. 27th Ann. Hawaii Int'l Conf. System Science*, pp.524-533.

[7] J.Brzezinski, J.M.Helary, M.Raynal, and M.Singhal. (1995), Deadlock Models and a General Algorithm for Distributed Deadlock Detection, *J. Parallel and Distributed Computing*, Vol.31, pp.112-125.

[8] S.Chen, Y.Deng, P.C.Attie, and W.Sun.(1996), Optimal deadlock detection in distributed systems based on locally constructed wait-for graphs, *Proc. Int'l Conf. Distributed Computing Systems*, pp.613–619.

[9] M.Roesler, and W.A.Burkhard.(1989), Resolution of Deadlocks in Object-Oriented Distributed Systems, *IEEE Trans. Computers*, Vol. 38, No. 8, pp.1212-1224.

[10] A.D.Kshemkalyani, and M.Singhal. (1989), Efficient detection and resolution of generalized distributed deadlocks, *IEEE Transactions on Software Engineering*, Vol.20, pp. 43–54.

[11] A.D.Kshemkalyani, and M.Singhal.(1997), Distributed detection of generalized deadlocks. *Proc. 17th Int'l Conf. Distributed Computing Systems*, pp.553–560.

[12] A.D.Kshemkalyani, and M.Singhal. (1999), A One-Phase Algorithm to Detect Distributed Deadlocks in Replicated Databases, *IEEE Trans. Knowledge and Data Eng.*, vol. 11, No. 6, pp. 880-895.

[13] S. Lee. and J.L. Kim.(1995), An Efficient Distributed Deadlock Detection Algorithm, *Proc. of the 15th Int. Conference on Distributed Computing System*, pp.169–178.

[14] S.Lee, and J.L.Kim.(2001), Performance Analysis of Distributed Deadlock Detection Algorithms, *IEEE Trans. Knowledge and Data Eng.*, vol. 13, no. 4,pp. 623-636 .

[15] S.Lee.(2001), Efficient Generalized Deadlock Detection and Resolution in Distributed Systems, *Proc. 21st Int. Conference on Distributed Computing Systems*, pp. 47-54.

[16] S.Lee(2004), Fast, Centralized Detection and Resolution of Distributed Deadlocks in the Generalized Model, *IEEE Trans. on Software Engineering*, Vol. 30, No.9,pp.561-573.

[17] Nacer Farajzadeh, Mehdi Hashemzadeh, Morteza Mousakhani, Abolfazl T. Haghighat. (2005), An Efficient Generalized Deadlock Detection and Resolution Algorithm in Distributed Systems, *Proc of the Fifth Int.Conference on Computer and Information Technology*

[18] Hashemzadeh, M. Farajzadeh, N. Haghighat, A.T. (2006)., Optimal detection and resolution of distributed deadlocks in the generalized model, *Proc of th 14th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*

[19] Srinivasan. S., Rajan Vidya, Rajaram Ramasamy. (2009), An Optimal, Distributed Deadlock Detection and Resolution Algorithm for Generalized Model in Distributed Systems, *CCIS* Vol.40, pp. 70–80.

Aspect-Oriented Reengineering of an Object-oriented Library in a Short Iteration Agile Process

Adrian O'Riordan
Computer Science Department, University College Cork, Cork, Ireland
E-mail: a.oriordan@cs.ucc.ie

Keywords: aspect-oriented software development, reengineering, agile development, software metrics, refactoring

Received: June 24, 2011

Aspect-oriented reengineering aims to modularize crosscutting concerns in an existing system using a new abstraction called an aspect. Code concerns may be tangled and scattered throughout an existing code base thus hampering maintenance. This paper describes the reengineering of an object-oriented software library called GEF using aspect-oriented techniques as an integral activity in an agile process. Graph Editing Framework (GEF) is a medium-sized open source Java library for the construction of graph editing applications. We evaluated both the original and reengineered code by applying a set of appropriate software metrics to measure to what extent aspect-oriented refactoring affected modularity attributes such as coupling, cohesion and complexity. To mirror a real world setting, analysis, re-design, and semi-automated refactoring was performed in three-week iterations typical of agile development using tools freely available on the Eclipse platform. We found that only marginal improvements in modularity were possible in that timeframe and argue that fully-automated aspect mining and refactoring tools are needed to bolster aspect-oriented reengineering.

Povzetek: Članek opisuje predelavo knjižnica z agilnim aspektno usmerjenim programiranjem.

1 Introduction

Aspect-oriented software development (AOSD) promises to improve the modularity of software by the separation of concerns into aspects during system development. This paper presents a study of aspect-oriented reengineering involving the analysis, re-design and refactoring of an existing medium-sized object-oriented library. This was done in a way that was authentic or faithful to industry practice where refactoring is an integral part of agile methods [1]; analysis, re-design, and refactoring are performed in short iterative cycles using tools widely available. Recently developed research tools in automated code transformation and aspect mining that are as yet not common in industry were thus not employed. The time spend in the refactoring phases of development was not changed from that commonly spend in conventional object-oriented refactoring despite the introduction of aspect technology.

We carried out the aspect-oriented refactoring or *aspectization* in a semi-automated manner as part of an agile development process. Agile methods have become popular and already incorporate refactoring in their development process and hence are a suitable approach for introducing aspect-oriented refactoring into a reengineering process. We employed the AspectJ language and associated development tools for refactoring. The two developers were experienced in Java development but only recently familiar with AOSD and AspectJ. We applied a metric suite to both the original and reengineered library, comparing the two sets of results in order to establish any improvements in the areas of reduced complexity, reusability, and maintainability. Conclusions are drawn on the efficacy of this approach.

1.1 Aspect-oriented software development

A reality of modern software is the requirement for continuous change. This change can be instigated externally by the discovery of bugs or changing customer needs or internally to an organization for technological or institutional reasons. Software evolution and maintenance is hampered by the types of decomposition used in coding and design: separation of concerns is a long standing challenge in software engineering [2]. A key problem in software evolution is that software designs tend to have a dominant kind of modularization. This could be feature-based (e.g. transactional) or paradigmatic (object-oriented). But changes that affect a particular feature or concern (such as security for example) may favour an alternate decomposition [3]. In particular, the limitations of object orientation are now becoming more apparent – such as in feature segregating or in applying domain-specific knowledge [4]. AOSD is a technology that addresses the separation of concerns in software at the code level.

The concept of an *aspect* originated at Xerox PARC in the form of aspect-oriented (AO) programming [5], and has gone on to receive significant attention in the software engineering research community [6]. AOSD developed out of work in object-oriented (OO) programming, reflection, and the meta-object protocol [5]. The aim of AOSD is to modularize crosscutting concerns in a system to manage the structural

relationship between representations of a concern. These code concerns or areas of interest can be scattered and tangled (intermixed) throughout the design and implementation; common examples include error handling, logging, and security. Concerns can relate to functional or non-functional requirements. Crosscutting concerns are claimed to make systems difficult to maintain, increase the complexity of the system and reduce the reusability of the code [7]. By applying AO techniques, these concerns can be put into separate modules called aspects, untangling them from each other. Though the AO approach was developed as a programming method, it has been extended to encompass more stages of the software development lifecycle [8].

AOSD tackles areas not addressed in a purely object-oriented OO approach to software development. For existing software to benefit it will be necessary to support the migration of legacy systems to AO solutions. Just as the adoption of OO software development lead to the need to reengineer legacy systems, as for example in [9]; the wider adoption of AOSD will require a similar effort. Laddad advocates a safe adaptation path for AOSD where AO refactoring is applied before AOSD is exploited from a project's inception [10]. There is less experience of applying AOSD in industry and few experience reports published as yet, see Section 5.

1.2 Overview of paper

The paper is structured as follows. Section 2 introduces background material on AO programming, the GEF library and the metric suite. Section 3 presents the reengineering implementation. Section 4 contains the evaluation. The paper finishes with a summary and conclusions.

2 Background

2.1 Aspect-oriented programming in AspectJ

AO programming introduces a number of unfamiliar concepts to programmers. These concepts offer additional functionality to assist with the modularization of crosscutting expressions by encapsulating a concern in one place that would otherwise cross existing units of modularity such as class, subprogram and package. We follow the formulation and terminology of AspectJ throughout this paper.

AspectJ is described as a seamless extension to the Java programming language [14]. AspectJ is free open source software available under an EPL (Eclipse Public License). The major Java extension called an aspect has a Java class style syntax. All legal Java programs are upwardly compatible with AspectJ and all AspectJ programs run on any Java Virtual Machine. The process of linking classes and aspects together is called weaving. In the case of AspectJ, this produces executable bytecode. The bytecode produced by the AspectJ compiler should be comparable to the bytecode produced by a Java compiler used on an equivalent (scattered and tangled) Java implementation [11]. The AspectJ Development Tools (AJDT) [1] provide Eclipse platform based tools for editing, building and debugging AspectJ programs. Whereas Eclipse has good support for AOSD, other IDEs have lagged behind. Alternatives to AspectJ include Hyper/J but AspectJ is by far the most widely deployed example of aspect technology at present.

Here is a brief summary of the operation of AspectJ; see [10] or [12] for a more detailed description. An aspect is a new unit of modularity providing encapsulation and abstraction and allowing tangled or scattered code to be removed from classes while still maintaining overall functionality. *Join points* are events that occur during the runtime execution of a program, for example each time a method or a constructor is called or a variable created. Each such run-time event is a separate join point visible to aspects during program execution.

A *pointcut* is used to identify, by matching, join points of interest. Examples of pointcut designators are *call*, *execution*, *target*, *this*, *get*, *set*, and *args*. There are both named pointcuts and property-based pointcuts that can have wildcard expressions. Pointcut expressions can be created with the &&, || and ! Java logical operators. Pointcuts can also expose contextual information at the join points that they match. Once a pointcut has matched a join point, advice specifies what is to occur.

Here we briefly explain the function of the designators that are used in the example code in Section 3.The *execution* designator picks out each method execution join point and *target* picks out each join point where the target object is of a specified type. The *within* designator limits the lexical scope of the join point and the *this* designator checks runtime type. A *cflow* picks out a join point within the dynamic context of another.

Advice is unnamed as it is implicitly invoked. There are three main types of advice. *Before advice* is advice that executes before a join point whereas *after advice* executes immediately after a join point. *Around advice* runs in place of the join point and is the most flexible type of advice since it can change contextual information. In general terms, an AO programming implementation is characterized by its join point model which dictates the location of joint points (where advice can run), quantifies joint points (how they are matched) and specifies what to do (for example run advice).

AspectJ also has *inter-type declarations (ITDs)*, formerly introductions. ITDs are declarations that affect a program's static structure. They are mainly used to provide definitions of fields and methods within an aspect on behalf of other classes. ITDs can be viewed as enabling open classes allowing structural additions. Note that aspects intercept base code without needing to modify it. This thus makes AO refactoring possible even when the base code cannot be changed.

2.2 Reengineering

Reengineering aims to restructure legacy software. Without comprehensive design specifications maintaining legacy code can be a major burden. Even where extensive documentation exists, reengineering and

software evolution can entail making changes throughout a software system, and has been found to be both difficult and tedious [13]. Reengineering is the examination and alteration of a system to reconstitute it in a new form and the subsequent implementation of this new form [14]. Reengineering generally consists of some reverse engineering or design discovery (often to achieve a more abstract representation) followed by restructuring. Existing OO reengineering does provide some techniques for dealing with tangled code. Refactoring [15] enables OO code restructuring and is an integral part of agile software development methods [16]. Agile methods such as Extreme Programming advocate a culture of continuous reengineering [17].

Many IDEs, such as Eclipse, now have support for a semi-automated refactoring process. Code refactoring includes techniques for renaming, decomposing, composing, relocating, and abstracting program code elements such as identifiers, methods, and classes. Two examples of code refactoring include extracting a method, and converting conditional code into polymorphic code. The aim is to improve quality measures such as "understandability", reusability, and maintainability; not to fix bugs or introduce new features.

But there are limits to the application of OO refactoring and the extent to which conventional refactoring can disentangle code [18]. To give just one indicative example, behaviour can be delegated to a separate class, but new problems can consequently be created because delegation decreases cohesion and adds additional components [19]. In addition, there are scenarios where it is very difficult to separate out a concern using conventional OO techniques, thus impacting ease of maintenance. This may lead to updates being required for unrelated modules for a minor change.

2.3 GEF library overview

The object-oriented software library that was reworked is GEF (Graph Editing Framework), a medium-sized free open source Java library for the construction of graph editing applications [2]. GEF is not a complete drawing program but it supports the construction of custom drawing programs. ArgoUML[3] is a popular open source UML modelling tool built using GEF. GEF (Version 0.12.3) was chosen for the reengineering project for two main reasons: (i) as a medium-sized application it is nontrivial but manageable: and (ii) because it is already well-designed using conventional OO design, any reengineering can focus on the benefits of AO restructuring.

Figure 1 shows screen captures of a simple demo application that uses GEF. GEF is designed using the Model-View-Controller architecture separating the graph models from the display information in Java SWING. GEF was developed to be easy to use and extend without modifying the underlying framework. A flexible Node-Port-Edge graph model is employed for drawing objects. Briefly stated GEF supports selection, grouping, layering and views but not zooming and undo. GEF specifies data as generic properties using JavaBeans. XML-based file formats are employed based on the PGML standard. GEF is a Java counterpart to graph editing libraries such as Unidraw (C++) and HotDraw (Smalltalk).

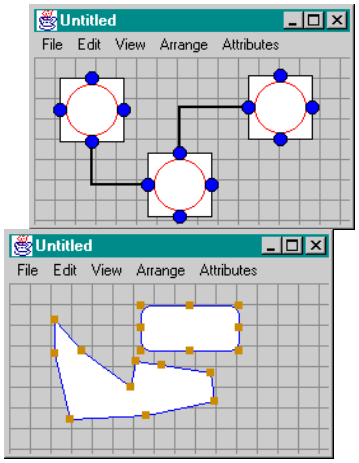


Figure 1: Screenshots of GEF demo application.

The most important classes are now briefly introduced; many of these are referred to in the refactoring in Section 3. Editor is the central class of the Graph Editing Framework. There is one instance of Editor for every diagram that is displayed on the screen. Editor does not handle input events, or modify a diagram; instead it passes events and messages to supporting objects. An Editor has a LayerManager which manages a stack of Layers. Layers contain the objects to be drawn, which are called Figs. Layers group Figs into transparent overlays. Figs are drawable objects that can be shown and manipulated in the Editor such as rectangles, lines, circles, and text. FigGroup is the class for groups of Figs to be treated as single items. When a Fig is selected the SelectionManager holds a selection object. Selections are objects used by the Editor when the user selects one or more Figs. Selections indicate the target of the next command. The behaviour of the Editor is determined by its current Mode. The Editors ModeManager keeps track of all the active Modes. Modes interpret user input events and decide how to change the state of the diagram. Examples of Modes are ModePopup which deals with right mouse button events and shows a popup menu and ModeSelect which allows one to select one or more figs. Cmd is an abstract class for all editor commands. Classes starting with Cmd (CmdSelectAll, CmdCopy, etc.) are classes that define a doIt() method that performs some action in the Editor. In total GEF consists of 302 classes and 30835 lines of code, broken up into 14 different packages. There is little documentation apart from the Javadoc API. Figure 2 shows the major classes of GEF in a reverse engineered MVC architectural design view that serves as the starting point for the re-design.

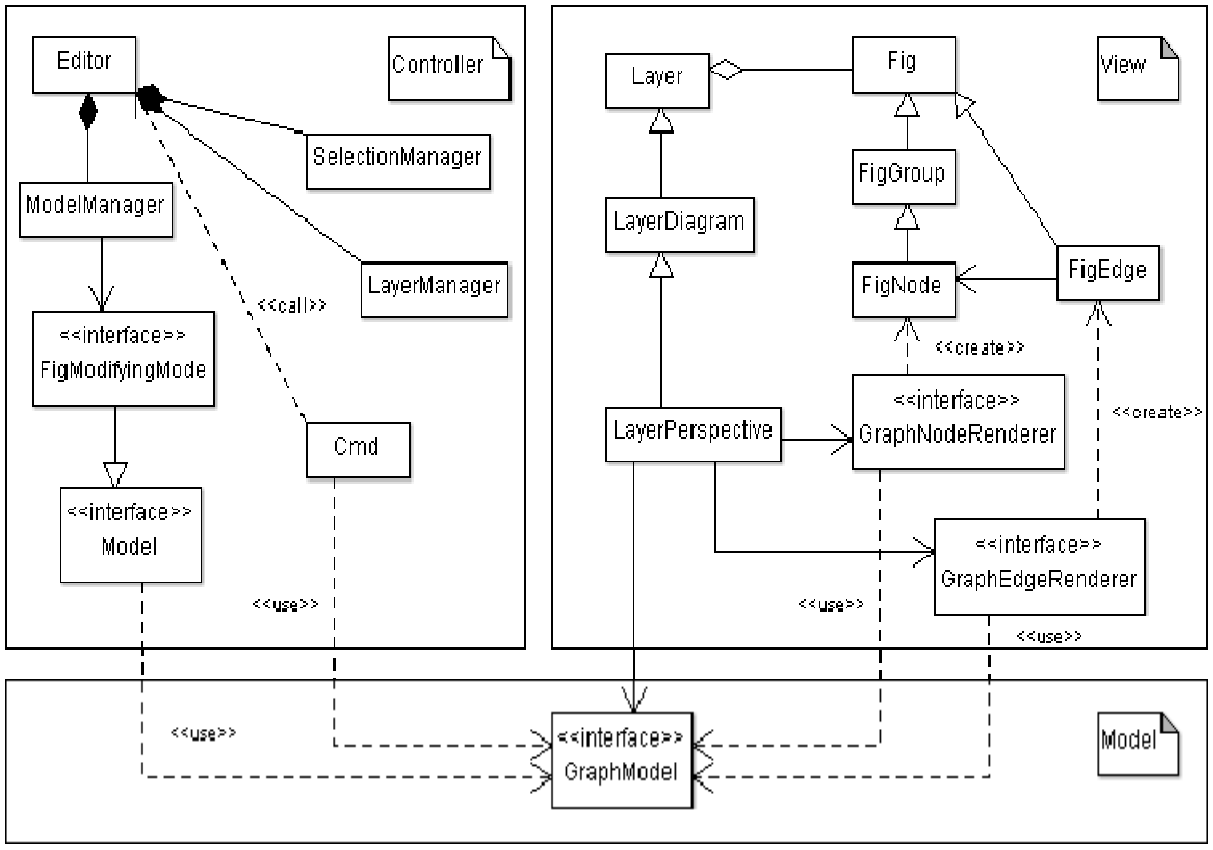


Figure 2. Reverse Engineered MVC Design of GEF.

2.4 Software metrics employed

Metrics for assessing modularity cannot be analyzed independently of other metrics of program quality. For example, a software system implemented as a single module has no inter-module communication but may be deficient in many other regards. Many software metrics have been devised based around the concepts of coupling, cohesion and complexity. In the broadest sense modularity relates to API compatibility, testability, maintainability and extensibility. A summary of the metrics employed in this study is given below. We employed Aopmetrics[4], an open source metrics tool for OO and AO programming. It provides AO extensions to many common OO metrics which can be used to measure the code base and make predictions on reuse and maintenance. Most of the metrics fit into the categories of size metrics, coupling metrics, cohesion metrics and complexity metrics, comparable to the Chidamber and Kemerer (C&K) OO metrics [20]. Additional package dependency and aspect-specific metrics are also present. Note that we use the Java terms *class* and *method* in the following summary descriptions where the Aopmetrics documentation has the terms *module* and *operation*.

Size metrics

Lines of Class Code (LOCC): LOCC gives the total non-blank and non-commented lines of class code.

Complexity metrics

Weighted Operations per Module (WOM): WOM counts the number of methods in a given class, capturing the internal complexity of a class which is an indicator of how much time and effort is required to maintain the class. Classes with a large number of methods may be too complicated or very application specific thus limiting reuse. Response for a Module (RFM): RFM of a class is the number of methods and advices that potentially can be executed in response to a message received by the class. If a large number of methods can be invoked in response to a message, the testing and debugging of the class becomes more complicated.

Coupling metrics

Coupling on Method Call (CMC): CMC is the number of classes or interfaces declaring methods that are possibly called by a given class. Usage of a high number of methods from many different classes indicates that the function of the given class cannot be easily isolated from the others. Coupling between Modules (CBM): CBM is the number of classes/aspect or interfaces declaring methods or fields that are possibly called or accessed by a given class. Excessive coupling between classes is detrimental to modular design and prevents reuse. Depth of Inheritance Tree (DIT): DIT is the length of the longest path from a given class/aspect to the class/aspect hierarchy root. The deeper a class is in the hierarchy, the greater the number of methods it is likely to inherit, making it more complex to predict its behaviour.

Afferent Coupling (Ca): Ca measures the number of classes outside a package that depend on classes inside the package [21]. Efferent Coupling (Ce): Ce measures

the number of classes inside a package that depend on classes outside the package.

Cohesion metrics
Lack of Cohesion in Operations (LCO): LCO measures the number of methods within a class that access one or more of the same attributes. Low LCO is desirable.

Package dependency metrics
Normalized Distance from Main Sequence (D): D is the distance of a package from the idealized line *Abstractness + Instability = 1* where *Abstractness* is defined as the ratio of the number of abstract classes to the total number of classes in the package and *Instability* is the ratio of efferent coupling (coupling outside package) to the total coupling. D is an indicator of a package's balance between abstractness and stability. This metric has a range $0 < D < 1$, where a 0 indicates ideal package design.

Aspect-oriented metrics
Crosscutting Degree of an Aspect (CDA) CDA is the number of classes affected by the pointcuts and by the inter-type declarations in a given aspect. CDA measures all classes possibly affected by an aspect. High values of CDA are usually desirable.

3 Reengineering the GEF Library

3.1 Adoption risks and process overview

There is significant adoption risk associated with AO technology: (i) lack of tool support; (ii) lack of education; (iii) implementation issues; (iv) unpredictable behaviour due to code injection; and (v) security issues [14]. Many of these issues will dissipate as tools and methods mature and gain wider acceptance. Issue (iv) is of particular relevance to AO reengineering as existing code bases with agreed upon contracts can be altered. Supporting processes and techniques, such as embodied in test-driven development help ensure unanticipated behaviour is not introduced. Introducing any major new technology has been found to cause an initial decrease in programmer productivity [22]. Laddad in [10] recommends a cautionary approach for AOSD adoption first employing simple AO techniques for common concerns such as logging and exception handling, to be followed by the more complex techniques for trickier concerns. Applying AO techniques to legacy systems can pose difficulties for various reasons: large code size, lack of documentation, complexity and inconsistencies of implementation and the need to preserve behaviour. A recent review endorses an incremental adoption path [23].

As yet there is no established process for software reengineering. Organisations have typically adapted their standard development process; for example, NASA [24]. We employed an agile process of short development iterations where refactoring is a major component. Adopting an agile approach, two engineers worked in four approximately three-week, development iterations consisting of analysis, design discovery and two iterations of AO code refactoring. Agile methods, such as Scrum and Extreme Programming, use refactoring to improve software quality and enhance project agility. The typical approach in agile development is to first write tests, by means of automated unit testing, which are subsequently used to verify that code transformations are behaviour preserving.

A refactoring process suitable for reengineering is described by Kataoka et al., where refactoring is done in iterations or "clumps" [25]. Pizka took a similar approach of short iterations of discovery, application and test [26]. In particular, similar approach we followed was: (i) Identify code to be refactored; (ii) Determine which changes to apply and to where; (iii) Write tests; (iv) Apply refactoring; (v) Assess effects and check that change is behaviour preserving. This mirrors the typical process required to manage incremental change in OO refactoring: determine change, locate relevant code and determine the change's extent, and carry out impact analysis.

For each new aspect we introduced we examined relevant source code call method calls, constructors, and blocks of code. Once an aspect was introduced, its functionality and purpose were reviewed and adjustments made to further refine how it interacted with and contributed to the existing classes, as well as adjusting the classes in the library that the newly created aspect was now advising. An aspect that starts off interposing a single class can be used to interpose multiple classes. This process was repeated for each crosscutting concern that was identified. Static and dynamic tests were performed to guarantee that software behaviour was preserved, that is to ensure that for the same set of input values, the resulting set of outputs were the same before and after refactoring. Stronger notions of behaviour preservation are needed for domains such as real-time and embedded systems where performance and other properties such as safety could be affected. This was not an issue with GEF.

3.2 Refactoring GEF

We carried out two iterations of AO refactoring using the semi-automated techniques. In the first refactoring iteration we concentrated on basic concerns and solutions within our limited time window. Common AO refactorings that we used included *Extract Feature into Aspect* and *Extract Fragment into Advice* [27]. We measured the restructured software after this phase. In the second iteration we did further refactoring primarily based on the AO implementation of established design patterns and re-ran the metric suite so the modularity of the software was again measured at this final stage. Design patterns are an attractive proposition for software design but while the benefits to the design are well documented [28], their implementations "tend to vanish in the code" [29], failing to "capture the concern explicitly" in the code [30]. AO implementations of design pattern retain explicitness while offering the desired benefit.

AO code refactoring differs from and is more pervasive than he conventional OO refactoring techniques mentioned in the introduction [31]; for example, while extract method simply moves code into a new method replacing it with a method invocation, using aspects you can take an additional step and take out those invocations in the source code altogether. You can make changes not possible with just a Java compiler such as moving out try/catch blocks into a separate aspect. Many of the guidelines and practices in OO refactoring, carry forward to AO refactoring. As stated in [32], AO refactoring "augments and not replaces conventional refactoring." AO refactoring techniques have been developed to modularize exception handling, concurrency, lazy initialization, contract enforcement, and a number of other design constraints. Catalogues of AO *refactorings* have been developed [27, 31, 33, 34]. These are often described in the template format popularized by the design pattern community.

We used the aforementioned Eclipse tools, AJDT and JUnit for writing unit tests. The AspectJ graphical structure browser and the Visualizer allowed us to identify concerns without the use of a dedicated aspect mining tool. Tool-supported refactoring can greatly reduces the effort of manually scanning and changing code. We return to the issue of automated aspect mining and automated refactoring in Section 5. Note that we used only a subset of the AspectJ language features. Indeed, as of 2010, most industrial applications of AOSD have used only basic features [23, 35]. We followed the guidelines given by Colyer [16] where pointcuts are named and individual pointcut definitions are kept simple. Named pointcuts can thus be reused. We placed all pointcuts in an aspect next to the associated advice. In AJDT you can handily associate run-time tests with each item of advice.

In particular the following separate of concerns (SoCs) were addressed wherein one or more aspects were introduced to deal with each.

- SoC1: Exception Handling
- SoC2: Logging
- SoC3: Notification Services
- SoC4: Event Handling
- SoC5: Design Pattern Concerns
    - Composite
    - Strategy
    - State

3.2.1 AO refactoring iteration 1

A summary of the five aspects introduced in the first iteration of refactoring are given next. *ThrowableException* is an aspect introduced to deal with SoC1. Following are additional aspects dealing with SoC2, SoC3 and SoC4 in turn.

ThrowableException aspect (SoC 1)
This is the simple aspect. Calls to printStackTrace() are made by some catch clauses in the original code. Such code snippets occur in multiple packages. We created a new package called exception, modularizing the crosscutting code into an aspect called ThrowableException. This will be single aspect instance – by default all aspects are singletons. Pointcut expressions are created matching join points that can occur in the Java source code. Around advice executes at the matched join points. As the program executes, the pointcuts match events in the runtime of the application triggering a stack trace method to execute. This allows duplicated source code to be removed, providing benefits such as improving the readability of the base code, having the exception throwing all in one place, and supporting future additions which may need to implement calls to printStackTrace(). This example uses the *execution*, *target* and *args* pointcut designators. The *args* designator used here captures contextual information, in this case the arguments passed to methods at an execution joinpoint.

```
package exception;

public aspect ThrowableException{

pointcut printingStackTrace(Throwable aCause):
   execution (* printStackTrace()) &&
target(aCause);

// other pointcuts elided

void around(Throwable cause)
:printingStackTrace(cause){
   proceed(cause);
   if (cause != null){
      System.out.println("Caused by:");
      cause.printStackTrace();
   }
}
//other advice elided
}
```

ExceptionHandler aspect (SoC1)
Exception handling occurs throughout a number of classes in the util package of the GEF library. Within this package we have modularized all try-catch clauses into a second aspect called ExceptionHandler. Exception handling also occurs in classes in other packages of the library but its use is applied in an inconsistent manner and so it was not possible to modularize into this aspect. Within the util package there existed a number of try-catch clauses in classes tangled with other logic in the class. We moved all this exception handling code into the aspect.

LoggingCalls aspect (SoC2)
Logging is used by a number of classes in the GEF library for debugging purposes. Logging is not applied uniformly throughout the library but instead is used on an ad hoc basis in a number of different classes.

It was possible to modularize checks that were done before a message was logged. Before a message is logged with debug priority (Log.debug("message")), a check is made to ensure that debug logging is enabled (Log.isDebugEnabled()). If the result of this check is true then the message is logged, if the result is false then logging is ignored. This check occurs in 70 different locations throughout the library, in a different classes and packages. Since this check is not a primary concern of

the classes in which it occurs, it was moved to an aspect. The aspect contains a pointcut that matches any join points that occur when a call is made to Log.debug() in GEF. When a match is made, contextual information is extracted from the join point; the Log object is extracted and made available to the aspect. A check is then made using around advice, which results in control returning to the join point if debug logging is enabled. If debug logging is not enabled, messages are not logged and control returns to the code after the join point.

PropertyChangeHandler aspect (SoC3)

In GEF the Globals class stores global information that is needed by all Editors. Within the Globals.java class, listener notification is implemented. A hashtable is created which keeps track of a number of PropertyChangeListeners for Figs. It allows for PropertyChangeListeners to be added to Figs. Any changes to the properties of a Fig will result in a notification being sent to its listeners. A Fig can have up to four listeners. There are five methods implemented in the Globals class that manage these propertyChangeListeners.

The methods for managing the properties are not scattered across the library but we decided to modularize these methods since they are specific to Figs and are not the primary concern of the Globals class. By modularizing them into an aspect, PropertyChangeHandler, the code in the Globals class becomes less complex and more robust if changes need to be made to the way listeners are handled.

```
public aspect PropertyChangeHandler{

private static Log Globals.LOG =
    LogFactory.getLog(Globals.class);
private static Hashtable Globals._pcListeners =
    new Hashtable();
private static PropertyChangeListener
    Globals.universalListener = null;
public static int Globals.MAX_LISTENERS = 4;

    public static void
Globals.addPropertyChangeListener (Object
    src, PropertyChangeListener l){
    PropertyChangeListener listeners[]=

(PropertyChangeListener[])_pcListeners.get(src);
if (listeners == null){
    listeners = new
    PropertyChangeListener[MAX_LISTENERS];
       _pcListeners.put(src, listeners);
}
for (int i = 0; i < MAX_LISTENERS; ++i)
    if(listeners[i] == null) {
       listeners[i] = l;
       return;
    }
}

public static void
Globals.addUniversalPropertyChangeListener
    (PropertyChangeListener pcl) {
    universalListener = pcl;
}

public static void
Globals.removeUniversalPropertyChangeListener(){
         // code cut for brevity
}

public static void
Globals.firePropChange(Object src, String
propName, boolean oldV, boolean newV){
    firePropChange(src, propName, new
       Boolean(oldV), new Boolean(newV));
}
// overloaded methods cut for brevity
}
```

This example also shows how the observer pattern is quite naturally implemented in AO programming. Also an alternative option, OO refactoring involving delegation, is not an attractive option here because of the added level of indirection and complexity. The trade-off between using and not using inheritance or delegation is an on-going area of debate. Empirically measuring generalization costs against reuse savings has proved difficult. An interesting proposed solution involves a cost-benefit approach to develop a suitable metric [36].

UseActionEvents aspect (SoC4)

There are classes in the GEF library, UseReshapeAction, UseResizeAction and UseRotateAction, which implement almost identical event listening methods. These classes deal with allowing an Editor to perform certain actions on groups of objects that are currently selected. These actions are a resize action, rotate action and reshape action. Since this is an area of the library where there may possibly be future additions of new classes that provide additional actions, we modularized this duplicated code which is crosscut among these classes into an aspect.

A new aspect called UseActionEvents was created which matches the execution of any actionPerformed() methods in the above classes. When the pointcut defined in the aspect matches a call to this method during the runtime execution of the class, control is passed to the aspect, which then executes some event listening logic depending on where the call originated from. Once the aspect is finished executing, control is passed back to the class.

```
public aspect UseActionEvents{

pointcut
handlingActionEvents(UseReshapeAction aReshape):
    execution (public void actionPerformed(..))
&& target(aReshape);

// other pointcuts elided

void around (UseReshapeAction reshape):
    handlingActionEvents1(reshape){
    Editor ce = Globals.curEditor();
    SelectionManager sm =
       ce.getSelectionManager();
    Enumeration sels = ((Vector)
       sm.selections().clone()).elements();

    while (sels.hasMoreElements()) {
        Selection s = (Selection)
           sels.nextElement();
        if (s instanceof Selection &&
        !(s instanceof SelectionReshape)){
           Fig f = s.getContent();
           if (f.isReshapable()){
              ce.damaged(s);
              sm.removeSelection(s);
                  SelectionReshape sr = new
```

```
                    SelectionReshape(f);
                sm.addSelection(sr);
                ce.damaged(sr);
            }
        }
    }
}
//other advice elided
}
```

3.2.2 AO refactoring iteration 2

A summary of the four additional aspects introduced in the second iteration of refactoring are given in the following subsections. The first of these related to the repaint method and addresses SoC3. The remaining three new aspects address SoC5. Code samples of the reengineered library are included for some of these.

Repainting (SoC3)
First we give a brief description of how the repainting of graphical objects takes place in GEF. Mouse events, screen damage, or changes to a figure's boundary necessitate repainting the screen. Damage is stored as a list of rectangles. This is part of the RedrawManager class's responsibility as well as determining the object under a given mouse point. In GEF, a Layer class can dictate the redraw order of a group of Figs. A Layer is responsible for notifying all dependent layers of changes. Different layers can be hidden, locked, or grayed out independently. A complex notification service maintains state. We introduced a new Repaint aspect that provides an aspect-based implementation of this notification mechanism. This is again based on the Observer pattern and operates similar to the PropertyChangeHandler described in Section 3.1. This necessitated moving and reworking code in RedrawManager.

Composite pattern for handling FigGroups (SoC5)
FigGroup has methods that perform various actions, such as setting and removing properties on all of the Figs in a FigGroup. In the original library different iterators process the list of Figs for each of these. We introduce an aspect to perform these updates. Note that the update operation requires contextual information in the form of the particular type of update operation.

```
static aspect UpdateAllFigs{
    pointcut updateOp (FigGroup fg):
    execution(    *    FigGroup.*(..))    &&    this
(FigGroup) && within (FigGroup);

pointcut FigGroupOperation(FigGroup fg):
    cflow (updateOp);

    // advice elided
    }
```

This example uses the *this*, *within* and *cfow* pointcut designators. The *cflow* designator specifies that the pointcut is in the control flow of each join point picked out by the updateOp pointcut. The pointcut expression with the execution designator matches all executions of any FigGroup method.

Strategies for different commands and state pattern for changing behaviour of Editor depending on FigModifyingMode (SoC5)
Depending on the context the various subclasses of Cmd can be used to perform a suitable action. This part of GEF isn't fully developed as operations such as Undo are not supported. During refactoring the various subclasses of Cmd were removed from the code simplifying the source class design by means of an aspect-oriented implementation of the strategy design pattern [33][37]. We attach advice corresponding to each command type as described in [33]. After advice is used to modularize the various states of FigModifyingMode. This has the advantage of localising future changes since this is extensively used.

4 Evaluation

The following sections presents the metrics after reengineering of the library was completed. Due to the large number of classes involved and the scattered nature of some concerns, for each metric we took average values for the entire library, to give an indication of what effect reengineering had on the library as a whole, with the exception of the *Lines of Class Code (LCC)* metric and the *Weighted Methods per Class (WMC)* metric.

4.1 Evaluation results

Table 1 gives the coupling and cohesion results.

Metrics	Original	Re. Iter 1	Re. Iter 2
CMC	3.205	3.140	3.013
CBM	3.246	3.181	3.126
DIT	1.383	1.383	1.383
Ca	14.81	14.67	14.67
Ce	10.90	10.90	10.90
LCO	117.6	117.6	117.9

Table 1: Coupling and cohesion results.

The coupling and cohesion results did not show dramatic changes between the original and the reengineered code, but the changes do give indications of the effect that the introduction of aspects had. Overall, there is a small reduction in coupling. The *Coupling on Method Call (CMC)* metric showed approximately a two and six percent average decrease in coupling for refactoring iteration 1 and 2 respectively. (Aopmetrics gives results to seven digits of precision but in all the tables here these are rounded down to four. The percentage increases and decreases are rounded to the nearest percentage.) The *Coupling between Modules (CBM)* metric showed a small average reduction of two and four percent between the original and reengineered library. *The Depth of Inheritance Tree (DIT)* metric remained the same for both versions of the library due to the fact that the introduction of aspects did not affect the class hierarchy in the way that subclassing would through OO refactoring. This observation has been previously published [38]. There were small reductions in *Afferent Coupling (Ca)* whereas *Efferent Coupling (Ce)* remained the same. There was a slight increase in the *Lack of Cohesion in Operations (LCO)* metric between the original and reengineered library. Generally high cohesion is a desirable property and so a reduction in lack of cohesion would have been the preferred result.

However, the increase is relatively minimal, and since *LCO* is a measure of the number of methods within a class that access one or more of the same attributes, the use of some inter-type declarations in aspects may have contributed to the increase.

Metric	Orig.	Re. Iter 1	Re. Iter 2
LCC	30835	30422	30355
RFM	3.246	3.181	2.952
WOM	7158	7023	7010

Table 2**:** Size and complexity results.

Table 2 has results related to size and complexity. The metrics Weighted Methods per Module (WOM) and Response for a Module (RFM) are good indications of both the internal complexity and overall complexity of classes. The RFM decreased for the reengineered version by approximately four and ten percent which indicates a small reduction in complexity. The LCC metric indicated a small reduction in code size. This small reduction is due to the removal of replicated code into aspects as well as the movement of some methods and fields.

The very slight increase in the LCO metric is not significant because the overall change in this metric was relatively small. Also there are uncertainties with respect to the level of confidence that can be put in this metric due to the varied results it has displayed in other studies; see Section 6. It is best to consider the results of a set of metrics rather than just one metric in isolation. The results obtained for RFM and WOM support claims of a reduction in complexity, which may have a knock on effect for encouraging reuse and simplifying maintenance. The D metric also provides reassurance that the reengineering has not caused any major stability issues in the library.

Table 3 below shows results for package stability and dependency where there was no significant movement.

Metrics	Orig.	Re. Iter 1	Re. Iter 2
D	0.426	0.427	0.427

Table 3: Package dependency results.

The crosscutting degree metric (CDA) displayed in Table 4, is not applicable for purely OO systems but comes into play when aspects have been used.

Metrics	Orig.	Re Iter 2	Re. Iter 2
CDA	0	44	89

Table 4: Aspect-oriented results.

5 Discussion

The use of AO techniques to reengineer the GEF library using semi-automated techniques in a tight timeframe proved only marginally beneficial to the overall design quality of the library in most areas. The results after applying the metrics support AO programming claims of reducing complexity and coupling but only to a small degree. We believe this was due to the fact that only a limited number of refactoring can be achieved in six weeks.

Similar negative results have been obtained from experiments on conventional refactoring; see for example [26], which used a medium sized Java code base and also a tight developer timeframe. Wilkin et al. also report disappointing results [39]. In another refactoring experiment, Bourquin and Kellen [40] note that code size reduced by ten percent but only after seven months of refactoring, though this involved a much larger code base (140 KLOC of Java) but the team size is not specified. Previous experience of more extensive reengineering, where a software system is modified by above 20 to 25 percent, has been found to be counterproductive [41]. Chen et al. has data on the human effort of OO refactoring, although this was restricted to exception handling [42]. 41 man-hours were spent refactoring 14 KLOC of Java with 371 LOC being modified. They deem the effort to be worthwhile based on a cost-benefit analysis calculated as the estimated savings in maintenance cost minus development costs (man-hours by engineer's pay per hour).

Though there are a number of case studies on aspect-oriented refactoring, see Section 6, unfortunately there is little concrete information provided in how many man-hours were involved in the various tasks. This early-stage work has so far, understandable, concentrated on methods and tools.

Some difficulties we encountered while reengineering are worth mentioning. A lot of time was spent analysing and re-designing GEF, for example identifying sites where an AO approach could be taken. Possibly because the system is a library as opposed to an actual application, a lot of classes were already relatively independent and modularized, limiting where aspects could be used. In many applications there are stand-out crosscutting concerns such as database access and security/authentication that are good candidates for AO refactoring. Persistence is another common concern that is amenable to an AO solution [30] that did not feature in the GEF library. In parts of the library it was difficult to cleanly remove all the code associated with some concerns such as logging. During the modularization of exception handling in the util package, additional lines of code and contextual data had to be extracted from the join point into the aspect, which was not ideal.

Here we briefly discuss two limitations of our methodological approach. While we did some we did not do widespread OO refactoring prior to the AO refactoring. It has been stated that initial code restructuring such as via OO refactoring can aid subsequent AO refactoring [43]. Capturing some concerns as aspects may necessitate restructuring of the base code to expose suitable join points. Second, we did not measure stability in the face of actual changes. Greenwood et al. performed an extensive empirical study of design stability in the face of system changes that are typically performed during software maintenance tasks finding that AO implementations tend to have a more stable design than purely OO implementations [44].

Tools to automate AO reengineering have begun to appear but are still at the research stage of development. Aspect mining techniques are vital to automate the aspect discovery phase. Kellens et al. provide a comprehensive survey of emerging aspect mining techniques [45]. Different approaches are being tried to help identify aspect candidates such as text analysis, dynamic program analysis, code slicing and natural language techniques. Research tools such as DynaAMiT, DelfSTof, Dynamo, and AOPMigrator have recently been developed [45][46]. Work is needed to make these more scalable, more usable and more widely known so as to transfer the technology to industry.

Fully-automated refactoring is the second major component needed to enable full automation. In automated refactoring, refactoring consists of program transformations that satisfied specified preconditions. At present AO refactoring is mostly done by hand or in the semi-automated way because of the immaturity of automated AO refactoring support tools and the fact that those that do exist cannot guarantee they are behaviour preserving [45]. IDEs such as Eclipse currently support a user-guided (or semi-automated) approach but a lot of human effort and expertise is still required. Research in fully automating OO refactoring is actively ongoing.

A property of software that can be affected by any type of refactoring is performance. Generally AO programming has been found to have a negligible effect on performance [10]. Some research has even shown unanticipated performance improvements after OO refactoring [47]. We ran the original and refactoring GEF Demo application are there was no noticeable performance differences.

5.1 Related studies of aspect-oriented reengineering

The majority of empirical studies have shown that applying AO concepts to applications can improve modularity and provide benefits in the areas of reduced complexity, maintainability and reusability but most of these studies don't explicitly state how much effort went into the reengineering.

The very small reduction in lines of code we observed is in line with similar studies [48, 49, 50]. Studies of the reengineering of AO software systems, such as those by Kendall [19], have shown improvements in modularization. This study entailed role modelling of intelligent agent protocols and concentrated on refactoring inter-agent communication and agent conversation/negotiation. Note that Kendall's reengineering used both traditional OO refactoring as well as AO refactoring. Work in the areas of exception handling [48, 49] have shown that the use of aspects helped reduce code tangling and loosen class coupling. Unlike our work, these two studies were restricted to one functional area, exception handling. Evaluations of AOP programming for real-time systems [50] also showed improved modularity for crosscutting concerns. Mixed results were obtained in a project reengineering the Hypercast system for multicast overlay networks [51]. The original Java implementation had 300 classes and was redesigned first using common AO programming methods, pointcut descriptions and advice. They found this approach led to programs that were "unnecessarily hard to develop, understand and change." They repeated the experiment with abstract interfaces that expose pointcut descriptors and impose contracts and found this easier and led to a clearer design. Zhang and Jacobson found a 22 percent decrease in coupling in reengineered middleware [52]. A study by Madeyski and Szala was inconclusive [53]. While other studies show a desirable change for the LCO metric [48], there are also studies where lack of cohesion increased [49]. This may indicate limitations of usefulness of this metric in AO systems or possibly calls for modifications on how the metric is calculated.

Using software metrics to mine aspects is a different way of applying metrics to the refactoring process. The explicit use of software metrics to locate problem code for (non-AO) refactoring has been tried [54]. Cole and Borba propose what they call AspectJ laws, a catalogue of code transformations [55].

JHotDraw, a Java version of the HotDraw library mentioned in Section 2.2, has been used as a test-bed for AOSD work. Note that HotDraw is similar to GEF in design, complexity and function. AJHotDraw is an open source AO reengineered version of JHotDraw created to test the feasibility of reengineering legacy code with aspects. Ceccato et al. used JHotDraw to compare aspect mining techniques [56]. A different development process from ours was used, a four step process consisting of mining, exploration, documentation, and refactoring based on so-called crosscutting concern sorts. Hannemann et al. show the viability of a role-based approach to semi-automate AO refactoring by refactoring three different design patterns - observer, singleton and template method – also in JHotDraw [57].

6 Conclusions and Future Directions

Having analyzed the empirical results and reviewed existing research in the area of aspect-oriented reengineering it is clear there is potential in the areas of reducing complexity, maintainability and promoting reuse. There are varying degrees of success depending on the extensiveness of the reengineering and the type of system it is being applied to. The results we obtained from applying a suitable metric suite to both the original library and the reengineered library suggest that the introduction of aspects did show slight improvements in many fundamental measures of software quality in our short iteration approach. The key question is if this improvement warranted the effort. Future work is needed on defining benefit in terms that factor in development costs. Extensive AO re-design may be difficult within or incompatible with the short iterations in the most common agile processes. We conclude that without greater automation in the form of tools and a supportive process, AO reengineering of working OO software in an agile process is hard to justify.

Constraints and limitations of this study where discussed in Section 5. Future work needs to look at issues surrounding the practical application of AO refactoring in agile development including team development, training, tool support, testing, and quality control. Beuche and Beushe highlighted major issues with transferring aspect technology into practice [58] that can serve as a guide to needed work in the area. They state that AO programming has yet to prove its value in terms of making software development cheaper and that AO programming might be useful for certain functions but not all. Ascertaining how AO refactoring can be most judiciously employed and incorporated into existing processes is an important factor. It is also worth noting that AO programming is still little used outside the Java community and large-scale success stories are few; but there are island of success, see [47, 59, 23] for the state-of-the-art in large-scale deployment. For large code bases it can be difficult to balance the amount of time spent investigating areas where AO can be introduced, and the overall benefit gained from doing so. In such cases prior developer knowledge of the system being reengineered could be advantageous to tip the balance in favour of AO refactoring as well as use of the automation tools discussed in Section 5.

Acknowledgments

I would like to thank the MSc student Mark Donnelly who worked with me on the AspectJ coding.

References

[1] Dyba, T., Dingsoyr, T. 2009. What do we know about agile software development? IEEE Software , 26(5), pp.6-9.

[2] Parnas, D. 1972. On the criteria to be used in decomposing systems into modules. Communications of the ACM, 15(1).

[3] Tarr, P., Ossher, H., Harrison, W., Sutton Jr., S.M., 1999. N degrees of separation: multi- dimensional separation of concerns, in: International Conference on Software Engineering, IEEE Press, New York, NY, pp. 107-119.

[4] Elrad, T., Filman, R.E., Bader, A., 2001. Aspect-oriented programming introduction. Communications of the ACM, 44(10), 2001.

[5] Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Videira Lopes, C., Loingtier, J.M., Irwin, J., 1997. Aspect-oriented programming, in: Aksit, M., Matsuoka, S. (Eds.), ECOOP 1997: LNCS, vol. 1241, Springer, Heidelberg, pp. 220 – 242.

[6] Katz, S., Mezini, M., Kienzle J., (Eds.), 2010. Transactions on Aspect-Oriented Software Development VII - A Common Case Study for Aspect-Oriented Modeling. Lecture Notes in Computer Science 6210, Springer, Heidelberg.

[7] Kulesza, U., Sant'Anna, C., Garcia, A., Coelho, R., von Staa, A., Lucena, C., 2006. Quantifying the effects of aspect-oriented programming: a maintenance study, in: Proc. IEEE International Conference on Software Maintenance, pp. 223-233.

[8] Araújo, J., Baniassad, E.L.A., 2007:.Guest editors' introduction: early aspects - analysis, visualization, conflicts and composition. T. Aspect-Oriented Software Development 3: 1-3

[9] Fanta, R., Rajlich, V., 1999. Restructuring legacy C code into C++. in: International Conference on Software Maintenance, IEEE Press, New York, NY, pp. 77-85.

[10] Ladded, R., 2003. AspectJ in Action, Manning, Greenwich, CT.

[11] Kiczales, G., 2004. The AOP report card, Software Development, January, CMP Media.

[12] Colyer, A., Clement, A., Harley, G., Webster, M., 2004. Eclipse AspectJ: Aspect-Oriented Programming with AspectJ and the Eclipse AspectJ Development Tools, Addison-Wesley Professional,.

[13] LaToza, T.D., Venolia, G., DeLine, R., 2006. Maintaining mental models: a study of developer work habits, in Proceedings of the 28th international Conference on Software Engineering (ICSE '06), ACM Press, New York, NY.

[14] Chikofsky, E., Cross, J., 1990. Reverse engineering and design recovery: A taxonomy, IEEE Software, 7(1), pp. 13-18.

[15] Fowler, M., Beck, K., Brant, J., Opdyke, W., Roberts, D., 1999. Refactoring: Improving the Design of Existing Code, Addison-Wesley Professional, Boston, MA.

[16] Martin, R.C., 2002. Agile Software Development: Principles, Patterns, and Practices, Prentice Hall, Upper Saddle River, NJ.

[17] Beck, K., 2000. Extreme Programming Explained: Embrace Change, Addison Wesley.

[18] Lippert, M., Roock, S., 2006. Refactoring in Large Software Projects: Performing Complex Restructurings Successfully, Wiley.

[19] Kendall, E.A., 2000. Reengineering for separation of concerns, in: Tarr, P., Finkelstein, A., Harrison, W., Nuseibeh, B., Ossher, H., Perry, D. (Eds.), Workshop on Multi-Dimensional Separation of Concerns in Software Engineering at ICSE 2000.

[20] Chidamber, S.R. Kemerer, C.F. 1996. A metric suite for object-oriented design, IEEE Transactions on Software Engineering, 20(6), pp. 476–493.

[21] Martin, R.C., 1994. OO design quality metrics: an analysis of dependencies, in: Workshop Pragmatic and Theoretical Directions in Object-Oriented Software Metrics, OOPSLA 1994, ACM Press, New York, NY.

[22] Weinberg, G., 1997. Quality Software Management: Anticipating Change, 4, Dorset House, New York, pp. 13-20.

[23] Rashid, A., Cottenier, T., Greenwood, P., Chitchyan, R., Meunier, R., Coelho, R., Sudholt, M., Joosen, W., 2010. Aspect-Oriented Software Development in Practice: Tales from AOSD-Europe, IEEE Computer, 43(2), pp.19-26.

[24] Rosenberg, L.H., Hyatt, L.E., 1997. Hybrid re-engineering, in: Third IEEE International

Symposium on Requirements Engineering (ISRE), IEEE Press, New York, NY.

[25] Kataoka, Y. Imai, T. Andou, and H. Fukaya, T., 2002. A quantitative evaluation of maintainability enhancement by refactoring, in: Proc. International Conference on Software Maintenance.

[26] Pizka. M., 2004. Straightening Spaghetti Code with Refactoring, in: Proc. of the Int. Conf. on Software Engineering Research and Practice - SERP, CSREA Press, pp 846- 852.

[27] Monteiro, M.P., Fernandez, J.M., 2006. Towards a catalogue of refactorings and code smells for AspectJ, in: A. Rashid and M. Aksit (Eds.), Transactions of Aspect-Oriented Software Development I: LNCS 3880, Springer, Heidelberg, pp. 214 – 258.

[28] Gamma, E,. Helm, R., Johnson, and R., Vlissides, J., 1994. Design Patterns: Elements of Reusable Object-Oriented Software, Addison Wesley Professional.

[29] Denier, S., Comte, P., 2006. Understanding design pattern density with aspects, Software Composition 5th international symposium.

[30] Rashid, A., Sawyer, P., 2001. Aspect-oriented and database systems: an effective customization approach, IEEE Software 148(5), pp. 156-164,

[31] Hanenberg, S., Oberschulte, and C., Unland, R., 2003. Refactoring of aspect-oriented software, in Unland, R. (Ed.), Lecture Notes in Computer Science, volume 2591, Springer, Heidelberg, 2003.

[32] Laddad, R., 2003. Aspect-oriented refactoring, Parts 1 and 2, The Server Side, http://www.theserverside.com/tt/articles/article.tss?l =AspectOrientedRefactoringPart1

[33] Demeyer, S., Ducasse, S., Nierstrasz,O., 2002. Object-Oriented Reengineering Patterns, Morgan Kaufmann, 2002.

[34] Binkley, D., Ceccato, M., Harman, M., Ricca, and F., Tonella, P., 2005. Automated refactoring of object-oriented code into aspects, in: 21st IEEE International Conference on Software Maintenance (ICSM 2005), IEEE Press, New York, NY, pp. 27–36.

[35] Apel, S., 2010. How AspectJ is Used: An Analysis of Eleven AspectJ Programs, Jounal of Object Technology (JOT), 9(1), pp. 117-142.

[36] Henderson-Sellers, B., 1994. Book Two of Object-Oriented Knowledge, Prentice Hall, Upper Saddle River, NJ.

[37] Hannemann, J., Kiczales, G., 2002. Design pattern implementation in Java and AspectJ, in: OOPSLA ‘02, ACM Press, New York, NY.

[38] Zakaria, A.A., Hosny, H., 2003. Metrics for aspect-oriented software design, in: Third International Workshop on Aspect Oriented Modeling at International Conference on Aspect-Oriented Software Development, ACM Press, New York, NY.

[39] Wilking, D., Khan, U.F., Kowalewski, S., 2007. An empirical evaluation of refactoring, e-Informatica Software Engineering Journal, 1(1).

[40] Bourqun, F., and Keller, R.K., 2007. High-impact refactoring based on architecture violations, in: Conference on Software Maintenance and Reengineering - CSMR , pp. 149-158.

[41] Thomas, W., Delis, A., Basili, V.R., 1997. An analysis of error in a reuse-oriented development Environment. Journal of Systems and Software, 38(3), 1997.

[42] Chen, C.-T., Cheng, Y.C., Hsieh, C.Y., Wu, I.-L., 2009. Exception handling refactorings: Directed by goals and driven by bug fixing, Journal of Systems and Software, 82(2), pp. 333-345.

[43] Murphy, G.C., Walker, R.J., Baniassad, E.L.A., Robillard, M.P., Lai, A., Kersten, M.A., Does aspect-oriented programming work? Communications of the ACM, 44(10), pp. 75-77.

[44] Greenwood, P., Bartolomei, T., Figueiredo, E., Dosea, M., Garcia, A., Cacho, N., Sant’Anna, C., Soares, S., Borba, P., Kulesza, U., Rashid, A., 2007. On the impact of aspectual decompositions on design stability: an empirical study, in: Ernst, E. (Ed.), ECOOP 2007: LNCS, vol. 4609, Springer, Heidelberg, pp. 176 - 200.

[45] Kellens, A., Mens, K., Tanella, P., 2007. Survey of automated code-level aspect mining Techniques, in: Rashid, A., Aksit, M. (Eds.), AOSD IV: LNCS 460, Springer, Heidelberg, pp. 14-162.

[46] Binkley, D., Ceccato, M., Harman, M., Tonella, P., 2006. Tool supported refactoring of existing object-oriented code into aspects, IEEE Transactions on Software Engineering.

[47] Colyer, A., Clement, A., 2004. Large-scale AOSD for middleware. in: Proceedings of the 3rd international conference on Aspect-oriented software development (AOSD '04). ACM Press, New York, NY, pp. 56-65.

[48] Filho, F.C., Rubira, C.M., Maranhão Ferreira, R., Garcia, A., 2006. Aspectization of exception handling: A quantitative study, in:Advanced Topics in Exception Handling Techniques: LNCS vol. 4119, Springer, Heidelberg, pp. 255-274.

[49] Lippert, M., V. Lopes, C., 2000. A study on exception detection and handling using aspect-oriented programming, in: International Conference Software Engineering (ICSE 2000), ACM Press, New York, NY, pp. 418-427.

[50] Tsang, S.L., Clarke, S., Baniassad, E., 2004. An evaluation of aspect-oriented programming for Java-based real-time systems development, in: 7th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, IEEE Press, New York, NY, pp. 291-300.

[51] Sullivan, K.J., Griswold, W.G., Song, Y., Cai, Y., Shonle, M., Tewari, N., Rajan, H., 2005. Information hiding interfaces for aspect-oriented designs, in: 10th European Software Engineering Conference, ACM Press, New York, NY, pp. 166–175.

[52] Zhang, C., Jacobsen, H., 2004. Resolving feature convolution in middleware systems, SIGPLAN

Notices, 39(10), ACM Press, New York, NY, pp. 188–205.
[53] Madeyski, L., Szala, L., 2007. Impact of aspect-oriented programming on software development and design quality, IET Software, 1(5), pp. 180-187
[54] Simon, F., Steinbreuckner, F.C., Lewerentz, C., 2001. Metrics based refactoring. in: European Conference on Software Maintenance and Reengineering, pp. 30-38.
[55] Cole, L. Borba, P., 2005. Deriving refactorings for AspectJ, in: AOSD IV, ACM Press, New York, NY, pp. 123-134.
[56] Ceccato, M., Marin, M., Mens, K., Moonen, L., Tonella, P., Tourwe, T., 2005. A qualitative comparison of three aspect mining techniques, in: 13th International Workshop on Program Comprehension, IEEE Press, New York, NY, pp. 13–22.
[57] Hannemann, J., Murphy, G., Kiczales, G., 2005. Role-based refactoring of crosscutting concerns, in: 4th International Conference on Aspect-Oriented Software Development, ACM Press, New York, NY, pp. 135 -146.
[58] Beuche, D., Beust, C., in: Colyer, A.M. , Kawakami Harrop Galvão, R., Johnson, R., Vasseur, A., Beuche, D., Beust, C., (Eds.) 2006. Point/counterpoint, IEEE Software 23(1), pp. 72-75.
[59] Wiese, D., Meunier, R., 2008. Large-scale application of AOP in the healthcare domain: A case study, in: 7th AOSD, ACM Press, New York, NY.

Web References

1. http://www.eclipse.org/ajdt/
2. http://gef.tigris.org/
3. http://argouml.tigris.org/
4. http://aopmetrics.tigris.org/

Evaluating the Effectiveness of Mutation Operators on the Behavior of Genetic Algorithms Applied to Non-deterministic Polynomial Problems

Basima Hani F. Hasan
Department of Computer Science
Yarmouk University, 21163 Irbid, Jordan
E-mail: basmah@yu.edu.jo

Moutaz Saleh M. Saleh
Department of Computer Science & Engineering
Qatar University, 2713 Doha, Qatar
E-mail: moutaz.saleh@qu.edu.qa

Keywords: evaluation, genetic, mutation operator, TSP, 0/1 knapsack problem, Shubert function, linear system

Received: April 25, 2011

Genetic Algorithms (GAs) are powerful general-purpose optimization search algorithms based upon the principles of evolution observed in nature. Mutation operator is one of the GA operators that used to produce new chromosomes or modify some features of them depending on some small probability value. The objective of this operator is to prevent falling of all solutions in population into a local optimum of solved problem. This paper evaluates the effect of applying well known mutation operators on selected non-deterministic polynomial (NP) hard problems and compares the results. The problems that will be introduced in this paper are: traveling salesman problem (TSP), 0/1 Knapsack problem, Shubert function, and system of linear equations.

Povzetek: Članek raziskuje učinkovitost mutacije v genetskih algoritmih pri reševanju NP-težkih problemov.

1 Introduction

GAs are powerful general purpose optimization search algorithms based upon the principles of evolution observed in nature. Even with today's high-powered computers, using an exhaustive search to find the optimal solution for even relatively small problems can be prohibitively expensive. For many problems, genetic algorithms can often find good solutions, near-optimal, in around 100 generations. This can be many times faster than an exhaustive search. Solution to a problem solved by genetic algorithms is evolved. Algorithm is started with a set of solutions, represented by chromosomes, called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions, offspring, are considered according to their fitness; the more suitable they are the more chances they have to reproduce. Then crossover and mutation are applied on them to find new points in the search space.

This paper will use GA to solve NP problems. A decision problem is called an NP problem if particular examples of it can be solved in polynomial time by a nondeterministic process i.e. by generating possible solutions at random guessing. There are many problems for which no polynomial-time algorithm is known, this paper will consider three of them; Traveling Salesman Problem (TSP), 0/1 Knapsack problem, and Shubert function.

Given a finite number of cities along with the cost of travel between each pair of them; TSP try to find the cheapest way of visiting all the cities each of which exactly once before returning back to the starting point. TSP difficulty comes from the fact that for N cities there are N!/2N possible paths. There are several algorithms, which approach this problem. This paper will use GA to solve this problem by applying several types of mutation methods and compare the results. These types, depending on the representation used are: Reciprocal exchange Mutation, Inversion mutation, Insertion mutation, Displacement mutation, Boundary mutation and Uniform random mutation.

The main idea of 0/1 Knapsack problem is how to fill the knapsack with the subset of given items to reach maximum profit without exceeding the knapsack capacity. The knapsack problem arises whenever there is resource allocation with financial constraints. Profit would be the importance of the item, while the cost is the amount of space it occupies in the knapsack. So we need to maximize our profit while minimizing our cost. This

problem shows how to deal with constraints. One way to achieve that is based on the application of special repair algorithms to correct any infeasible solutions. This paper will use GA to solve this problem by applying several types of mutation methods and compare the results. These mutation types depending on the representation used are: Flip bit, Boundary, Non-uniform, Inversion, Insertion and Displacement.

In Shubert function the target is to benchmark global optimization methods. The formulation of the global optimization problem is to find the absolute minimum for a given function over the allowed range of its variables. So, Shubert function is used as an indicator for the efficiency of these methods. The Shubert function could be 1, 2, 3… or n dimension. The object function for one-dimension Shubert function can be given as:

$$f(x_1) = \sum_{j=1}^{5} j \cdot \cos[(j+1)x_1 + j]$$
$$\text{where} \quad -10 \le x_1 \le 10$$

For n-dimension Shubert function, there are n*pow(3,n) global solutions. The object function for n-dimension Shubert function can then be given as:

$$f(\mathbf{x}) = \prod_{i=1}^{n} \sum_{j=1}^{5} j \cdot \cos[(j+1)x_i + j]$$
$$\text{where} \quad -10 \le x_i \le 10, \ \mathrm{i} = 1, \cdots, \mathrm{n}$$

GAs are general purpose search algorithms, so we can use them for searching the global minima of the Shubert function. This paper will use GA to solve this problem by several types of mutation methods and compare the results. These types depending on the representation used are: Boundary mutation, Uniform random mutation and Non-uniform mutation.

Many problems lend themselves to being solved with systems of linear equations. However, solving systems of linear equations is a common computational problem well known to mathematicians, scientists and engineers. Several algorithms exist for solving this problem. But, when the equations contain interval coefficients, i.e. intervals in which the desired coefficient values are known to lie, the problem may not be solvable in any reasonable sense. In fact, it has been shown that the general problem of solving systems of linear equations with interval coefficients is NP-hard which is extremely difficult to be solved. Hence, this paper will use GA to solve this problem by applying several types of mutation methods and compare the results. These types, depending on the representation used are: Boundary mutation, Uniform random mutation, Non-uniform mutation, Reciprocal mutation, Inversion mutation, Insertion mutation, and Displacement mutation.

2 Related Work

In recent years, the genetic algorithms for solving NP hard problems have achieved great results [1] [2] [3] [4]. In particular, the Traveling Salesman Problem (TSP) has been receiving continuous and growing attention in artificial intelligence, computational mathematics and optimization. For instance, the work in [5] proposed an improved GA to solve TSP through adopting an untwist operator which can unite the knots of route effectively, so it can shorten the length of route and quicken the convergent speed. In [6], a new selection strategy is incorporated into the conventional genetic algorithm to improve the performance of genetic algorithm in solving TSP. The results show that the number of evolutional iterations to reach an optimal solution can be significantly reduced. Liu and Huang [7] proposed a novel genetic algorithm to overcome the defections of slow convergence of traditional GA. The algorithm creates crossover and mutation by merging two kinds of heuristics. Simulation results indicated that it can get high-quality solution while consume less running time.

The 0/1 Knapsack Problem is a well-known NP hard problem [8] [9] [10] [11] as it appears in many real life world with different application. In [12] an evolutionary genetic algorithm for solving multi objective 0/1 Knapsack Problem is introduced. Experimental outcome show that the proposed algorithm outperforms the existing evolutionary approach. In addition, the work in [13] proposed a genetic algorithm using greedy approach to solve this problem. The experiments prove the feasibility and validity of the algorithm.

Shubert function optimization problem has also been studied in literature. Based on the principle of free energy minimization of thermodynamics, a new thermodynamics evolutionary algorithm (TDEA) for solving Shubert function optimization problem has been proposed in [14]. The results show that thermodynamics evolutionary algorithm is of potential to obtain global optimum or more accurate solutions than other evolutionary methods.

The problem of solving systems of linear equations with use of AI based approaches has been studied by many researches for decades. Different definitions for the solution of such problem have been considered and various AI techniques have been successfully developed [15] [16] [17]. Recently, the research in [18] aimed to approximate the exact algebraic solution of this problem through minimizing its cost function. To do so, two different AI approaches were adopted: the neural networks (NN) based approach and the genetic algorithms (GA) based one. The results shows that it will be interesting to combine both GA based and NN based approaches into one single method. GA can be used at the beginning phase for global search, whereas the NN based technique can be used in the final, local search phase to improve the solution obtained by GA.

3 GA Operation

GA is an iterative procedure that consists of a constant population size of individuals which are decoded and evaluated according to a fitness function. To form a new population, individuals are selected according to their fitness. A crossover and mutation are then applied on them to find new points in the search space. Figure 1 shows the iterative procedure of a general GA. For our research work, we will adopt the GA procedure steps to solve four well known NP problems: Traveling Salesman Problem (TSP), 0/1 Knapsack problem, Shubert function,

and system of linear equations. To do so, Table 1 shows how these steps are applied on the four problems.

```
Generate random population of n chromosomes
Loop (until stop condition)
    Evaluate fitness f(x) of each chromosome
    Loop (until new population is completed)
        Select 2 parent chromosome
        Crossover the parents to form offspring
        Mutate new offspring at each locus
        Place new offspring in a new population
    End Loop
Adopt newly generated population
Test for best solution
End Loop
```

Figure 1: A general GA procedure.

GA Step	TSP	0/1 Knapsack	Shubert Function	System of Linear Equations
Encoding	integer	binary	floating point	floating point
Initial Population	Heuristic (greedy algorithm from different 17 cities)	15 items	Two genes in each individual	Generated randomly within a specified interval for each variable
Parameters	Population size=50 Probability of crossover=80% Probability of mutation=5% Maximum number of Generations=5000			
Evaluation Function	total cost of travel between cities	summation of the profits of the items selected	smallest value we get is the best	how far obtained solution from the correct one
Selection Operator	roulette wheel	roulette wheel	roulette wheel	roulette wheel
Crossover Operator	heuristic	single point	single point	single point
Mutation Operator	• Reciprocal • Inversion • Insertion • Displacement • Boundary • Uniform	• Flip bit • Inversion • Insertion • Displacement • Boundary • Non uniform	• Boundary • Uniform random • Non-uniform	• Boundary • Uniform random • Non-uniform • Reciprocal • Inversion • Insertion • Displacement

Table 1: Choice of parameters for each GA step.

4 Mutation Operator

Text of the conclusion From biology view, mutation is any change of DNA material that can be reproduced. From computer science view, mutation is a genetic operator that follows crossover operator. It usually acts on only one individual chosen based on a probability or fitness function. One or more genetic components of the individual are scanned. And this component is modified based on some user-definable probability or condition. Without mutation, offspring chromosomes would be limited to only the genes available within the initial population. Mutation should be able to introduce new genetic material as well as modify existing one. With these new gene values, the genetic algorithm may be able to arrive at better solution than was previously possible. Mutation operator prevents premature convergence to local optima by randomly sampling new points in the search space. There are many types of mutation and these types depend on the representation itself.

4.1 Applying Mutation on TSP

Since integer representation is the best representation for TSP, the following mutation types were applied to this problem:

- *Reciprocal*: two cities are exchanged, swapped, after they are selected randomly.
- *Inversion:* two cut points are selected randomly, and then the sub tour between them is inverted.
- *Insertion*: a city and a place to be inserted in it are selected randomly.
- *Boundary:* a city is chosen and replaced randomly with either the upper or lower bound for that city. The chromosome is then searched for the upper or lower bound and that city is replaced with the bound as shown in Figure 2.

```
x = rand()%1000/1000.0;
if (x < PMUTATION)
{
        //to decide upper or lower boundary
        p = rand()%1000/1000.0;
        if (p < 0.5)
        //replace with lower boundary
        {
                currentpop = i;
                index = search(lbound,currentpop);
                if (index != -1)
                        swap(&current,&population[i].gene[index]);
                population[i].gene[j]=current;
        }
        else
        //replace with upper boundary
        {
                currentpop = i;
                index = search(ubound,currentpop);
                if (index != -1)
                        swap(&current,&population[i].gene[index]);
                population[i].gene[j]=current;
        }
}
```

Figure 2: Boundary mutation.

- *Displacement:* a sub tour and a place to be inserted in it are selected randomly as shown in Figure 3.

```
CityNum = ( rand() % CitiesNum);
InsertPos = (rand() % CitiesNum);
do
//length of the tour
length1 = (rand()%CitiesNum+1);
while ((InsertPos+length1-1>=CitiesNum)||(CityNum+length1-1>=CitiesNum));
p = rand()%1000/1000.0;
if (p < PMUTATION){
        w=0;
        for (int h = CityNum;h<=(CityNum+length1-1);h++)
        {
                w = w + 1;
                SubList[w]=population[i].gene[h];
        }
        if (InsertPos > CityNum){
                for (int j = CityNum + 1; j <=InsertPos; j++){
                        population[i].gene[j-1] = population[i].gene[j+length11];
                        w=0;
                        j=j-1;
                }
                for (int jj = j; jj<(j+length1); jj++){
                        w = w+1;
                        population[i].gene[jj]=SubList[w];
                }
        }
        if (InsertPos < CityNum){
                for (int j=CityNum - 1; j>=InsertPos; j--)
                {
                        population[i].gene[j+length1] = population[i].gene[j];
                        w=0;
                        j=j+1;
                }
                for (int jj = j; jj<(j+length1);jj++)
                {
                        w = w+1;
                        population[i].gene[jj]=SubList[w];
                }
        }
}
```

Figure 3: Displacement mutation.

- *Uniform:* a city selected for mutation is replaced with a uniform random value between the user-specified upper and lower bounds for that city. Then the chromosome is searched for the uniform random value found and replaces it with that city.

4.2 Applying Mutation on 0/1 Knapsack

since binary representation is the best representation for 0/1 Knapsack problem, the following mutation types were applied to this problem:

- *Flip bit*: an item is selected randomly and its value is inverted from 0, selected, to 1, unselected, or vice versa as shown in Figure 4.

```
x = (rand()%ItemsNum);
p = rand()%1000/1000.0;
if (p < PMUTATION)
       population[i].gene[x] = 1 - population[i].gene[x];
```

Figure 4: Flip mutation.

- *Inversion*: items between two randomly chosen points in the individual are reversed in order.
- *Insertion*: an item is taken at random and inserted randomly into another position in the sequence.
- *Displacement:* A randomly selected section of the individual is moved as a block to another location in the individual.
- *Boundary:* an item is selected randomly and its value is replaced randomly either by the upper bound (1) or the lower bound (0).
- *Non-uniform:* This type increases the probability that the amount of the mutation will be close to 0 as the generation number increases. This mutation operator keeps the population from stagnating in the early stages of the evolution then allows the genetic algorithm to fine tune the solution in the later stages of evolution as shown in Figure 5.

```
x = rand()%1000/1000.0;
if (x < PMUTATION)
{
        p = rand()%2;
        y = rand()%1000/1000.0;
        if (p == 0){
                pos=floor(ItemsNum*(1-pow(y,pow(1-generation/MAXGENS,5))));
                population[i].gene[pos]=1-population[i].gene[pos];
        }
        else if (p == 1){
                pos=ceil(ItemsNum*(1-pow(y,pow(1-generation/MAXGENS,5))));
                population[i].gene[pos]=1-population[i].gene[pos];
        }
}
```

Figure 5: Non-uniform mutation.

If $s_v^t = \langle v_1,\ldots,v_m\rangle$ is a chromosome (t is the generation number) and the element vk was selected for this mutation, the result is a vector $s_v^{t+1} = \langle v_1,\ldots,v_k',\ldots,v_m\rangle$, where $v_k' = \text{mutate}(v_k, \nabla(t,n))$, where n is the number of bits per one element of a chromosome, mutate(v_k, pos) means mutate the k-th value element on pos bit, and:

$$\nabla(t,n) = \begin{cases} \lfloor \Delta(t,n) \rfloor & \text{if a random digit is 0} \\ \lceil \Delta(t,n) \rceil & \text{if a random digit is 1} \end{cases}$$

4.3 Applying Mutation on Shubert Function

for best practices, floating point representation is commonly used for encoding Shubert function with two genes, variables, in each individual. Accordingly, a 2-dimension Shubert function is adopted. This function has 760 local minima, 18 of which are global minima with value -186.73067. Its object function is:

$$f(x_1,x_2) = \sum_{j=1}^{5} j\cdot\cos[(j+1)x_1 + j] \bullet \sum_{j=1}^{5} j\cdot\cos[(j+1)x_2 + j]$$

where $-10 \le x_i \le 10 \quad i = 1,2$

Since this is a two dimension function, with only two variables, the following mutation types are used:

- *Boundary:* a gene is selected and replaced randomly by the upper (10) or lower (-10) bound.
- *Uniform random*: a gene is selected randomly and replaced by a random number from the interval of [-10.0, 10.0].
- *Non-uniform*: If $s_v^t = \langle v_1,\ldots, v_m\rangle$ is a chromosome, *t* is the generation number, and the element v_k was selected for this mutation, the result is a vector $s_v^{t+1} = \langle v_1,\ldots,v_k',\ldots,v_m\rangle$, where,

$$v_k' = \begin{cases} v_k + \Delta(t, UB-v_k) \text{ if a random digit is 0} \\ v_k - \Delta(t, v_k-LB) \text{ if a random digit is 1} \end{cases}$$

LB and UB are lower and upper domain bounds of the variable v_k. In addition, the following function is used:

$$\Delta(t, y) = y.\,(1-r\,(1-t/T)\,b)$$

Where *r* is a random number from [0, 1], *T* is the maximum generation number, and *b* is a system parameter determining the degree of dependency of the iteration number.

4.4 Applying Mutation on Linear System Equation

The floating-point representation is used for encoding individuals in such system with every gene represents the value of one variable. The individuals of the initial population can be generated randomly within a specified interval for each variable, and the evaluation function is used to indicate how far the obtained solution from the correct one. For the general system of linear equations:

$$a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n = b_1$$
$$a_{21}X_1 + a_{22}X_2 + \ldots + a_{2n}X_n = b_2$$
$$\ldots \qquad \ldots \qquad \ldots$$
$$a_{n1}X_1 + a_{n2}X_2 + \ldots + a_{nn}X_n = b_n$$

suppose that the vector $V = (v_1,v_2,\ldots,v_n)$ represents a solution. Then the evaluation function will be:

$$\frac{1.0}{\sum_{i=1}^{n}\left|\left(\sum_{j=1}^{n} aij.vj\right)-bi\right|}$$

Eventually, the following mutation types will be applied for this linear equation system: Boundary, Uniform random, Non-uniform, Reciprocal, Inversion, Insertion, and Displacement.

5 Test and Results

After running the program 50 times for each type of mutation, we get the results represented in tables 2, 3, 4, and 5 for TSP, 0/1 Knapsack, Shubert function, and linear equation system problems respectively.

	Reciprocal	Inversion	Insertion	Displacement	Boundary	Uniform
Mean	1818.8	1791.9	1718.2	1967.5	2168.9	2071.6
Std Div	429.7	301.5	138.3	370.0	470.5	473.8

Table 2: Output per mutation type applied to TSP.

	Flip Bit	Boundary	Non-Uniform	Inversion	Insertion	Displacement
Mean	1514.8	1508.2	1519.9	1538.6	1520.1	1530.0
Std Div	17.6	6.7	27.6	90.1	28.7	47.0

Table 3: Output per mutation type applied to 0/1 Knapsack.

	Boundary	Uniform	Non-Uniform
Mean	1002.2	2078.5	1002.88
Std Div	1.7	7.3	4.0

Table 4: Output per mutation type applied to Shubert function (x1, x2).

	Boundary	Uniform	Non-Uniform	Reciprocal	Inversion	Insertion	Displacement
Mean	0.1812	4.9277	0.1770	0.2398	0.3253	0.3416	0.2414
Std Div	0.0987	3.9398	0.0883	0.1509	0.3473	0.3396	0.1180

Table 5: Output per mutation type applied to Linear Equation System.

Now, for evaluating the effectiveness of each mutation type applied to TSP, two hypotheses are considered:
First Hypothesis (H0): $M_{insertion} = M_{inversion} = M_{reciprocal} = M_{displacement} = M_{uniform} = M_{boundary}$
Second Hypothesis (H1): $M_{insertion} <= M_{inversion} <= M_{reciprocal} <= M_{displacement} <= M_{uniform} <= M_{boundary}$
Here, M_X is the mean of the number of generations required to reach the desired solution using the X mutation. In H0 we assume that all types of mutation are having the same mean M, but in H1 we assume that there is a difference between them; such that the mean in the first type is less than, better, the mean in the second type and so on. Now, we apply the Jonckheere-Terpstra test of the statistical package SPSS on the collected data for TSP to either accept or reject the above hypotheses. The results in Table 6 give some statistics. The very important part of these results is the value of P (in the last line). Since P = 0.000 which is < 0.01, we reject H0 and accept H1. This means that the proposed order of mutation types in H1 is correct.

Parameter	Value
Number of Levels	6
Population Size	300
Observed J-T Statistic	25605.500
Mean J-T Statistic	18750.000
Std. Deviation of J-T Statistic	855.730
Std. J-T Statistic	8.011
Asymp. Sig. (2-tailed)	0.000

Table 6: Results of Jonckheere-Terpsta Test on TSP.

For evaluating the effectiveness of each mutation type applied to 0/1 Knapsack, two hypotheses are considered:
First Hypothesis (H0): $M_{boundary} = M_{FlipBit} = M_{nonUniform} = M_{insertion} = M_{displacement} = M_{inversion}$
Second Hypothesis (H1): $M_{boundary} <= M_{FlipBit} <= M_{nonUniform} <= M_{insertion} <= M_{displacement} <= M_{inversion}$
After applying the Jonckheere-Terpstra test, we get the results shown in Table 7. Since P = 0.016 which is > 0.01, we reject H1 and accept H0.

Parameter	Value
Number of Levels	6
Population Size	300
Observed J-T Statistic	20801.500
Mean J-T Statistic	18750.000
Std. Deviation of J-T Statistic	854.717
Std. J-T Statistic	2.400
Asymp. Sig. (2-tailed)	0.016

Table 7: Results of Jonckheere-Terpsta Test on 0/1 Knapsack..

To evaluate the effectiveness of each mutation type applied to Shubert, the following two hypotheses are considered:
First Hypothesis (H0): $M_{boundary} = M_{Non\text{-}Uniformt} = M_{Uniform}$
Second Hypothesis (H1): $M_{boundary} <= M_{Uniform} <= M_{Non\text{-}Uniform}$
After applying the Jonckheere-Terpstra test, we get the results shown in Table 8. Accordingly, since P = 0.000 which is < 0.01, we reject H0 and accept H1.

Parameter	Value
Number of Levels	3
Population Size	150
Observed J-T Statistic	6382.000
Mean J-T Statistic	3750.000
Std. Deviation of J-T Statistic	287.284
Std. J-T Statistic	9.162
Asymp. Sig. (2-tailed)	0.000

Table 8: Results of Jonckheere-Terpsta Test on Shubert f(x1, x2)

To evaluate the effectiveness of each mutation type applied to the linear equation system, the following two hypotheses are considered:
First Hypothesis (H0): $M_{nonUniform} = M_{Boundary} = M_{displacement} = M_{reciprocal} = M_{inversion} = M_{insertion} = M_{uniform}$

Second Hypothesis (H1): $M_{nonUniform} <= M_{Boundary} <= M_{displacement} <= M_{reciprocal} <= M_{inversion} <= M_{insertion} <= M_{uniform}$ After applying the Jonckheere-Terpstra test, we get the results shown in Table 9. Accordingly, since P = 0.000 which is < 0.01, we reject H0 and accept H1.

Parameter	Value
Number of Levels	7
Population Size	350
Observed J-T Statistic	38576.000
Mean J-T Statistic	26250.000
Std. Deviation of J-T Statistic	1082.035
Std. J-T Statistic	11.392
Asymp. Sig. (2-tailed)	0.000

Table 9: Results of Jonckheere-Terpsta Test on Linear Equation System

6 Conclusions and Future Work

Genetic algorithms are an effective way to solve many problems especially NP-hard problem. In this paper, genetic algorithms were used to solve TSP, 0/1-Knapsack problem Shubert Function, and linear equation system. Mutation is one of the important operators of genetic algorithms since the type of mutation used often has great effects on the results. The research study shows that insertion mutation is the best suite for TSP, Boundary and non-uniform mutations are the best to use for Shubert function and linear equation system, but for 0/1 knapsack problem all mutation types used gave nearly the same result. For future work, other NP problems can be solved with genetic algorithms, and new mutations can be obtained by combining two or more types of mutation operators.

References

[1] M. Fangfang, and L. Han, “An Algorithm in Solving the TSP Based on the Improved Genetic Algorithm,” 1st IEEE International Conference on Information Science and Engineering (ICISE), 2009, pp. 106-108.

[2] Y.Yi, and Q. Fang, “The improved hybrid genetic algorithm for solving TSP based on Handel-C”, 3rd IEEE International Conference on Advanced Computer Theory and Engineering (ICACTE), vol. 3, 2010, pp. 330-333.

[3] Z. Tao, “TSP Problem Solution Based on Improved Genetic Algorithm”, 4th IEEE International Conference on Natural Computation, vol. 1, 2008, pp. 686-690.

[4] T. Hong, W. Lin, S. Liu, and J. Lin, “Experimental analysis of dynamic migration intervals on 0/1 knapsack problems”, IEEE Congress on Evolutionary Computation, 2007, pp. 1163-1167.

[5] L. Wang, J. Zhang, and H. Li, “An Improved Genetic Algorithm for TSP”, IEEE International Conference on Machine Learning and Cybernetics, vol. 2, 2007, pp. 925-928

[6] J. Lu, N.Fang, D. Shao, and C. Liu, “An Improved Immune-Genetic Algorithm for the Traveling Salesman Problem”, 3rd IEEE International Conference on Natural Computation, 2007, pp. 297-301.

[7] Y. Liu, and J. Huang, “A Novel Genetic Algorithm and Its Application in TSP ”, IEEE IFIP International Conference on Network and Parallel Computing, 2008, pp. 263-266.

[8] H. Ishibuchi, and K. Narukawa, “Performance evaluation of simple multiobjective genetic local search algorithms on multiobjective 0/1 knapsack problems”, IEEE Congress on Evolutionary Computation, vol. 1, 2004, pp. 441-448.

[9] D.S. Vianna, and J.E.C. Arroyo, “A GRASP algorithm for the multi-objective knapsack problem”, 24th IEEE International Conference of the Chilean Computer Science Society, 2004, pp. 69-75.

[10] H.H. Yang, S.W. Wang, H.T. Ko, and J.C. Lin, “A novel approach for crossover based on attribute reduction - a case of 0/1 knapsack problem”, IEEE International Conference on Industrial Engineering and Engineering Management, 2009, pp. 1733-1737.

[11] C.L. Mumford, ”Comparing representations & recombination operators for the multi-objective 0/1 knapsack problem”, The IEEE 2003 Congress on Evolutionary Computation, vol. 2, 2003, pp. 854-861.

[12] S.N. Mohanty, and R. Satapathy, “An evolutionary multiobjective genetic algorithm to solve 0/1 Knapsack Problem”, 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 397-399.

[13] S. Kaystha, and S. Agarwal, “Greedy genetic algorithm to Bounded Knapsack Problem”, 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 6, 2010, pp. 301-305.

[14] W. Xuan, W. Shao-song, and X. Li, “Solving Shubert Function Optimization Problem by Using Thermodynamics Evolutionary Algorithm”, IEEE International Conference on Biomedical Engineering and Computer Science (ICBECS), 2010, pp. 1-4.

[15] G. Alefeld, V. Kreinovich, and G. Mayer , “On the Solution Sets of Particular Classes of Linear Interval Systems”, Journal of Computaional and Applied Mathematics, 2003, pp. 1–15.

[16] S. Markov, “An Iterative Method for Algebraic Solution to Interval Equations”, Journal of Applied Numerical Mathematics, 1999, pp. 225–239.

[17] S. Ning, and R.B. Kearfott, “A comparison of some methods for solving linear interval equations”, SIAM Journal of Numerical Analysis, 1997, vol. 34(4), pp. 1289–1305.

[18] N.H. Viet and M. Kleiber, “AI Methods in Solving Systems of Interval Linear Equations”, In proccedings of the ICAISC 2006, Springer-Verlag, LNAI 4029, pp. 150–159.

Short Signcryption Scheme for the Internet of Things

Xuanwu Zhou[1,2], Zhigang Jin [1], Yan Fu[1,3], Huaiwei Zhou[3], Lianmin Qin[3]
[1] School of Electronics and Information Engineering, Tianjin University, Tianjin 300072, China
[2] Command College of the Chinese Armed Police Forces, Tianjin 300250, China
[3] Information Technology Research Center, Huarong Corporation, Yantai 265400, China
E-mail: schwoodchow@163.com

Keywords: internet of things, signcryption, provable security, distributed key management, system efficiency

Received: June 23, 2011

Signcryption is an effective cryptographic primitive, which simultaneously fulfils both the functions of encryption and signature with much lower cost than traditional schemes; it is an ideal method to provide confidentiality and unforgeability and ensure secure data storage and transmission in the IOT (Internet of things). In the paper, we propose a publicly verifiable short signcryption scheme S-ECSC for the Internet of things based on elliptic curves cryptosystem; and prove the provable security of S-ECSC under the Random Oracle model, including confidentiality in IND-CCA2 model, unforgeability in UF-CMA model and non-repudiation security. As per the efficiency analysis, S-ECSC achieves an average 80% reduction in computation cost compared with typical discrete logarithm, RSA based signcryption schemes, and has the lowest communication cost in Elgamal type signcryptions. With its superiority in efficiency and security, S-ECSC proves to be more suitable for resource-restricted environment in IOT and better satisfies the requirement of secure protocols in IOT, such as key management, secure routing, etc. At last, we take key generating and distributing protocol of distributed key management in IOT as an application example, and analyse the method and importance to apply S-ECSC into secure protocols in IOT.

Povzetek: Članek opisuje šifrirno shemo za internet stvari.

1 Introduction

The concept of IOT (Internet of Things) was first put forward by Ashton of the former MIT Auto-ID Center in 1999 when he was working on RFID (Radio Frequency Identification). Presently, the most widely-accepted definition of IOT is as follows [1, 2, 3, 4]. IOT is a self-configuring network in which things are connected with network according to certain protocols with RFID, ultra-red sensor, GPS(Global Positioning System), laser scanner, etc to interchange and transmit data, and ultimately achieve intelligent identification, positioning, tracing, supervision and management. IOT is the new direction of future computer and communication technology, and is regarded as the third landmark in the development of information technology after computer science and Internet.

According to the function classification, the hierarchical structure of IOT is composed of application layer, network layer and sensor layer. The basic function of network layer is secure and reliable interconnection between things via wire-based or wireless technology, in which the secure and dynamic interconnection via wireless network has been the overwhelming trend. In wireless network technology, many researchers have focused on IEEE802.11 WLAN (Wireless Local Area Network) , which is mainly composed of wireless Ad hoc network, WSN (Wireless Sensor Network) and WMN (Wireless Mesh Network). As a new wireless network, IEEE802.11 WLAN proves to be suitable for commercial, medical, domestic, military, and other applications with its superiority, such as inexpensiveness, adaptability and reliability, etc. In the Internet of things, IEEE802.11 WLAN has been playing an increasingly important role in secure and reliable connection between different objects. Whereas, the distributed network management and restricted network resources in IEEE802.11 WLAN have rendered many problems as to the security of confidentiality, integrity, non-repudiation and availability for data storage and transmission in IOT. Besides, security measures designed for traditional network, which has relatively abundant network resources, fixed connection, stable topology, special routing and comprehensive network service, are not completely applicable to wireless network environment in the Internet of thing. Therefore, it is of great necessity to design special security technology, protocols and corresponding algorithms for the secure and dynamic wireless communication in the Internet of things.

The confidentiality and integrity of message is the basic requirement for secure communication in IOT; in the symmetric setting, efforts focused on the composition of symmetric key encryption and message authentication code (MAC). In asymmetric settings, the composition method of “signature-then-encryption” has been

employed. But these have all proved impractical not only for the insecurity in case of arbitrary schemes but also for the low efficiency regarding application into resource-restricted environment in IOT, which results from the sum cost of encryption and signature.

In 1997, Zheng proposed a cryptographic primitive "signcryption" [5], which simultaneously fulfils the integrated function of public encryption and digital signature with a computing and communication cost significantly smaller than that required by the "signature-then-encryption" method. Since then, signcryption has been a focus of cryptography as an ideal method to simultaneously provide confidentiality and unforgeability and many researchers have explored the application of signcryption in different security protocols [6, 7, 8, 9, 10, 11]. The study of signcryption algorithms suitable for IOT network environment and its application in IOT security schemes is an important direction in cryptography; it is more of a requirement from the rapid development of the Internet of things than just a requirement from the theoretical or applied cryptography research.

In order to improve the security and efficiency of communication in the Internet of things, we propose a publicly verifiable short signcryption scheme *S-ECSC* for the Internet of things based on elliptic curves cryptosystem; and prove the provable security of *S-ECSC* under the Random Oracle model, including confidentiality in *IND-CCA2* model, unforgeability in *UF-CMA* model and non-repudiation security. At last, we take key generating and distributing protocol for different terminals of distributed key management in IOT as an application example, and analyse the method and importance in the application of *S-ECSC* into secure protocols in IOT. Compared with other typical discrete logarithm, RSA and elliptic curves based signcryption schemes, *S-ECSC* is more suitable for resource-restricted environment in IOT communication with its superiority in computing and communication cost and can better satisfy the requirement of secure protocols in IOT, such as key management, secure routing, etc.

2 Short Signcryption Scheme on Elliptic Curves

First, we pin down the basic notions concerning signcryption which will facilitate the design and analysis of the short signcryption scheme.

2.1 Basic Notions in Signcryption

Definition 2.1.1 (Elliptic Curve) An elliptic curve $E(F_q)$ over finite field F_q is a sextuple: $T = (q, a, b, P, l, h)$, where $P = (x_P, y_P)$ is the base point of $E(F_q)$, prime l is the order of P. As to $t \in Z_l^*$, Q and $G \in E(F_q)$, $Q = tG$ denotes multiple double additions on elliptic curve. O is the point at infinity, satisfying $lP = O$ and $G + O = G$ for any point $G \in E(F_q)$.

Definition 2.1.2 (ECDLP, Elliptic Curve Discrete Logarithm Problem). ECDLP is the following computation:

$$x \leftarrow ECDLP(Q, P).$$

P is a base point and $Q \in \langle P \rangle$, $x \in Z_l^*$, $Q = xP$.

Definition 2.1.3 (Signcryption Scheme) A signcryption scheme $\Sigma = (GC, GK, SC, USC)$ consists of the following algorithms:

1. Probabilistic common parameters generation algorithm $GC(1^k)$ takes security parameter 1^k as input and returns a sequence of common parameters such as description of computational groups and hash functions.

2. Key generation algorithm $GK(ID, 1^k)$, which is also probabilistic, takes identity and security parameter as input and returns secret/public key-pair (sk_{ID}, PK_{ID}).

$$(sk_{ID}, PK_{ID}) \leftarrow GK(ID, 1^k).$$

3. Signcryption algorithm $SC(sk_A, PK_B, m)$ that takes sender's secret key sk_A, receiver's public key pk_B and message $m \in SP_m$ (SP_m is the message space) as input and returns signcryption text C or $\perp$ (a reject symbol). It is also probabilistic algorithm.

$$C \cup \{\perp\} \leftarrow SC(sk_A, PK_B, m).$$

4. Deterministic unsigncryption algorithm $USC(sk_B, PK_A, C)$ takes as input receiver's secret key sk_B, sender's public key PK_A and signcryption text C, and returns either message m or $\perp$.

$$m \cup \{\perp\} \leftarrow USC(sk_B, PK_A, C).$$

If the signcryption scheme is publicly verifiable, it is composed of an additional public verification algorithm PV.

5. Deterministic public verification algorithm $PV(PK_A, PK_B, C, R)$ takes as input public key pair (PK_B, PK_A), signcryption text C and parameter R, and returns either "*true*" or $\perp$.

$$\text{"}True\text{"} \cup \{\perp\} \leftarrow PV(PK_A, PK_B, C, R).$$

2.2 *S-ECSC* Signcryption Algorithm

Short signcryption scheme $S-ECSC = (GC, GK, SC, USC, PV)$

Common parameters generation

$GC(1^k)$ = "On input (1^k):

$K : E(F_q) \rightarrow \{0,1\}^{L_K(1^k)}$, $H : \{0,1\}^* \rightarrow Z_l^*$,

$(K, H, T) \leftarrow GC(1^k)$."

$T = (q, a, b, P, l, h)$,where $P = (x_P, y_P)$ is the base point of $E(F_q)$, $ord(P) = l$ is a prime, O is the point at infinity.

Key pair generation

$GK(A, 1^k)$ ="On input $(A, 1^k)$:

$sk_A \xleftarrow{\$} Z_l^*, PK_A = sk_A P \neq O$,

$(sk_A, PK_A) \leftarrow$."

$GK(B, 1^k)$ ="On input $(B, 1^k)$:

$sk_B \xleftarrow{\$} Z_l^*, PK_B = sk_B P \neq O$,

$(sk_B, PK_B) \leftarrow$."

Signcryption

$SC(sk_A, PK_B, m)$="On input (sk_A, PK_B, m):

If $sk_A \notin Z_l^*$ or $PK_B \notin < P >$ return $\perp$,

$r \xleftarrow{\$} Z_l^*$, $R = (x_R, y_R) \leftarrow rPK_B$,

$\sigma \leftarrow K(R), c \leftarrow E_\sigma(m)$,

$h \leftarrow H(m \| PK_A \| PK_B \| R)$,

$s = (hsk_A + r) \bmod l$,

$C = (c, h, s)$."

Symmetric encryption scheme $\Upsilon = (E, D)$ is an encryption scheme with passive indistinguishability defined in Definition 3. 2.8.

Unsigncryption

$USC(sk_B, PK_A, C)$="On input (sk_B, PK_A, C):

If $sk_B \notin Z_l^*$ or $PK_A \notin < P >$ return $\perp$,

Parse C into (c, h, s),

If $h, s \notin Z_l^*$ or $c \notin SP_E$ return $\perp$,

Else $I = sP - hPK_A$, $R = sk_B I = (x_R, y_R)$,

$\sigma \leftarrow K(R), m \leftarrow D_\sigma(c)$,

$h' \leftarrow H(m \| PK_A \| PK_B \| R)$,

If $h = h'$ return m, else return $\perp$."

Public Verification

If controversy arises between signcryption senders and receiver, a trusted third party can solve the repudiation. The third party evaluates the following formula after signcryption receiver publishing (R, C).

$PV(PK_A, PK_B, C, R)$=

"On input PK_A, PK_B, C, R):

Parse C into (c, h, s),

$\sigma \leftarrow K(R), m \leftarrow D_\sigma(c)$,

$h' \leftarrow H(m \| PK_A \| PK_B \| R)$,

If $h = h'$ return *true*, else return $\perp$."

3 Provable Security of *S-ECSC*

In this section, we will analyse the provable security of the signcryption in random oracle model, including confidentiality, unforgeability and non-repudiation.

3.1 Correctness of *S-ECSC*

Definition 3.1.1 Message space $Message(sk_A, PK_B)$ is the set of all m associated to each private/public key pair (sk_A, PK_B) output by $GK(ID, 1^k)$ for which $SC(sk_A, PK_B, m)$ never returns $\perp$.

Definition 3.1.2 A signcryption scheme $\Sigma = (GC, GK, SC, USC)$ is correct if $USC(sk_B, PK_A, C) = m$ for any private/public key pair (sk_A, PK_B) output by $GK(ID, 1^k)$, any message $m \in Message(sk_A, PK_B)$, and any $C \neq \perp$ that might be output by $SC(sk_A, PK_B, m)$.

Theorem 3.1.1 *S-ECSC* is correct for any private/public key pair (sk_A, PK_B) output by $GK(ID, 1^k)$, any message $m \in Message(sk_A, PK_B)$ and any $C \neq \perp$ that might be output by $SC(sk_A, PK_B, m)$.

Proof of correctness: Obviously, the signcryption scheme *S-ECSC* is correct if and only if

$$USC(SC(sk_A, PK_B, m)) = m.$$

As per the formula in the scheme,

$$\begin{aligned} sk_B I &= sk_B(sP - hPK_A) \\ &= sk_B(sP - hsk_A P) = sk_B(s - hsk_A)P \\ &= sk_B rP = rPK_B = R = (x_R, y_R), \end{aligned}$$

$\sigma \leftarrow K(R)$.

$\Rightarrow m \leftarrow D_\sigma(c)$,

$h' \leftarrow H(m \| PK_A \| PK_B \| R)$,

$\Rightarrow h = h'$, $m \leftarrow USC(sk_B, PK_A, C)$.

Thus $USC(SC(sk_A, PK_B, m)) = m$, the short signcryption *S-ECSC* is correct, as desired.

3.2 Confidentiality of *S-ECSC*

Definition 3.2.1 Computational Elliptic Curve Problem (CECP). Let $T = (q, a, b, P, l, h)$ be an elliptic curve and *AC* an attacker on CECP , CECP is defined as the following:

Experiment $EXP_T^{CECP}(AC)$

$d, e \xleftarrow{\$} Z_l^*$,

$D = dP, E = eP$,

$F \in \langle P \rangle \leftarrow AC^T(D, E)$,

If $F = deP$ return 1 else return 0.

Note that $F = deP = dE = eD$. (1)

Definition 3.2.2 Decisional Elliptic Curve Problem (DECP). Let $T = (q, a, b, P, l, h)$ be an elliptic curve and *AD* an attacker on DECP , DECP is defined as the following:

Experiment $EXP_T^{DECP}(AD)$

$b \xleftarrow{\$} \{0,1\}$,

If b =0 $Sce_0 : d, e, f \xleftarrow{\$} Z_l^*$,

If b =1 $Sce_1 : d, e \xleftarrow{\$} Z_l^*, f = de(\mathrm{mod}\, l)$,

$D = dP, E = eP, F = fP$,

$b' \leftarrow AD^T(D, E, F)$,

If $b' = b$ return 1 else return 0.

Definition 3.2.3 DECP Oracle O_T^{DECP} . Let $T = (q, a, b, P, l, h)$ be an elliptic curve, DECP Oracle is defined as the following:

O_T^{DECP} ="on input (P, D, E, F)

If $D, E, F \notin < P >$ return $\perp$, else

If *DCEP* (D, E, F) =1 return 1,

If *DCEP* (D, E, F) =0 return 0."

Definition 3.2.4 Elliptic Curve Gap Problem (ECGP). Let $T = (q, a, b, P, l, h)$ be an elliptic curve and *AECG* an attacker on ECGP , let's consider the following experiment:

Experiment $EXP_T^{ECGP}(AECG)$

$d, e \xleftarrow{\$} Z_l^*$,

$F = fP \leftarrow AECG_T^{O_T^{DECP}(\cdot,\cdot)}(d, e)$,

If $f = de(\mathrm{mod}\, l)$ return 1 else return 0.

The *ECGP advantage* of *AECG* is defined as

$Adv_T^{ECGP}(AECG) = \Pr(EXP_T^{ECGP}(AECG) = 1)$. (2)

Hypothesis 3.2.1 (ECGP is hard). Given elliptic curve T and secure parameter 1^k , the probability of solving ECGP in time t is $\xi(1^k, T)$ which is negligible, that is

$$\xi(1^k, T) = \Pr[1^k, d, e \xleftarrow{\$} Z_l^*, Q \leftarrow xP : EXP_T^{ECGP}(AECG) = 1]. \quad (3)$$

Definition 3.2.5 Left-or-right signcryption oracle. Let $\Sigma = (GC, GK, SC, USC)$ be a signcryption scheme, a left-or-right signcryption oracle is defined as follows.

Oracle $SC_{sk_A, PK_B}(LR(m_0, m_1, b)) =$

"On input (m_0, m_1):

$b \in \{0,1\}, m_0, m_1 \in SP_m$,

$C \leftarrow SC(sk_A, PK_B, m_b)$,

Return C ."

Definition 3.2.6 Confidentiality of signcryption. Let *ASC* be an algorithm against the confidentiality of signcryption scheme Σ that has access to a left-or-right signcryption oracle and returns a bit. We consider the following experiment:

Experiment $EXP_{SGC}^{ind-cca2}(ASC)$

$(K, H, T) \leftarrow GC(1^k)$

$,(sk_A, PK_A) \leftarrow GK(A, 1^k)$,

$(sk_B, PK_B) \leftarrow GK(B, 1^k)$,

$C' \leftarrow SC_{sk_A, PK_B}(LR(m_0, m_1, b)), b \leftarrow \{0,1\}$,

$b' \leftarrow ASC^{SC_{sk_A, PK_B}(LR(\cdot,\cdot,b)), USC(sk_A, PK_B, \cdot)}$

If *ASC* queried $USC(sk_A, PK_B, \cdot)$ on a signcryption text previously returned by $SC_{sk_A, PK_B}(LR(m_0, m_1, b))$ then return 0,

If $b' = b$ return 1 else return 0.

The *IND-CCA2 advantage* of *ASC* is defined as

$$\delta(k) = Adv_{SGC}^{ind-cca2}(ASC) = \Pr(EXP_{SGC}^{ind-cca2}(ASC) = 1). \quad (4)$$

A signcryption scheme is indistinguishable under adaptive chosen cipher-text attack if the *IND-CCA2 advantage* of any attacker *ASC* with reasonably restricted resources (time-complexity, frequency and length of queries) is negligible.

Definition 3.2.7 Left-or-right Encryption Oracle. Let $\Upsilon = (E, D)$ be the symmetric encryption algorithm in the signcryption scheme, a left-or-right encryption Oracle is defined as:

Oracle $E_\sigma(LR(m_0, m_1, b))$ ="On input (m_0, m_1):

$b \in \{0,1\}, m_0, m_1 \in SP_m$,

$C \leftarrow E_\sigma(m_b)$,

Return C ."

Definition 3.2.8 Passive Indistinguishability. Let *AI* be an algorithm against the passive indistinguishability of symmetric encryption scheme Υ , which has access to a left-or-right encryption oracle and returns a bit. We consider the following experiment:

Experiment $EXP_\Upsilon^{pi}(AI)$

$C' \leftarrow E_\sigma(LR(m_0, m_1, b))$,

$b \leftarrow \{0,1\}, b' \leftarrow AI^{E_\sigma(LR(\cdot,b))}$,

If $b' = b$ return 1 else return 0.

The *pi advantage* of *AI* is defined as

$$\upsilon(k) = Adv_\Upsilon^{pi}(AI) = \Pr(EXP_\Upsilon^{pi}(AI) = 1). \quad (5)$$

An encryption scheme is passively indistinguishable if the *pi advantage* of any attacker *AI* with reasonably restricted resources (time-complexity, frequency and length of queries) is negligible.

Hypothesis 3.2.2 (Ideal Hash Function). Hash function has the property of Random Oracle. Namely, the outputs of hash function are randomly and uniformly distributed.

Theorem 3.2.1 If there exists an algorithm *ASC* against the *IND-CCA2* property of signcryption scheme Σ in time t with non-negligible advantage $\delta(k)$, using q_{SC}

queries to its signcryption oracle and (q_H, q_M) queries to its random oracles. Then we can thus formulate an *AECG* attacker on ECGP with non-negligible advantage ξ (1^k, T) in time t', using q_{SC} queries to its signcryption oracle and $q_{SC}+q_H$ queries to its random oracles.

Proof of confidentiality: In our proof, the random oracles K and H are replace by the random oracle simulators with two types of "query-answer" lists. For example, Sim_K simulates random oracle K with two types of "query-answer" lists L_1^K and L_2^K. L_1^K consists of simple "query-answer" (R, σ) entries from K ,while L_2^K consists of special input-output entries ($PK_B \| \Omega_i \| (?,\sigma)$) which implies $\sigma = K(\Omega^{sk_B})$ with the implicit input Ω^{sk_B} stored and denoted as "?".

$\mathrm{Sim}\_K(L^K, R)$ ="on input (L^K, R):

If $1 \leftarrow DECP(D, PK_B, R)$ return $\perp$,

Else if $1 \leftarrow DECP(\Omega_i, PK_B, R)$ return σ_i,

//there is an entry ($PK_B \| \Omega_i \| (?,\sigma_i)$) *in* L_2^K

Else if $R = R_i$, return σ_i *//there is an entry* (R_i, σ_i) *in* L_1^K

Else $\sigma_i \xleftarrow{\$} \{0,1\}^{L_K(1^k)}$, $R_i \leftarrow R$,

Add (R_i, σ_i) into L_1^K."

If $EXP_{SGC}^{ind-cca2}(ASC)$ makes queries to random Oracle O^H, *AECG* will reply with simulator Sim_H.

$\mathrm{Sim}\_H(L^H, \hbar)$ ="on input (L^H, $\hbar$): // $\hbar$ =($m \| PK_A \| PK_B \| R$)

If $1 \leftarrow DECP(D, PK_B, R)$ return $\perp$,

Else if $1 \leftarrow DECP(\Omega_i, PK_B, R)$ and $m \| PK_A \| PK_B = m_i \| (PK_A)_i \| (PK_B)_i$

return σ_i *//there is an entry* ($\Omega_i \| (m_i \| (PK_A)_i \| (PK_B)_i \| ?)$ *in* L_2^H

Else if $\hbar = \hbar_i$, return $\hbar_i$ *//there is an entry* $(\hbar_i, h_i)$ *in* L_1^H

Else $h_i \xleftarrow{\$} Z_l^*$, $\hbar_i \leftarrow \hbar$, Add ($\hbar_i$, h_i) into L_1^H."

If $EXP_{SGC}^{ind-cca2}(ASC)$ makes queries to random Oracle O^{SC}, *AECG* will reply with simulator Sim_SC.

$\mathrm{Sim}\_SC(L_2^K, L_2^H, PK_A, PK_B, m)$ =

"On input (L_2^K, L_2^H, PK_A, PK_B, m): // $\hbar$ =($m \| PK_A \| PK_B \| R$)

$\sigma \xleftarrow{\$} \{0,1\}^{L^K(1^k)}$, $c \leftarrow E_\sigma(m)$,

$h \xleftarrow{\$} Z_l^*$, $s \xleftarrow{\$} Z_l^*$,

$\Omega \leftarrow sP - hPK_A$, $\Omega_i \leftarrow \Omega$,

$\sigma_i \leftarrow \sigma$, $m_i \leftarrow m$, $h_i \leftarrow h$,

Add entry $\Omega_i \| (?,\sigma_i)$ into L_2^K,

Add entry ($m \| PK_A \| PK_B \| (?, h_i)$) into L_2^H,

C = (c, h, s) and return C."

If $EXP_{SGC}^{ind-cca2}(ASC)$ makes queries to O^{USC}, *AECG* will reply with simulator Sim_USC.

$\mathrm{Sim}\_USC(L^K, L^H, D, PK_A, PK_B, C)$ =

"On input (L^K, L^H, D, PK_A, PK_B, C):

Parse C into (c, h, s), $\Omega \leftarrow sP - hPK_A$,

If $\Omega = D$ return $\perp$,

If $\exists$ (R_i, σ_i) in L_1^K s.t. $1 \leftarrow DECP(\Omega_i, PK_B, R_i)$ or $\exists\ \Omega_i \| (?,\sigma_i)$ in L_2^K s.t. $\Omega_i = \Omega$

Then $\sigma' \leftarrow \sigma_i$,

Else $\sigma' \xleftarrow{\$} \{0,1\}^{L_K(1^k)}$, $\Omega_i \leftarrow \Omega$, $\sigma_i \leftarrow \sigma'$,

Add entry $\Omega_i \| (?,\sigma_i)$ into L_2^K, $m \leftarrow D_{\sigma'}(c)$,

If $\exists (\hbar_i, h_i)$ in L_1^H s.t. $1 \leftarrow DECP(\Omega_i, PK_B, R_i)$ or $\exists\ (\Omega_i \| (m_i \| (PK_A)_i \| (PK_B)_i \| ?), h_i)$ *in* L_2^H s.t. $\Omega_i = \Omega$, $m_i = m$, $(PK_A)_i \| (PK_B)_i = (PK_A) \| (PK_B)$,

Then $h' \leftarrow h_i$,

Else $\Omega_i \leftarrow \Omega$, $(PK_A) \| (PK_B) \leftarrow (PK_A)_i \| (PK_B)_i$, $m_i \leftarrow m$, $h_i \xleftarrow{\$} Z_l^*$,

Add entry ($\Omega_i \| (m_i \| (PK_A \| PK_B)_i \| (?, h_i))$ into L_2^H,

If $h = h_i$ return m, else return $\perp$."

Based on Theorem 3.2.1 we formulate an *AECG* attacker on ECGP; apparently contradicting Hypothesis 3.2.1, thus prove the confidentiality of the improved signcryption scheme.

The *AECG* attacker on ECGP is formulated as follows.

$AECG(T, D, E)$=

"On input (T, $D = dP$, $E = eP$):

$h^*, s^* \xleftarrow{\$} Z_l^*$, $PK_A \leftarrow (h^*)^{-1}(s^*P + D)$,

$PK_B \leftarrow E$, $\sigma^* \xleftarrow{\$} Z_l^*$,

$C^* = (c^*, h^*, s^*) \leftarrow SC_{sk_A, PK_B}(LR(m_0, m_1, b))$,

// $c^* \leftarrow E_{\sigma^*}(m_b)$, and *the random oracle queries* O^* *are replaced with random oracle simulator* Sim_O^*.

If SC_{sk_A,PK_B} has ever queried $\mathrm{Sim}\_K(R) = \perp$,

Halt and return R,

If SC_{sk_A,PK_B} has ever queried $\mathrm{Sim}\_H(\hbar) = \perp$,

Halt and return $\hbar$ (the rightmost $|R|$ bits of $\hbar$),

$EXP_{SGC}^{ind-cca2}(ASC) =$ "

// *random oracle queries* O^* *are also replaced with random oracle simulator* Sim_O^*.

If $EXP_{SGC}^{ind-cca2}(ASC)$ has ever queried $\mathrm{Sim}\_K(R) = \perp$,

Halt and return R,

If $EXP_{SGC}^{ind-cca2}(ASC)$ has ever queried $\mathrm{Sim}\_H(\hbar) = \perp$,

Halt and return $\hbar$ (the rightmost $|R|$ bits of $\hbar$),

$b \leftarrow EXP_{SGC}^{ind-cca2}(ASC)$,

Return R."

Let *ASC* be an attacker against *IND-CCA2* security of signcryption in time t , using q_{sc} queries to its signcryption oracle, q_{usc} queries to its unsigncryption oracle and (q_k , q_h) queries to its random oracles. *AECG* is an attacker against *ECGP* security of elliptic curve in time t' , using $q_{O^{DECP}}$ queries to its DECP Oracle O_T^{DECP}. *AI* is an attacker against *PI* security of the symmetric key encryption in time t'' . From the *AECG* algorithm formulated above, the following bound holds. More details about the probability proof of the theorem can be found in [5, 12, 13, 14].

$$Adv_{SGC}^{ind-cca2}(t, q_{sc}, q_{usc}, q_k, q_h) \le 2\,Adv_T^{ECGP}(t', q_{O^{DECP}}) + 2\,Adv_{\Upsilon}^{pi}(t'') + q_{sc}\left(\frac{q_k + q_h + q_{sc} + q_{usc} + 2}{2^{L^K(1^k)-1}}\right) + \frac{q_h + 2q_{usc}}{2^{L^K(1^k)-1}}. \quad (6)$$

$$\Rightarrow Adv_{SGC}^{ind-cca2}(t, q_{sc}, q_{usc}, q_k, q_h)/2 - q_{sc}\left(\frac{q_k + q_h + q_{sc} + q_{usc} + 2}{2^{L^K(1^k)}}\right) - \frac{q_h + 2q_{usc}}{2^{L^K(1^k)}} - Adv_{\Upsilon}^{pi}(t'') \le Adv_T^{ECGP}(t', q_{O^{DECP}}). \quad (7)$$

As *ASC* and *AECG* are reasonably resource bounded,

$$\Rightarrow q_{sc}\left(\frac{q_k + q_h + q_{sc} + q_{usc} + 2}{2^{L^K(1^k)}}\right) - \frac{q_h + 2q_{usc}}{2^{L^K(1^k)}}$$ is negligible.

And with the assumption Υ is passive indistinguishable, $Adv_{\Upsilon}^{pi}(t'')$ is negligible too.

$$\Rightarrow Adv_{SGC}^{ind-cca2}(t, q_{sc}, q_{usc}, q_k, q_h)/2 \le Adv_T^{ECGP}(t', q_{O^{DECP}}). \quad (8)$$

On account of all the above analyses, if the *IND-CCA2*security of signcryption will be broken by *ASC* with non-negligible advantage, so will the *ECGP* security of elliptic curve by *AECG* with non-negligible advantage. Therefore, *S-ECSC* achieves confidentiality in the *IND-CCA2* model, as desired.

3.3 Unforgeability of *S-ECSC*

Definition 3.3.1 Unforgeability of Signcryption. Let $\Sigma = (GC, GK, SC, USC)$ be a signcryption scheme, and let A be an algorithm that has access to a signcryption oracle and returns a pair of strings. We consider the following experiment:

Experiment $EXP_{SGC}^{uf-cma}(A)$

$(sk_A, PK_A) \leftarrow GK(A, 1^k)$,

$(sk_B, PK_B) \leftarrow GK(B, 1^k)$,

$(m, C') \xleftarrow{\$} A^{SGC(sk_A, PK_B,)}(PK_A, PK_B)$.

If the following are true return 1 else return 0:

1. $m \leftarrow USC(sk_B, PK_A, C')$,
2. $m \in Message(sk_A, PK_B)$,
3. m is not a query of A to its signcryption oracle.

The *UF-CMA advantage* of A is defined as

$$Adv_{SGC}^{uf-cma}(A) = \Pr(EXP_{SGC}^{uf-cma}(A) = 1). \quad (9)$$

To be specific, the *UF-CMA advantage* can be concluded as a function $\varepsilon(k)$ defined by

$$\varepsilon(k) = \Pr[(sk_A, PK_A) \leftarrow GK(A, 1^k), (sk_B, PK_B) \leftarrow GK(B, 1^k), (m, C') \xleftarrow{\$} A^{SGC(sk_A, PK_B,)}(PK_A, PK_B): m \leftarrow USC(sk_B, PK_A, C')]. \quad (10)$$

A signcryption is un-forgeable under chosen message attack if the *UF-CMA advantage* of any attacker *A* with reasonably restricted resources (time-complexity, frequency and length of queries) is negligible.

Hypothesis 3.3.1 (ECDLP is hard). Let T be an elliptic curve, and let *A* be an algorithm that has access to a elliptic curve oracle and returns a string. We consider the following experiment:

Experiment $EXP_T^{ECDLP}(A)$

$x \xleftarrow{\$} Z_l^*, Q \leftarrow xP$,

$x' \leftarrow A^T(P, Q)$.

If $x' = x$ return 1 and return 0 otherwise.

The *ECDLP advantage* of A is defined as

$$Adv_T^{ECDLP}(A) = \Pr(EXP_T^{ECDLP}(A) = 1). \quad (11)$$

Given elliptic curve T and secure parameter 1^k, the probability of solving ECDLP in time t is $\delta(1^k, T)$ which is negligible, that is

$$\delta(1^k, T) = \Pr[1^k, x \xleftarrow{\$} Z_l^*, Q \leftarrow xP : EXP_T^{ECDLP}(A) = 1]. \quad (12)$$

Definition 3.3.2 Gap Elliptic Curve Discrete Logarithm (GECDL). Let $T = (q, a, b, P, l, h)$ be an elliptic curve and *AGL* an attacker on GECDL, O_T^{DECP} is DECP Oracle, let's consider the following experiment:

Experiment $EXP_T^{GECDL}(AGL)$

$d \xleftarrow{\$} Z_l^*, D \leftarrow dP,$

$d' \leftarrow AGL_T^{O_T^{DECP}(\cdot,\cdot)}(D),$

If $d' = d$ return 1 else return 0.

The *GECDL advantage* of *AGL* is defined as

$$Adv_T^{GECDL}(AGL) = \Pr(EXP_T^{GECDL}(AGL) = 1). \quad (13)$$

Hypothesis 3.3.2 (GECDL is hard). Given elliptic curve T and secure parameter 1^k, the probability of solving GECDL in time t is negligible, that is

$$\delta(1^k, T) = \Pr[1^k, d \xleftarrow{\$} Z_l^*, d' \leftarrow AGL_T^{O_T^{DECP}(\cdot,\cdot)}(D), EXP_T^{GECDL}(AGL) = 1]. \quad (14)$$

Proof of unforgeability: Let *ASC* be an attacker against *UF-CMA* security of signcryption executing in time t, using q_{sc} queries to its signcryption oracle, q_{usc} queries to its unsigncryption oracle and (q_k, q_h) queries to its random oracles. *AGL* is an attacker against *GECDL* security of elliptic curve executing in time t', using $q_{O^{DECP}}$ queries to its DECP Oracle O_T^{DECP}. From the algorithm formulated above, the following bound holds. Similarly, more details about the probability proof of the theorem can be found in [12, 13, 14].

$$Adv_{SGC}^{uf-cma}(t, q_{sc}, q_{usc}, q_k, q_h) \leq 2\sqrt{Adv_T^{GECDL}(t', q_{O^{DECP}})} + \left(\frac{q_{sc}(q_k + q_h + q_{sc}) + q_h + 1}{2^{L^K(1^k)-1}}\right). \quad (15)$$

As *ASC* is reasonably resource bounded,

$$\Rightarrow \frac{q_{sc}(q_k + q_h + q_{sc}) + q_h + 1}{2^{L^K(1^k)-1}} \text{ is negligible}$$

$$\Rightarrow Adv_{SGC}^{uf-cma}(t, q_{sc}, q_{usc}, q_k, q_h) \leq 2\sqrt{Adv_T^{GECDL}(t', q_{O^{DECP}})}. \quad (16)$$

If the *UF-CMA* security of signcryption will be broken by *ASC* with non-negligible advantage, so will the *GECDL* security of elliptic curve by *AGL* with non-negligible advantage. Therefore, *S-ECSC* achieves unforgeability in the *UF-CMA* model, as desired.

3.4 Nonrepudiation of *S-ECSC*

Definition 3.4.1 Non-repudiation of signcryption. It is computationally feasible for a third party to settle a dispute between signcryption sender and receiver in an event where sender denies the fact that he is the originator of signcryption.

Definition 3.4.2 Relation Map. A relation is a map defined as

$$\Re_{E,\pi}^H : \{0,1\}^* \times \{0,1\}^* \to \{0,1\}. \quad (17)$$

For every string $x \in \{0,1\}^*$, random oracle $H \in 2^\infty$ and $E, \pi \in \{0,1\}^*$, it satisfies

$$\Re_{E,\pi}^H(x,x) = \Re_{E,\pi}^H(x,0^*) = 0. \quad (18)$$

Besides, $\Re_{E,\pi}^H$ must be computable by a deterministic polynomial time algorithm $A^H(x, y, E, \pi)$. A malleability adversary s is a pair of probabilistic polynomial time algorithms (P, Q) with access to random oracle $H \in 2^\infty$.

The security notion of non-malleability for encryption scheme was introduced by Dolev, Dwork and Naor[15]. In this section, we generalize non-malleability into a more comprehensive security notion applicable to signcryption as well.

Definition 3.4.3 Non-malleability of Signcryption. A signcryption scheme $\Sigma = (GC, GK, SC, USC)$ is non-malleable if any adversary can not by witnessing signcryption generating of a message m or querying a signcryption oracle, produce the signcryption text of a related message m'.

To be specific, a signcryption scheme is non-malleable if for every relation $\Re$ and every malleability adversary $s = (P, Q)$, there is a deterministic time algorithm Q' so that $|\tau(k) - \tau_*(k)|$ defined as follows is negligible.

$$\tau(k) = \Pr[H \leftarrow 2^\infty; (SC, USC) \leftarrow K(1^k); \pi \leftarrow P^H(SC); x \leftarrow \pi^H(1^k); \beta \leftarrow SC^H(x); \beta' \leftarrow Q^H(SC, \pi, \beta) : \Re_{E,\pi}^H(x, USC^H(\beta') = 1)], \quad (19)$$

$$\tau_*(k) = \Pr[H \leftarrow 2^\infty; (SC, USC) \leftarrow K(1^k); \pi \leftarrow P^H(SC); x \leftarrow \pi^H(1^k); \beta'_* \leftarrow Q'^H(SC, \pi) : \Re_{E,\pi}^H(x, USC^H(\beta'_*) = 1)]. \quad (20)$$

Theorem 3.4.1 The short signcryption scheme *S-ECSC* achieves non-repudiation security.

Proof of non-repudiation: In signcryption schemes, unforgeability implies non-repudiation if there is no duplication of the signcryption text. If the signcryption scheme is forgeable or malleable, the signcryption generator will have opportunity to repudiate.

In *S-ECSC,* the map $K : E(F_q) \to \{0,1\}^{L_K(1^k)}$ and $H : \{0,1\}^* \to Z_l^*$ are both unique, distinct (m_1 , r) and(m_2 , r) will generate different signcryption text C =(c , h , s). Furthermore, the scheme can be reinforced by state padding. The state padding not only ensures different signcryption text for distinct (m_1 , r) and (m_2 , r), but for the same original message (m , r) with different state information. Thus, the above signcryption scheme satisfies: as to $|\tau(k) - \tau_*(k)|$ for every c there is a k_c such that $|\tau(k) - \tau_*(k)| \le k^{-c}$ for every $k \ge k_c$. Thus the signcryption text C produced by SC (sk_A , PK_B , m) is not duplicable, and with the unforgeability proof of *S-ECSC* in *UF-CMA* model in section 3.3, we can come to the conclusion that *S-ECSC* achieves non-repudiation security, as desired.

4 Efficiency of *S-ECSC*

In this section, the short signcryption scheme *S-ECSC* will be compared with other typical schemes including discrete logarithm based signcryption *SCS* [5], *B&D* [16], *KCDSA*[17], *SC-DSA*[18] and RSA based signcryption *TBOS*[19] and elliptic curve based scheme *ECSCS*[20] and *ECGSC*[21].

In these schemes, such computing as modular exponential, modular inverse and elliptic curve addition ,elliptic curve scalar multiplication should be taken into comparison for computing complexity, while computing cost of modular addition, modular multiplication, hash, symmetric encryption/decryption are negligible. To ensure the security of the basic cryptographic primitives, the minimum security parameters of these cryptosystems recommended for the current practice are as follows: for DLP, $|p|$ =1024bits, $|q|$ =160bits. For RSA, $|N|$ =1024bits; for ECC, $|q|$ =131bits (79, 109 may also be chosen), $|l|$ =160bits. The block length of the block cipher is 64bits. The length of secure hash function is 128bits.

Schemes	*GC+GK*	*SC*	*USC*	*EC*	*PV*	*Length of C*
SCS	2*E*	1*E*+1 *I*	2*E*	/	/	|*D* (·) |+|*KH* (·)| + |*q*|
B&D	2*E*	2*E*+1 *I*	3*E*	0	2*E*	|*D* (·) |+ |*h*|+ |*q*|
KCDSA	2*E*	2*E*	3*E*	Save *r,s* or 3*E*	2*E*	|*D* (·) |+ |*h*|+ |*q*|
SC-DSA	2*E*	2*E*+2 *I*	3*E*+1 *I*	Save *r,s* or 2*E*+1 *I*	2*E*+1 *I*	|*D* (·) |+2|*q*|
TBOS	2*E*+2*I*	2*E*	2*E*	0	2*E*	| *N* |
ECSCS	2*kP*	1 *kP* +1 *I*	2*kP*	/	/	|*D* (·) |+ |*h*|+ |*n*|
ECGSC	2*kP*	2*kP*+1 *I*	3*kP*+1 *I*	0	2*kP*+1 *I*	|*l*|+ |*LH* (.)|+ 2|*q*|
S-ECSC	2*kP*	1*kP*	2*kP*+1 *I*	0	2*kP*+1 *I*	|*D* (·) |+|*h*|+ 2|*p*|

Table 1: Comparison of computing and communication cost

Notes of notations: 1. *GC* denotes the common parameters generation algorithm, *GK* denotes the keys generation algorithm; *SC* denotes the signcryption algorithm; *USC* denotes the unsigncryption algorithm; *EC* denotes the extra computation to accomplish public verifiability; *PV* denotes the public verification by a third party. *Length of C* denotes the length of signcryption text. 2. *E* denotes modular exponential; *I* denotes modular inverse; *KP* denotes scalar multiplication on elliptic curve. / denotes there is no relevant computation. 3. |*D* (·) |denotes the block length of block cipher, |*h*| denotes the outputs length of secure hash function, |*KH* (·)| denotes the length of key hash function in *SCS*, the same as |*h*|, |*LH* (.)| denotes the length of hash function with long message digest, much larger than |*h*|.

Remark 1. (Comparison with DLP based signcryption schemes). *SCS* is the fastest scheme in all of the four DLP based schemes (*SCS, B&D, KCDSA* and *SC-DSA*). Based on the result of Koblitz and Menezes [22], the computing cost in key generation in our scheme is 1/8 of that in *SCS*; signcryption operation in ours is about 1/8 of that in *SCS*, and unsigncrption is about 1/8 of that in *SCS*. To sum up, *S-ECSC* reduces about 87%computating cost compared with *SCS*.

Remark 2. (Comparison with RSA based signcryption scheme). As per the result of [22], the computing cost in key generation in our scheme is about 1/8 of that in *TBOS*; signcryption operation in ours is about 1/16 of that in *TBOS*, and unsigncrption is about 1/8 of that in TBOS, achieving a total 89%computating cost reduction over *TBOS*.

Remark 3. (Comparison with other ECC based schemes). The computing cost in key generation are the same; signcryption cost in ours is slightly lower than that in *ECSCS* while unsigncrption of *ECSCS* is slightly lower than ours, total resulting an equal computing cost, yet *ECSCS* proves to be unsuitable for public verifying. Although *ECGSC* is publicly verifiable, its computing cost in signcryption and unsigncryption is much larger than *S-ECSC*, resulting in a much higher total computing cost.

Remark 4. (Comparison of communication cost). As per the comparison of signcryption text length, except for RSA based *TBOS* signcryption, *S-ECSC* has the lowest communication cost in Elgamal type signcryption schemes.

Therefore, we may come to the conclusion that our short signcryption scheme *S-ECSC* has the highest efficiency and the lowest communication cost in all of the publicly verifiable schemes.

5 Application of *S-ECSC* in Secure Communication of IOT

In order to achieve confidentiality and integrity for secret key in the Internet of things, a secure channel should be established for the distribution and transmission in key management schemes. Meanwhile, the special network environment in IOT, such as wireless connection, micro-terminals and restricted resources, makes it necessary to design schemes of high efficiency which can fulfil the same function with much smaller computing and communication cost than traditional schemes. With its superiority in computing and communication, *S-ECSC* greatly improves the efficiency in key management in IOT in terms of key distributing time and bandwidth resources and better satisfies the requirement of secure protocols in key management. In this section, we will take the key management schemes in [23,24,25] as an example and propose a key applying and distributing scheme in key management of IOT based on *S-ECSC*. With this *S-ECSC* based scheme, we analyse the method and importance to apply *S-ECSC* in secure wireless communication for the Internet of things. The key applying and distributing protocol in key management of IOT based on *S-ECSC* is as follows.

(1)Initializing

PKG selects the system parameter, including the parameters in *S-ECSC*.

$sk_A \in Z_l^*$, $PK_A = sk_A P \neq O$, (sk_A, PK_A) is the private/public key pair for one of the distributed terminals *A*. $sk_B \in Z_l^*$, $PK_B = sk_B P \neq O$, (sk_B, PK_B)is the private / public key pair of PKG.

(2) Key applying

Step1: Terminal *A* encodes the applying request data {ID_A, Message} into plaintext $m \in Message$ (sk_A, PK_B), and applies the signcryption algorithm on m.

$$C \leftarrow SC\ (sk_A, PK_B, m).$$

Then signcryption text C will be transmitted to PKG.

Step2: PKG applies the unsigncryption algorithm on signcryption text C.

$$m \leftarrow USC\ (sk_B, PK_A, C).$$

Thus, PKG recovers plaintext m, and simultaneously fulfils authentication on identity of terminal *A* and examines the integrity of message m.

(3)Key generating and distributing

Step1: PKG generates secret key $k \in Message$ (sk_B, PK_A) with the key generating algorithm $KG(1^k)$, and applies the signcryption algorithm on k.

$$C' \leftarrow SC\ (sk_B, PK_A, k).$$

Then the signcryption text C' will be transmitted to terminal *A*.

Step2: Terminal *A* applies the unsigncryption algorithm on C'.

$$k \leftarrow USC\ (sk_A, PK_B, C').$$

Thus, *A* recovers secret key k, and fulfils authentication on identity of PKG and examines the integrity of secret key k.

With the application of *S-ECSC* in key applying and distributing, the above scheme achieves secure and efficient transmission of terminal secret key via the public channel of IOT. The scheme fulfils the integrated functions of encryption and digital signature in a single step and simultaneously achieves confidentiality, integrity and non-repudiation for the secret terminal key and other signcrypted message; whereas, the computing and communication cost is significantly smaller than traditional schemes.

6 Conclusions

The study of signcryption algorithms suitable for IOT network environment and its application in IOT security schemes is an important direction in cryptography; it is more of a requirement from the rapid development of the Internet of things than just a requirement from the theoretical or applied cryptography research. In the paper, we propose a publicly verifiable short signcryption scheme *S-ECSC* suitable for secure communication in the Internet of things; and prove the provable security of *S-ECSC* under the Random Oracle model, including confidentiality, unforgeability and non-repudiation security. At last, we take key generating and distributing protocol for different terminals of distributed key management in IOT as an example, and analyze the method and importance in the application of *S-ECSC* into secure protocols in IOT.

Compared with other typical discrete logarithm, RSA and elliptic curve based signcryption schemes; *S-ECSC* achieves about 87% reduction in computing cost than DLP signcryption schemes and about 89% reduction compared with RSA schemes. And it has the lowest communication cost in the Elgamal type schemes. Therefore, security schemes based on *S-ECSC* are most suitable for such circumstances as with restricted computation ability and integrated space, circumstances with limited bandwidth yet requiring for high-speed operation. Besides, the computational problems ECGP and GECDL in the paper can also be basis of security proof for other elliptic curve based schemes.

Acknowledgement

The authors should thank the anonymous reviewers for their constructive advice and comments to the paper, with which we can greatly improve our work.

References

[1] ITU(2005). ITU Internet Report 2005: The Internet of Things. ITU.

[2] Atzori L,Iera A,Giacomo M(2010). The Internet of Things : a survey. Computer Networks, pp.2787-2805.

[3] EpoSS (2010).Internet of Things in 2020:Roadmap for the Future.EpoSS.

[4] Zhu Hongbo,Yang Longxiang,Yu Quan(2010). Investigation of Technical Thought and Application Strategy for the Internet of Things . Journal of Communication, pp. 2-9.

[5] Zheng Y(1997). Digital signcryption or how to achieve cost (signature &encryption) << cost (signature) + cost(encryption). Advances in Cryptoloty-CRYPTO'97, Lecture Notes in Computer Science vol.1294,Springer-Verlag, Berlin ,pp.165-179.

[6] Zhang Chuanrong. Zhang Yuqing . Li Fageng and Xiao Hong(2010). New Signcryption Algorithm for Secure Communication of ad hoc Networks. Journal of Communications, pp.19-24.

[7] Luo Ming, Zuo Chunhua and Wen Yingyou (2010).Signcryption-based fair exchange protocol. Journal of Communications, pp.87-93.

[8] Kim H, Song J, Yoon H(2007). A practical approach of ID-based cryptosystem in ad hoc networks. Wireless Communications and Mobile Computing. pp.909-917.

[9] Li F G, Hu Y P, Zhang C R(2007). An identity-based signcryption scheme for multi-domain ad hoc networks. ACNS 2007, LNCS 4521, Springer-Verlag, Berlin , pp.373-384.

[10] Chen, Weidong. Feng Dengguo(2005). Some Apiplications of Signcryption to Distributed Protocols. Chinese Journal of Computers, pp.1421-1430.

[11] Kamat P, Baliga A, Trappe W(2006). An identity-based security framework for VANETs. VANET'06. Los Angeles, California, USA, pp.94-95.

[12] Baek J, Steinfeld R and Zheng, Y(2002). Formal Proofs for the Security of Signcryption. Public Key Cryptography'02, Lecture Notes in Computer Science vol.2274, Springer-Verlag,Berlin,pp.80-98.

[13] Shoup V (2004). Sequences of Games: A Tool for Taming Complexity in Security Proofs , International Association for Cryptographic Research (IACR) ePrint Archive: Report 2004/332.

[14] Bellare M and Rogaway P (2004).The Game-Playing Technique, International Association for Cryptographic Research (IACR) ePrint Archive: Report 2004/331.

[15] Dolev D, Dwork C and Naor M (1991).Non-malleable cryptography. 23rd ACM Symposium on Theory of Computing. IEEE ,New York.

[16] Bao, F and Deng R H(1998). A signcryption scheme with signature directly verifiable by public key. Public Key Cryptography'98, Lecture Notes in Computer Science vol.1431, Springer-Verlag, Berlin,pp.55-59

[17] Yum D H and Lee P J(2002). New Signcryption Schemes based on KCDSA. Proceedings of the 4thInternational Conference on Information Security and Cryptology, Seoul, South Korea, pp. 305-317.

[18] Shin J B, Lee Kand Shim K(2003). New DSA-Verifiable Signcryption Schemes. Proceedings of the 5th International Conference on Information Security and Cryptology, Seoul, South Korea, pp.35-47.

[19] Malone-Lee J and Mao W (2003). Two birds one stone: Signcryption using RSA. Topics in Cryptology – Cryptographers'Track, RSA Conference 2003, Lecture Notes in Computer Science vol.2612, Springer-Verlag, Berlin,pp.210-224.

[20] Zheng Y and Imai H(1998). How to construct efficient signcryption schemes on elliptic curves. Information Processing Letters, pp.227-233.

[21] Han Yiliang, Yang Xiaoyuan and etc(2006). ECGSC:Elliptic Curve Based Generalized Signcryption. Proceedings of The 3th International Conference on Ubiquitous Intelligence and Computing, Springer-Verlag, Berlin,pp.956-965.

[22] Koblitz N , Menezes A and Vanstone S(2000). The state of elliptic curve cryptography. Designs, Codes and Cryptography, pp.173-193.

[23] Li G S, Han W B(2005). A new scheme for key manegement in ad hoc networks. ICN2005, LNCS 3421,Springer-Verlag, Berlin,pp.242-249.

[24] Gu Jing-jing,Chen Song-Can,Zhuang Yi(2010). Wireless Sensor Networks-Based Topology Structure for the Internet of Things Location.Chinese Journal of Computer ,pp.1548-1556.

[25] Chen Juan,Fang Binxing,Yin Lihua(2010). A Source-Location Privacy Preservation Protocol in Wireless Sensor Networks Using Source-Based Restriced Flooding.Chinese Journal of Computer, pp.1736-1747.

Conference report

Information Society 2011
14th International Multiconference
10–14 October 2011, Ljubljana, Slovenia

IS 2011

http://is.ijs.si

Each fall for the last 14 years the Jožef Stefan Institute has hosted the Information Society multiconference. The multiconference usually consists of 10 or so conferences, each with its own topic and program committee, but all related to the information society in the technological and sociological sense. Most of the conferences are organized every year, some take place biannually and some are one-time events. In 2011 the multiconference comprised the following conferences:

- Cognitive Sciences
- Cognitonics
- Collaboration, Software and Services in Information Society
- Data Mining and Data Warehouses
- Education in Information Society
- Facing Demographic Challenges
- Intelligent Systems
- Internet and Slovenia: 1985–1995
- Robotics.

193 papers were presented by 299 authors from 13 countries. The multiconference opened with the talks by Zoran Stančič, the representative of the EU Directorate-General for the Information Society and Media, and Norbert Kroó, a distinguished member of the Hungarian Academy of Sciences. Information-society awards were given to Vladimir Batagelj for life work in theoretical computer science and the shaping of the Slovenian computer-science community, and Janez Brank for recent achievements in the organization of ACM computer-science competitions. For the first time information strawberry and lemon were awarded for the best and worst information-society public services. The lemon was given to the countrywide real-estate inventory which was flawed in numerous ways, while the strawberry was awarded for the streamlining of the access to healthcare data.

In the conference Facing Demographic Challenges have – after five years since it was organized for the first time – participated not only scientists, but also representatives of political parties. They were shown basic demographic projections indicating that while the total population of Slovenia will remain stable due to immigration, the fraction of the native population will halve in about 80 years. Such »demographic winter«, a term coined by the French scientist Gèrard François Dumont, is typical for most of Europe. The panel composed of the representatives of the parties discussed how their policies may address this phenomenon. While the panel did not reach any insightful conclusions, we welcome it as an important step in realizing that the governing bodies must listen to the scientific community.

The conference Internet and Slovenia: 1985–1995 focused on the turbulent events during the introduction of the Internet and the World Wide Web in Slovenia, and in 1991 when Slovenia declared independence from the former Yugoslavia. In the opinion of several independent observers, the contribution of the Slovenian computer scientists and professionals to communicating the situation during the struggle for independence may have been more important than that of the politicians. The fact that the Internet was better established in Slovenia than in the rest of the former Yugoslavia gave Slovenia a crucial advantage in the communication with the rest of Europe and USA.

Attending the multiconference, one must acknowledge that 2011 was another year of steady advance of the information society, seemingly unaffected by the economic crisis that plagues the world. Moore's law still holds, making computers and other electronic devices ever faster, smaller and more ubiquitous. Virtually everybody in the developed world has the access to numerous computing devices, robotics and automation are increasingly present in the industry and are beginning to trickle into everyday life in the form of robotic vacuum cleaners and lawnmowers. Like the progress in hardware, the amount of information and knowledge generated, transmitted and stored is likewise increasing rapidly. The ability of every person to educate him- or herself and make informed decisions is greater than ever before. This gives us the opportunity to make a better live for everybody, if only we seize it. To do so, the awareness of what information society can do for those who live in it must be spread, which is one of the main goal of the multiconference.

Organization committee:
Matjaž Gams (chair), Mitja Luštrek (co-chair)

Call for Papers:

Information Society 2012

15th International Multiconference

8–12 October 2012, Ljubljana, Slovenia

IS 2012

http://is.ijs.si

The concepts of information society, information era, infosphere and infostress have by now been widely accepted. But what does it really mean for the society, science, technology, education, governments, our lives? The Information Society multiconference deals with information technologies, which are of major importance for the development of Europe. Inormation Society 2012 will serve as a forum for the world-wide and national community to explore further research directions, business opportunities and governmental policies. For these reasons we host a scientific meeting in the form of a multiconference, which will consist of several independent conferences with themes essential for the development of the information society. The main objective is the exchange of ideas and developing visions for the future of information society. IS 2012 is a high-quality multidisciplinary conference covering major recent scientific achievements.

Independent conferences within IS 2012

- Cognitive Sciences
- Collaboration, Software and Services in Information Society
- Data Mining and Data Warehouses
- Education in Information Society
- Facing Demographic Challenges
- Intelligent Systems
- Internet and Slovenia: 1985–1995
- Language Technologies
- Robotics
- 100 years of Turing and 20 years of SLAIS

Submission and dates

Submitted papers must be original and not currently under review by another conference or journal. They should address the topic of one of the conferences within IS 2012, and should preferably be accessible to a wide audience. All papers will be peer-reviewed.

The papers can be in English or in Slovene, depending on the conference, and should be up to four pages long in most cases. The deadlines for submission are still to be determined, but are expected to range from late spring to the end of August, again depending on the conference. They will be posted on the conference website (http://is.ijs.si). Both printed and electronic proceedings on a USB flash drive will be available.

Organization

Main organizer: Jožef Stefan Institute, Department of Intelligent Systems
Organization committee: Matjaž Gams (chair), Mitja Luštrek (co-chair), Vedrana Vidulin and others

CONTENTS OF *Informatica* Volume 35 (2011) pp. 1–533

Papers

Editorials

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 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 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 a manuscript at: http://www.informatica.si/Editors/PaperUpload.asp. 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 seventeen 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 intelligent systems in the European computer science, informatics and cognitive 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, Raheel Ahmad, Cutting Alfredo, Hameed Al-Qaheri, Gonzalo Alvarez, Wolfram Amme, Nicolas Anciaux, Rajan Arora, Costin Badica, Zoltán Balogh, Andrea Baruzzo, Borut Batagelj, Norman Beaulieu, Paolo Bellavista, Steven Bishop, Marko Bohanec, Zbigniew Bonikowski, Borko Bosković, Marco Botta, Pavel Brazdil, Johan Brichau, Andrej Brodnik, Ivan Bruha, Maurice Bruynooghe, Wray Buntine, Dumitru Dan Burdescu, Yunlong Cai, Juan Carlos Cano, Tianyu Cao, Norman Carver, Marc Cavazza, Jianwen Chen, L.M. Cheng, Chou Cheng-Fu, Girija Chetty, G. Chiola, Yu-Chiun Chiou, Ivan Chorbev, Shauvik Roy Choudhary, Sherman S.M. Chow, Lawrence Chung, Mojca Ciglarič, Jean-Noël Colin, Vittorio Cortellessa, Jinsong Cui, Alfredo Cuzzocrea, Darko Čerepnalkoski, Gunetti Daniele, Grégoire Danoy, Manoranjan Dash, Paul Debevec, Fathi Debili, Carl James Debono, Joze Dedic, Abdelkader Dekdouk, Bart Demoen, Sareewan Dendamrongvit, Tingquan Deng, Anna Derezinska, Gaël Dias, Ivica Dimitrovski, Jana Dittmann, Simon Dobrišek, Quansheng Dou, Jeroen Doumen, Erik Dovgan, Branko Dragovich, Dejan Drajic, Jozo Dujmovic, Umut Riza ErtÃijrk, CHEN Fei, Ling Feng, YiXiong Feng, Bogdan Filipič, Iztok Fister, Andres Flores, Vladimir Fomichov, Stefano Forli, Massimo Franceschet, Alberto Freitas, Jessica Fridrich, Scott Friedman, Chong Fu, Gabriel Fung, David Galindo, Andrea Gambarara, Matjaž Gams, Maria Ganzha, Juan Garbajosa, Rosella Gennari, David S. Goodsell, Jaydeep Gore, Miha Grčar, Daniel Grosse, Zhi-Hong Guan, Donatella Gubiani, Bidyut Gupta, Marjan Gusev, Zhu Haiping, Kathryn Hempstalk, Gareth Howells, Juha Hyvärinen, Dino Ienco, Natarajan Jaisankar, Domagoj Jakobovic, Imad Jawhar, Yue Jia, Ivan Jureta, Dani Juričić, Zdravko Kačič, Slobodan Kalajdziski, Yannis Kalantidis, Boštjan Kaluža, Dimitris Kanellopoulos, Rishi Kapoor, Andreas Kassler, Daniel S. Katz, Samee U. Khan, Mustafa Khattak, Elham Sahebkar Khorasani, Ivan Kitanovski, Tomaž Klobučar, Ján Kollár, Peter Korošec, Valery Korzhik, Agnes Koschmider, Jure Kovač, Andrej Krajnc, Miroslav Kubat, Matjaz Kukar, Anthony Kulis, Chi-Sung Laih, Niels Landwehr, Andreas Lang, Mohamed Layouni, Gregor Leban, Alex Lee, Yung-Chuan Lee, John Leggett, Aleš Leonardis, Guohui Li, Guo-Zheng Li, Jen Li, Xiang Li, Xue Li, Yinsheng Li, Yuanping Li, Shiguo Lian, Lejian Liao, Ja-Chen Lin, Huan Liu, Jun Liu, Xin Liu, Suzana Loskovska, Zhiguo Lu, Hongen Lu, Mitja Luštrek, Inga V. Lyustig, Luiza de Macedo, Matt Mahoney, Domen Marinčič, Dirk Marwede, Maja Matijasevic, Andrew C. McPherson, Andrew McPherson, Zuqiang Meng, France Mihelič, Nasro Min-Allah, Vojislav Misic, Vojislav Mišić, Mihai L. Mocanu, Angelo Montanari, Jesper Mosegaard, Martin Možina, 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, Denis Obrul, Barbara Oliboni, Matjaž Pančur, Wei Pang, Gregor Papa, Marcin Paprzycki, Marek Paralič, Byung-Kwon Park, Torben Bach Pedersen, Gert Schmeltz Pedersen, Zhiyong Peng, Ruggero G. Pensa, Dana Petcu, Marko Petkovšek, Rok Piltaver, Vid Podpečan, Macario Polo, Victor Pomponiu, Elvira Popescu, Božidar Potočnik, S. R. M. Prasanna, Kresimir Pripuzic, Gabriele Puppis, HaiFeng Qian, Lin Qiao, Jean-Jacques Quisquater, Vladislav Rajkovič, Dejan Rakovic, Jean Ramaekers, Jan Ramon, Robert Ravnik, Wilfried Reimche, Blagoj Ristevski, Juan Antonio Rodriguez-Aguilar, Pankaj Rohatgi, Wilhelm Rossak, Eng. Sattar Sadkhan, 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, Zhongzhi Shi, LIAN Shiguo, Heung-Yeung Shum, Tian Song, Andrea Soppera, Alessandro Sorniotti, Liana Stanescu, Martin Steinebach, Damjan Strnad, Xinghua Sun, Marko Robnik Šikonja, Jurij Šilc, Igor Škrjanc, Hotaka Takizawa, Carolyn Talcott, Camillo J. Taylor, Drago Torkar, Christos Tranoris, Denis Trček, Katarina Trojacanec, Mike Tschierschke, Filip De Turck, Aleš Ude, Wim Vanhoof, Alessia Visconti, Vuk Vojisavljevic, Petar Vračar, Valentino Vranić, Chih-Hung Wang, Huaqing Wang, Hao Wang, Hui Wang, YunHong Wang, Anita Wasilewska, Sigrid Wenzel, Woldemar Wolynski, Jennifer Wong, 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, Borut Zalik, Aleš Zamuda, Mansour Zand, Zheng Zhao, Dong Zheng, Jinhua Zheng, Albrecht Zimmermann, Blaž 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 2011 (Volume 35) 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 35 Number 4 December 2011 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **35** (2011) Number 4, pp. 399–533