Journal of Emerging Technologies in Web Intelligence

ISSN 1798-0461

Volume 2, Number 3, August 2010

Contents

Special Issue: Selected Best Papers of the International Conference on Information and Communication Systems (ICICS 2009)

Guest Editors: Mohammad Al-Rousan, Wei Li, and Ahmed Al-Dubai

Introducing the New Associate Editor-in-Chief

I am happy to welcome the new Associate Editor-in-Chief of the JETWI, Dr. Jiehan Zhou, to the editorial board.

Sabah Mohammed
Editor-in-Chief
August 2010

Jiehan Zhou is a joint research scientist and project manager at Intelligent System Group and MediaTeam, Computer Science and Engineering Lab, University of Oulu. He obtained his PhD in manufacturing and automation from the Huazhong University of Science and Technology, Wuhan, China in 2001. He did 2-year postdoctoral research in CIMS, Department of Automation, Tsinghua University, Beijing, China. He worked in VTT/Oulu Finland (Technical Research Center of Finland) and INRIA/Sophia Antipolis France (French national institute for research in computer science and control) for an 18-month ERCIM (European Research Consortium for Informatics and Mathematics) fellowship. He is a member of the ACM, IEEE, HUMAINE Association at Emotion-Research.Net, and Pattern Recognition Society of Finland. His current research interests include Pervasive Computing, Pervasive Service Computing, Multimedia Telecommunication, Community Coordinated Multimedia, Service-Oriented Computing, Semantic Web, and Emotion-Oriented Computing. His research has been published in Journals of Personal and Ubiquitous Computing, IEEE Multimedia, etc.

doi:10.4304/jetwi.2.3.157-157

Special Issue on Selected Best Papers of the International Conference on Information and Communication Systems (ICICS 09)

Guest Editorial

Mohammad Al-Rousan
College of Computer and Information Technology, Jordan University of Science & Technology, Jordan
Email: alrousan@just.edu.jo

Wei Li
Computer Science Department, University of Alabama in Huntsville, USA
Email: wli@cs.uah.edu

Ahmed Al-Dubai
School of Computing, Edinburgh Napier University, 10 Colinton Road, Edinburgh, EH10 5DT, U.K
Email: a.al-dubai@napier.ac.uk

There is too much content on the Internet, so people who consult it can lose sight of their own objective without strong intention. A similar thing can also occur on a Web site. To solve this issue, some Web sites have employed recommender systems such as collaborative filtering to help a user browsing the Web site.

Thus, using a recommender system for computer-based content selection and presentation is a way to strike a balance between content generated by other people and a user. Usually, the target of a recommender system is a single Web site.

The International Conference on Information and Communications Systems (ICICS2009) is a forum for Industry Professionals and Academics from Companies, Governmental Agencies, and Universities around the world to present their latest research results, ideas, developments and applications in all areas of Computer and Information Sciences. The topics that have been covered in the ICICS2009 include, but are not limited to: Artificial Intelligence, Mobile Computing, Networking, Information Security and Cryptography, Intrusion Detection and Computer Forensics, Web Content Mining, Bioinformatics and IT Applications, Database Technology, Systems Integration, Information Systems Analysis and Specification, Telecommunications, and Human-computer Interaction. We selected 12 high quality papers (out of 62 papers, which were presented at the ICICS2009) and invited the authors of the selected papers to extend them and submit them for a complete new peer-review for consideration in this Special Issue (SI). The final decision for the inclusion in the SI has been strictly based on the outcome of the review process. The main objective of the SI is to make available the latest results in the field to the research community and report state-of-the-art and in-progress research on all aspects of information and communication systems. The selected papers span a broad range on the information retrieval, E-business and Internet. The contributions of these papers are outlined below.

Muath Alzghool and Diana Inkpen have studied the data fusion method that should be able to combine the results that have high retrieval effectiveness with the results that have low retrieval effectiveness. They have conducted a number of experiments. Their results show that the new proposed technique is significantly better than CombSUM or WCombSUM in combing results with high quality variation. On the other hand, Nadia Bouassida and Hanêne Ben-Abdallah have focused on the Pattern and spoiled pattern detection. They have presented a method that identifies design patterns and spoiled patterns through an XML document retrieval approach. This latter provides for the possibility of tolerating structural variations between the design and the searched pattern. Further, their proposed pattern identification method can be parameterized in order to delimit the degree of acceptable variations. Moreover, Adnan A. Tala'a aims at combining the power of wireless communications with the strength of database information keeping through Internet as a media to make it possible to implement an automated record keeping system. The proposed system is based on wireless communications for schools.

Most previous works were focused on looking at mobile phone as terminal device. However, Meisam Hejazinia and Mohammadreza Razzazi have a different view. They look at mobile phone as server; they defined the concept of society network, which could be leveraged to enhance mobile phone mobile agent capability. Their contribution was redefining mechanisms of mobile agents for specific characteristics of mobile phone. Reusing and remixing contents are keys to expressing activities. Within this context, Kosuke Numa, Katsuaki Tanaka, Mina Akaishi and Koichi Hori present a new framework for content circulation. The author concluded that the framework could help content recomposition and this framework is applicable to various manners of expressions like Web content creation. On the other hand, Katsuaki Tanaka, Koichi Hori and Masato Yamamoto address another important issue. They propose new recommender system based on extending contexts of content and personal history. This systems captures personal context from a history of personal online and offline activities, treats information on Web sites as a large set of context, and discovers and extends the overlaps of personal activities and Web sites, then recommends information located in the Web sites.

doi:10.4304/jetwi.2.3.158-159

E-buisnees has been a major avenue for many researchers in IT. With the global emergence of e-government and its potential benefits to citizens, there has been a growing need adopting the e-government services. Rand A. Obeidat and Emad A. Abu-Shanab address this issue with a particular attention paid to the Jordanian experience in e-business.

As for the software and requirements engineering, Andrea De Lucia and Abdallah Qusef address discusses problems concerned with the conduction of requirements engineering activities in agile software development processes and suggests some improvements to solve some challenges caused by agile requirements engineering practices in large projects. William J. Tastle, Amjad Abdullat and Mark J. Wierman introduce a new approach in requirements elicitation analysis. In this study, it is commonly acknowledged that difficulties in understanding challenging and complex problems, sometimes even perceived as being rather intractable, usually presage the more interesting efforts of building information systems in organizations.

Home networking is a growing area of research. Most of the home networks are currently used to connect PCs for tasks such as printing and shared Internet connectivity. However, Sandeep Kumar, Saiful Islam, Archana Gupta and Himanshu Bhardwaj tend to take this field further as they have designed a new system for controlling electrical devices but the design can be extended to control mechanical devices. They have simulated the system for controlling the LEDs and this is working properly. This SI also considers the needs of different communities, including the blind people. Ameer H. Morad designs a small device, very easy to use called, GPS Talking Blind People, help blind people to navigate around camps, cities and get voice messages notifying them concerning their locations.

Finally, as guest –co-editors of this SI, we would like to express our deepest thanks to the Editor-in-Chief, Professor Sabah Mohammed for hosting this Issue in the JETWI and for his continued support and helpful guidance throughout all the stages of preparing this SI. Our sincere thanks also go to the Editorial-office staff of the journal for their excellent job during the course of preparing this special issue. We also thank the authors for their contributions, including those whose papers were not included. We thank and greatly appreciate the thoughtful work of many reviewers who provided invaluable evaluations and recommendations.

Mohammad Al-Rousan is currently an associate professor at the Department of Network Engineering and Security, Jordan University of Science and Technology (JUST). He was educated in KSA, Jordan and USA. He received his BSc in Computer Engineering from King Saud University, Saudi Arabia, in 1986. He received his M.S. in Electrical and Computer Engineering from University of Missouri-Columbia, MI, USA in 1992. In 1996 , he was awarded the PhD in Electrical and Computer Engineering from Brigham Young University, UT, USA. He was then an assistant professor at JUST, Jordan. In 2002, he joint the Computer Engineering Department at American University of Sharjah, UAE. Since 2008, he has been the Dean of College of Computer and Information Technology at JUST. He is the Director of the Artificial Inelegant and Robotics Laboratory, and a co-founder for the Nano-bio laboratory, at JUST. His search interests include wireless networking, System protocols, intelligent systems, computer applications, and Nanotechnology, Internet computing. Dr. Al-Rousan served on organising and program committees for many prestigious international conferences. He is the recipient of several prestigious awards and recognitions. He co-chaired international conferences on Information and Communication Systems (ICICS09).

Wei Li received the Doctor of Philosophy degree from Virginia Polytechnic Institute and State University and has been working on software measurement, design, and formal method. His recent work focused on validating software metrics using open-source data and formalizing metrics.

Ahmed Y. Al-Dubai is currently a lecturer in the School of Computing at Edinburgh Napier University. He was educated in Yemen, Jordan and UK. He received his BSc and MSc in Computer Science from Mutah University and Al al-Bayt University, Jordan in 1996, 1999, respectively. In 2004, he was awarded the PhD in computing from the Department of Computing Science, University of Glasgow (Outstanding PhD studentship award). He was then a full time lecturer at Thames Valley University-London, 2004-2005 before joining Edinburgh Napier University. His research interests include communication algorithms, parallel & distributed computing and next generation wired and wireless networks. His research is funded by different sources, including EU, Universities UK and the Royal Society. He served on organising and program committees for many prestigious IEEE and ACM international conferences. He is the recipient of several prestigious awards and recognitions. He has been the Guest Co-Editor of 10 international journals. He chaired and co-chaired 12 IEEE/ACM international conferences/workshops. Dr Al-Dubai is Senior Member of the IEEE, and Member of the IEEE Computer Society and the ACM.

A Novel Class-Based Data Fusion Technique for Information Retrieval

Muath Alzghool
Al-Balqa' Applied University, Alsalt, Jordan
Email: alzghool@site.uottawa.ca

Diana Inkpen
University of Ottawa, Ottawa, Canada
Email: diana@site.uottawa.ca

***Abstract*— Data fusion in information retrieval combines the results from multiple retrieval models or document representations. The achievement of data fusion technique is dependent on the quality of the inputs; classical data fusion techniques fail to improve the retrieval if the quality of the retrieval results varies from low to high quality. In order to tackle this problem, in this paper we address the issue of high variation among the retrieval strategies or document representations which affect the combination of their outputs. Our investigation on the MALACH speech collection – in which different segment representations are available – shows that neither the classical data fusion (CombSUM) nor the weighted version (WCombSum) improve the retrieval. We propose a novel class-based data fusion technique to deal with this issue. The segments retrieved by models based on different document representations are classified according to the quality of the segment into three classes: high, intermediate, and low quality class; then the similarity scores of each segment are fused using the classical CombSUM. Our experimental results show that the new technique is significantly better than CombSUM or WCombSUM in combing results with high quality variation.**

***Index Terms*—Information storage and retrieval, searching spontaneous speech transcriptions, data fusion.**

I. Introduction

Conversational speech such as recordings of interviews or teleconferences is difficult to search through. The transcripts produced with Automatic Speech Recognition (ASR) systems tend to contain many recognition errors, leading to low Information Retrieval (IR) performance [1] unlike the retrieval from broadcast speech, where the lower word error rate did not harm the retrieval [2].

A large number of IR systems and retrieval strategies have been proposed and implemented in the last 30 years. All these approaches differ one from another in several issues such as the preprocessing process, the data representation, the weighting scheme and the similarity measure. There is a tremendous need to benefit from the strategies. One way to benefit from them is to combine their results by a data fusion technique.

Users tend to express their queries in various ways: sometimes they use more general terms, sometimes more specific terms, or a combination of both. IR systems need to be able to accommodate this variety of user needs; there is also variation among the collections (if it is a special collection like the one we use or a general collection like the news collection). Some retrieval models or weighting schemes perform better when the queries are general, others perform better when the queries are more specific, and others when a combination is available. In this paper we are looking for a system that will perform well in all these cases such as data fusion, where the system attempts to combine the results from multiple retrieval models.

Although the application of data fusion in IR has yielded good results in the majority of the cases, it has been noticed that its achievement is dependent on the quality of the input result lists [3-6]. Classical data fusion technique fail to improve the retrieval when the quality of the inputs varies from low to high quality; the presence of some poor-quality inputs (containing very few relevant documents in the top part of the list) causes a significant drop in the fusion performance

Lee [7] analyzed the overlap values of result sets from six different participants in TREC-3; he found that low overlap in non-relevant and high overlap in relevant documents is critical to improving effectiveness. We believe that the data fusion method should be able to combine the results that have high retrieval effectiveness with the results that have low retrieval effectiveness. Therefore, we propose a novel data fusion technique to fuse the results of different document representations, where the quality of the retrieval results varies from low quality to high quality.

We applied our data fusion techniques to Multilingual Access to Large spoken ArCHives collection (MALACH) [8] that used in the Cross-Language Speech Retrieval (CLSR) task at Cross-Language Evaluation Forum (CLEF) 2007. See Section 5 for a brief description of the collection.

The remainder of this paper is organized as follows. Section 2 is pointing to the most important work in model fusion. Section 3 describes the two IR systems that we used to provide candidate weighting schemes (retrieval

doi:10.4304/jetwi.2.3.160-166

strategies) for our model fusion technique. Section 4 describes the data fusion technique proposed in this paper. Section 5 outlines the CLEF CL-SR test collection. Section 6 presents our experimental results. Finally, Section 9 presents conclusions and future work.

II. Related Work

Model fusion combines the results from multiple retrieval models. Since different models may have different strengths, combining information extracted by multiple retrieval models can bring performance improvements. Fusion of retrieval results from different models for improving retrieval performance has been reported in works like [7, 9-12]. Retrieval results from different systems [10] or retrieval results using different document representations [11] were fused together for performance improvement.

Another way to differentiate data fusion methods is the way they compute the final score of documents. Some methods directly use the similarity values of the documents across the lists [9, 13], other consider their rank [3, 7], and others their probability of occurring in a predefined segment of the lists [14, 15]. In addition, some methods are based on the Social Choice Theory [16], he used pair wise contests of documents to determine their final score.

In general, a linear combination (CombSUM) of the retrieval results was found to be the simplest and most effective way for fusing multiple information sources in order to improve retrieval performance.

The application of data fusion in information retrieval has shown relevant results in the majority of the cases; nevertheless, it has been noticed that it is sensitive to several factors, such as the quality of the input lists[3-6].

III. System Description

The weighting schemes for our fusion system were provided by two IR systems: SMART [17] and Terrier [18] .

SMART was originally developed at Cornell University in the 1960s. SMART is based on the vector space model of IR. We use the standard notation from SMART: the weighting scheme for the documents, followed by dot, followed by the weighting scheme for the query, where the schemes are abbreviated by the type of normalization (n means no normalization, c cosine, t idf, l log, etc.). We used the nnc.ntc, ntc.ntc, lnc.ntc, ntn.ntn, lnn.ntn, ltn.ntn, lsn.ntn weighting schemes[17]. We chose these schemes because they performed well on the training data in our last experiments[19].

Terrier was originally developed at the University of Glasgow. It is based on Divergence from Randomness models (DFR) where IR is seen as a probabilistic process.[18] We experimented with all the weighting schemes implemented in Terrier (BB2, BM25, DFR_BM25, DFRee, DLH13, DLH, IFB2, In_expB2, In_expC2, InL2, PL2, LemurTF_IDF, and TF_IDF).

IV. Model Fusion

A. CombSUM

Fox and Shaw [10] proposed several fusion methods for combining multiple scores. The most simple and effective one was called CombSUM, which sums up all the scores of a document, as in formula 1:

$$CombSUM = \sum_{i \in IR\ schemes} score_i \quad (1)$$

where $score_i$ is the similarity score of the document to the query for the weighting scheme i which retrieved this document.

Since there are different weighting schemes from different systems, these schemes will generate different ranges of similarity scores, so it is necessary to normalize the similarity scores of the documents. Lee [7] proposed a normalization method by utilizing the maximum and minimum scores for each weighting scheme as defined by formula 2.

$$NormalizedScore = \frac{score - MinScore}{MaxScore - MinScore} \quad (2)$$

B. Weighted CombSUM

When training data is available, many researchers experimented with updated versions of CombSUM, where a weight is assigned to each retrieval strategy according to performance on the training data. Then, they applied the determined fusion formula to the test data. This fusion method is called WCombSUM, represented by formula 3.

$$WCombSUM = \sum_{i \in IR\ schemes} W_{ik} * NormalizedScore_i \quad (3)$$

where W_{ik} is a pre-calculated weight associated with each retrieval strategy, and the $NormalizedScore_i$ is calculated by formula 2 as described before.

In the literature, there are different ways to assign a weight (W_{ik}) for each retrieval strategy:

- Manually-weighted scheme [12, 20], where the researchers try different weight values for each retrieval strategy and select the best combination. We believe this technique is an unsystematic way to derive the weights.
- MAP-based weighted scheme [16, 21, 22], where the Mean Average Precision (MAP) score for each retrieval strategy on training data is considered as a weight for that strategy. This technique is simple and proves to be effective for some cases when there is no performance variation between the retrieval strategies on different data.
- MAP-Recall weighted scheme [19], where the MAP and the recall score are combined to derive the weight for each retrieval strategy so that the best weighting scheme contribute the most, and the others only support it.

In our experiments, we will use CombSUM and WCombSUM as baseline method, to compare it to our new technique. As a base case, we will consider the MAP

TABLE I. The retrieval results on MALACH collection using the weighting scheme DLH13 from the Terrier IR system on training data.

	Training	Test
Auto	0.1041	0.0735
Manual	0.3321	0.2560
Auto+Manual	0.2837	0.1606
CombSUM	0.2844	0.1953
WCombSUM	0.3272	0.2393

scores as the weights in the training phase for WCombSUM.

C. Class-Based Fusion

In this section we will discuss the case when we have different retrieval strategies and there are large differences in the effectiveness (significant difference), or we have one retrieval strategy and different representations for the documents, so that when we apply the retrieval strategy to the different representations, there are a significant differences among the different representations. Because of these differences, the basic fusion methods fail to improve the retrieval due to the noises from bad strategies or representations.

For example, the MALACH test collection contains 8104 segments from 272 interviews with Holocaust survivors and each segment contains different versions of automatic transcriptions, two sets of automatically-generated thesaurus terms, manually generated summaries, and manually-generated thesaurus terms. Each of them can be viewed as a representation for the segment. The first representation is when we index the automatically-generated data (Auto). The second one, is when we index the manually-generated data (Manual), and the third one is when we index the automatic and the manually-generated data together (Auto+Manual). If we apply any retrieval strategy to each representation, there are big differences among the representations, for example as shown in Table I, the MAP score for Auto, Manual, and Auto+Manual are 0.1041, 0.3321, and 0.2837, respectively. When the basic fusion methods like CombSUM or WCombSUM where the weights are the MAP scores on training data are applied, the MAP scores for the fusion methods are 0.2844 and 0.3272, respectively.

We are looking for a fusion technique that can handle the variations among the retrieval strategies or the document representations.

To achieve this goal, we will divide the retrieved documents from all the retrieval strategies or the document representations into three classes: the first one is expected to have the best precision values, the second one has intermediate precision values, and the last one has low precision; we will call these classes high, intermediate, and low class, respectively. Since the Manual experiment has the best MAP, we will assume the high class will have the top n documents from the Manual experiment. The intermediate class will have the next m documents from Manual and the top m documents from Auto+Manual. Finally, the low class will have the remaining documents from Manual, Auto+Manual, and all the documents from Auto experiment. Note that the intersection between the three classes has to be mutually exclusive, i.e., if a document d appears in the top n documents from Manual and in the top m documents from Auto+Manual, d will be included in the high class, not in the intermediate class.

TABLE II. 11-level interpolated recall-precision values for the three experiments: Manual, Auto+Manual, and Auto. We show how to derive n and m, as explained in the text.

	Manual	Auto+Manual	Auto
Recall	**Precision**	**Precision**	**Precision**
0%	0.722	**0.697**	**0.424**
10%	0.577	0.504	0.247
20%	0.507	0.439	0.189
30%	0.435	0.353	0.146
40%	0.405	0.315	0.115
50%	0.353	0.282	0.091
60%	0.301	0.256	0.061
70%	0.242	0.200	0.041
80%	0.154	0.152	0.017
90%	0.090	0.088	0.023
100%	0.032	0.025	0.001

The next step shows how to estimate the values for n and m (n is the separation cut-off point between the high and the intermediate class, and m is the separation cut-off point between the intermediate and the low class). We use the evaluation of the three experiments on training data; for this stage we choose interpolated precision values at 11 recall points. To estimate n, for separating the high class from the intermediate class, we choose the maximum precision on Auto+Manual experiment, then find the level of recall that represents this value in the manual experiment, which is actually the same as looking at the length of the document list at the cut-off point; finally, we multiply this recall level by 1000 to calculate n (since the number of retrieved documents for each retrieval strategy is 1000, we take a portion of this number, which is proportional to the recall level). We use the same procedure for m; we chose the maximum precision on the Auto experiment, then find the level of recall on Auto+Manual and multiply it by 1000.

For example, Table II represents the precision at the 11-levels of recall for the three experiments mentioned in Table I. To estimate n, first we have to find the best precision in Auto+Manual, which is 0.697; then we have to find the level of recall that represents this value in the Manual experiment (0.1), and finally multiply this recall level by 1000; therefore, the estimated value for n is 100. We do the same thing for m; the maximum precision value in Auto is 0.424; the level of recall that represents this value according to the evaluation of the Auto+Manual experiment is 0.3; therefore, m is equal to 300. The high class will contain the top 100 documents

TABLE III. Some statistics about the number of terms and the number of tokens for the three experiments.

	Number of terms	Number of tokens	Average term frequency
Auto	13,605	1,711,684	125.8
Manual	7,131	278,717	39
Auto + Manual	15,884	1,990,401	125.3

from Manual; the intermediate class will contain the next 300 documents from Manual and the top 300 from Auto+Manual; finally, the low class will contain the remaining documents from Manual and Auto+Manual (600 and 700, respectively) and all the documents from Auto that were not included neither in the high class nor in the intermediate class. The three classes are mutually exclusive. In the above example, if one of the top 100 documents from Manual happens to be in the set of top 300 from Auto+Manual, then this document will be in the high class, not in the intermediate one.

The final step is to fuse the similarity scores of each document and to sort them in decreasing order in each class separately, then arrange the documents for the high class first, then the intermediate class, and finally the low class. To fuse the similarity scores, we could use CombSUM or WCombSUM. We have to normalize the similarity scores according to the maximum and minimum in each class. In our experiments, for any run that uses the class-based fusion, we will use the prefix "WC" before the method name, i.e., WCCombSUM.

V. THE CLEF CL-SR TEST COLLECTION

This section describes the data that we used. The MALACH collection contains 8104 "documents" which are manually-determined topically-coherent segments taken from 272 interviews with Holocaust survivors, witnesses and rescuers, totaling 589 hours of speech. Two ASR transcripts are available for this data, in this work we use the ASRTEXT2006B field provided by IBM research with a word error rate of 25%. Additional meta-data fields for each document include: two sets of 20 automatically assigned keywords determined using two different k-nearest neighbor classifiers (AK1 and AK2), a set of a varying number of manually-assigned keywords (MK), and a manual 3-sentence summary written by an expert in the field. A set of 63 training topics and 33 test topics were generated for this task. The topics provided with the collection were created in English from actual user requests. Topics were structured using the standard Text Retrieval conference (TREC) format of Title, Description and Narrative fields. For cross-language experiments, the topics were translated into Czech, German, French, and Spanish by native speakers. Relevance judgments were generated using search-guided procedure and standard pooling methods. See [8] for full details of the collection design.

TABLE IV. The average idf values, and number of missing search terms from title and description fields, for training (681 terms) and test (356 terms) topics

	IDF Training	IDF Test	Missing Training	Missing Test
Auto	1.22	1.08	28	8
Manual	1.75	1.74	27	9
Auto+Manual	1.22	1.05	10	5

VI. EXPERIMENTAL RESULTS

The candidate retrieval strategies (weighting schemes) for our fusion system were provided by two IR systems: SMART [17, 23] and Terrier [18, 24].

We conducted three types of experiments, based on the fields which were indexed. In the first one, the automatic transcripts (ASRTEXT2006B), and two automatic keywords (AK1 and AK2) were used for indexing the documents; we call this experiment Auto. In the second experiment, we indexed the manual keywords and the manual summaries for each document; we named this experiment Manual. In the last experiment we indexed the automatic transcripts, the two automatic keywords fields, the manual summaries, and the manual keywords, we call this experiment Auto+Manual. The title and description fields from each topic are used as query. Table III shows some statistics about each experiment.

One interesting observation is that the number of terms (distinct words) in the manual fields is about half of the number of terms in the automatic fields. The number of tokens (total number of words) in the manual fields is about 16% of the number of tokens in the automatic fields. The average term frequencies are 39, 125, and 125 for Manual, Auto, and Auto+Manual, respectively. This ratio is very high: about four times more in the Auto fields. We also note that combining Auto and Manual brings about 14% of the terms to the Auto+Manual list of terms, which means that there is more information in the combined fields.

A. Manual Summaries and Keywords versus Automatic Transcripts

Experiments on manual keywords and manual summaries (Manual) available in the test collection showed high improvements over automatic transcripts and automatic keywords (Auto). The MAP score jumped from 0.0779 to 0.2727 on the test data. Also, if we indexed the Manual fields and the Automatic fields together (Auto+Manual), the MAP score jumped to 0.161, but it is far from the results on the Manual. This was also the case in the systems that participated in CLEF-CLSR. We are looking for a justification of why the difference is so big between the results of the Auto experiment and the Manual experiment, and why when we merge the Auto with Manual we do not reach the performance of the Manual fields. Since there are no manual transcripts available for the segments, we cannot

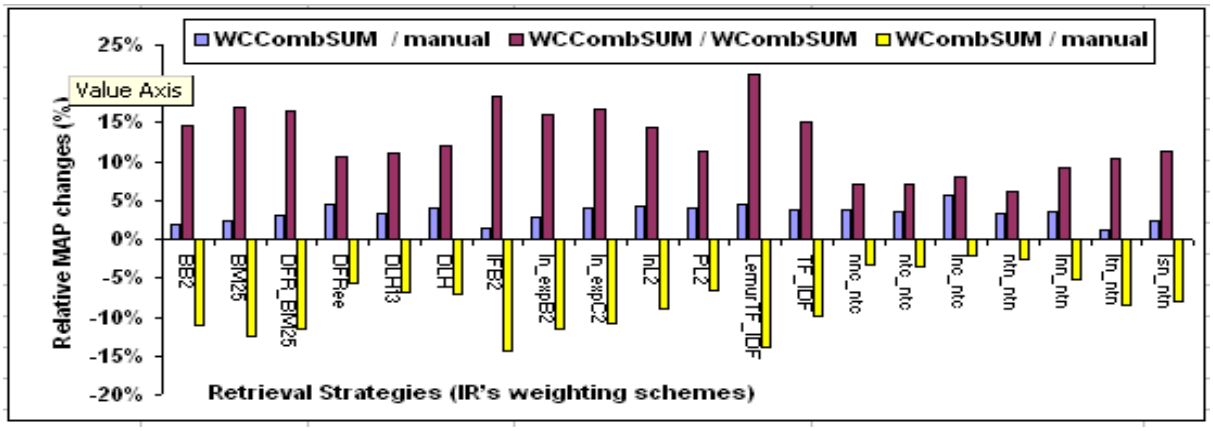


Figure 1. Relative MAP changes between WCCombSUM and the best pre-fusion run (manual representation), WCCombSUM and WCombSUM, and WCombSUM and the best pre-fusion run (manual representation). The fusion applied to 20 retrieval strategies from SMART and Terrier to fuse 3 segment representations: auto, manual, auto-manual.

know how the word error rate (WER) affects the retrieval.

We think that there are several factors that may affect the retrieval. The manual summaries are very concise representations of the segments; they tend to use different language than the segments. The automatic transcripts or the manual summaries cover the search terms from the training and test topics in different ways. Table IV counts the missing terms for each experiment in the training and test topics. We noticed that the number of the missing terms is approximately the same for Manual and Auto, and for Auto+Manual is approximately half the missing number of terms from Manual or Auto. Therefore, we cannot consider the missing terms as the factor which affects the large difference in MAP score between Auto and Manual. Another factor could be related to the ability of the search terms to discriminate among the documents. The classic discrimination measure is the idf value for the search terms. Table IV shows the average idf for the training and test topics. We notice that the average idf for Auto and Auto+Manual is less than for Manual. Therefore, the topics ability to discriminate the documents in the Manual experiments is higher than for Auto or Auto+Manual. A last factor that we mention is

TABLE V. RESULTS (MAP SCORES, AND NUMBER OF RELEVANT DOCUMENTS RETRIEVED) FOR 20 WEIGHTING SCHEMES FROM SMART AND TERRIER, AND THE RESULTS OF THE FUSION METHODS (WCOMBSUM AND WCCOMBSUM), ON THE TEST DATA.

Weighting scheme	Auto		Manual		Auto+Manual		WCombSUM		WCCombSUM	
	MAP	Rel-Ret	MAP	Rel.-Ret	MAP	Rel-Ret	MAP	Rel-Ret	MAP	Rel-Ret
BB2	0.0441	972	0.2699	1826	0.0970	1133	0.2402	1869	0.2752	1888
BM25	0.0567	1120	0.2490	1824	0.1404	1381	0.2182	1874	0.2551	1882
DFR_BM25	0.0580	1122	0.2558	1818	0.1408	1407	0.2261	1878	0.2635	1889
DFRee	0.0695	1298	0.2527	1822	0.1586	1697	0.2387	1897	0.2640	1900
DLH13	0.0735	1335	0.2560	1825	0.1606	1720	0.2384	1898	0.2647	1890
DLH	0.0719	1325	0.2460	1812	0.1606	1707	0.2287	1875	0.2560	1878
IFB2	0.0605	1080	0.2705	1824	0.135	1335	0.2320	1899	0.2747	1900
In_expB2	0.0657	1259	0.2727	1826	0.1537	1581	0.2416	1918	0.2805	1925
In_expC2	0.0700	1288	0.2704	1826	0.1551	1609	0.2409	1915	0.2812	1911
InL2	0.0629	1259	0.2575	1826	0.1521	1570	0.2346	1898	0.2685	1898
PL2	0.0730	1295	0.2510	1803	0.1575	1658	0.2347	1876	0.2613	1873
Lemur TF_IDF	0.0517	1146	0.2269	1814	0.1319	1425	0.1956	1867	0.2371	1874
TF_IDF	0.0651	1302	0.2525	1818	0.1452	1627	0.2277	1890	0.2620	1884
nnc_ntc	0.0779	1270	0.2190	1760	0.161	1698	0.2119	1837	0.2271	1845
ntc_ntc	0.0630	1235	0.2154	1776	0.1525	1623	0.2080	1845	0.2230	1873
lnc_ntc	0.0722	1269	0.2270	1784	0.1585	1667	0.2222	1865	0.2398	1879
ntn_ntn	0.0649	1250	0.2140	1792	0.1464	1643	0.2084	1857	0.2212	1867
lnn_ntn	0.0658	1284	0.2346	1789	0.1527	1684	0.2226	1880	0.2429	1897
ltn_ntn	0.0512	1166	0.2167	1785	0.1297	1511	0.1984	1844	0.2191	1880
lsn_ntn	0.0426	1028	0.1856	1787	0.1140	1376	0.1706	1795	0.1899	1832

the average term frequency, which is much larger in Auto and Auto+Manual (125) than in Manual (39), as previously shown in Table III.

Since the manual summaries and the automatic transcripts complement each other, each one brings new terms to the document structure as shown also in Table I. Mixing the two fields is supposed to improve the retrieval, in theory. From the results, it is clear that simple merging technique - during the indexing - does not help. A better way to combine or fuse the two fields during the indexing was addressed by [25].

In the next section, we will present the experimental results for the class-based fusion, which improve the retrieval, and benefits from the different information included in different segment representations.

B. Class-Based Fusion Experiments

We have applied our class-based fusion proposed in section IV.C to fuse the results from the three segments representations Auto, Manual, and Auto+Manual for each retrieval strategy (weighting scheme) from SMART or Terrier.

The baselines are the best retrieval run (the results from the Manual representation run) and the classical retrieval *WCombSUM*, where the weights in *WCombSUM* are represented by the MAP of each run on training data.

As shown in Table V and Figure 1 (see the previous page), the classical fusion technique (*WCombSUM*) does not improve the results comparing to the best run involved in the fusion process for the 20 retrieval strategies; but our method was better than the best run involved in the fusion for all the 20 retrieval strategy. For 15 out of 20 runs, the improvement was significant, based on a one-tailed Wilcoxon signed rank test with ($p < 0.05$). Also, our method was significantly better than the classical *WCombSUM* for all the 20 retrieval strategies. Based on number of relevant document retrieved, the *WCCombSUM* method is significantly better than the best run, *CombSUM*, and *WCombSUM*.

We conclude from our experiments that the information in meta-data like manual summaries and keywords complement the information contained automatic transcriptions and automatic keywords, and we could benefit from this feature to post-fuse the results of each representation and improve the retrieval.

VII. CONCLUSIONS

We have addressed the case when there are large differences in the effectiveness between the retrieval strategies or the document representations involved in the fusion, where classical techniques failed badly to improve the results. The solution was a class based method.

Finally, we have showed that meta-data complemented the error-full transcription, and we could benefit from the class-based fusion to improve the retrieval.

REFERENCES

[1] P. Pecina, P. Hoffmannov´a, G. J. F. Jones, Y. Zhang, and D. W. Oard, "Overview of the CLEF-2007 Cross Language Speech Retrieval Track," *Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, vol. 5152, pp. 674-686, 2008.

[2] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," in *Proceedings of the RIAO: Content-Based Multimedia Information Access* PARIS, 2000.

[3] E. A. Fox and J. A. Shaw, "Combination of Multiple Searches," in *Proceedings of the Second Text REtrieval Conference (TREC 2)* Gaithersburg, Maryland, USA: National Institute of Standards and Technology (NIST), 1993.

[4] C. C. Vogt and G. W. Cottrell, "Predicting the performance of linearly combined IR systems," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* Melbourne, Australia: ACM, 1998.

[5] K. Ng and B. K. Paul, "Predicting the effectiveness of Naive data fusion on the basis of system characteristics." vol. 51: John Wiley & Sons, Inc., 2000, pp. 1177-1189.

[6] S. Wu and S. McClean, "Performance prediction of data fusion for information retrieval," *Information Processing & Management*, vol. 42, pp. 899-915 2006.

[7] J. H. Lee, "Combining multiple evidence from different properties of weighting schemes," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* Seattle, Washington, United States: ACM, 1995.

[8] D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, B. Ramabhadran, M. Franz, and S. Gustman, "Building an information retrieval test collection for spontaneous conversational speech," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* Sheffield, United Kingdom: ACM, 2004.

[9] B. T. Bartell, G. W. Cottrell, and R. K. Belew, "Automatic Combination of Multiple Ranked Retrieval Systems," in *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* Dublin,Ireland: ACM/Springer, 1994.

[10] J. A. Shaw and E. A. Fox, "Combination of Multiple Searches," *Proceedings of the third Text REtrieval Conference (TREC-3)*, pp. 105-108, 1994.

[11] G. J. F. Jones, J. T. Foote, K. Spark Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* Zurich, Switzerland: ACM, 1996.

[12] K. Ng, "Information fusion for spoken document retrieval," in *Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference - Volume 04* Istanbul, Turkey: IEEE Computer Society, 2000.

[13] C. C. Vogt, G. W. Cottrell, R. K. Belew, and B. T. Bartell, "Using relevance to train a linear mixture of experts," in *The Fifth Text REtrieval Conference (TREC-6)* Gaithersburg, Maryland, 1997.

[14] G. Lebanon and J. Lafferty, "Cranking: Combining Rankings Using Conditional Probability Models On Permutations," in *Proceedings of the 19th International Conference on Machine Learning* Sydney, Australia, 2002.

[15] D. Lillis, F. Toolan, R. Collier, and J. Dunnion, "ProbFuse: a probabilistic approach to data fusion," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* Seattle, Washington, USA: ACM, 2006.

[16] M. Montague, "Metasearch: Data Fusion for Document Retrieval," in *Computer science*. vol. Ph.D. Hanover: Dartmouth College, 2002.

[17] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, pp. 513-523, 1988.

[18] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson, "Terrier Information Retrieval Platform " *Lecture notes in computer science: Advances in Information Retrieval*, vol. 3408/2005, pp. 517-519, 2005.

[19] M. Alzghool and D. Inkpen, "Model Fusion Experiments for the CLSR Task at CLEF 2007," *Lecture Notes in Computer Science: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, vol. 5152, pp. 695-702, 2008.

[20] D. He and J.-W. Ahn, "Pitt at CLEF05: Data Fusion for Spoken Document Retrieval," *Accessing Multilingual Information Repositories*, vol. 4022/2006, pp. 773–782, 2006.

[21] D. He and D. Wu, "Toward a Robust Data Fusion for Document Retrieval," in *Proceedings of the 2008 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'08)* Beijing, China, 2008.

[22] M. Alzghool and D. Inkpen, "Cluster-based Model Fusion for Spontaneous Speech Retrieval," in *Proceedings of the Workshop on Searching Spontaneous Conversational Speech, SIGIR 2008* Singapore, 2008.

[23] C. Buckley, G. Salton, and J. Allan, "Automatic Retrieval With Locality Information Using SMART," *Proceedings of theText REtrieval Conferenc (TREC-1)*, pp. 59-72, 1992.

[24] G. Amati and C. J. V. Rijsbergen, *Probabilistic models of information retrieval based on measuring the divergence from randomness* vol. 20. New York: ACM, 2002.

[25] G. J. F. Jones, K. Zhang, E. Newman, and A. M. Lam-Adesina, "Examining the Contributions of Automatic Speech Transcriptions and Metadata Sources for Searching Spontaneous Conversational Speech," in *SIGIR 2007 workshop: Search in Spontaneous Conversational Speech workshop* Amsterdam, 2007.

Pattern and Spoiled Pattern Detection through an Information Retrieval Approach

Nadia Bouassida
Institut Supérieur d'Informatique et de Multimédia
Université de Sfax, Tunisie
Email: Nadia.Bouassida @isimsf.rnu.tn

Hanêne Ben-Abdallah
Faculté des Sciences Economiques et de Gestion
Université de Sfax, Tunisie
Email:Hanene.BenAbdallah@fsegs.rnu.tn

Abstract—**Design patterns provide for a higher software quality and a reduced development cost. However, to reach these benefits, designers are expected to have a good understanding and experience with design patterns, which is not evident to acquire. Another way to benefit from design patterns is by assisting designers in their detection/identification within a given design in order to improve it.**

Since the *exact* structural instantiation of a pattern is less frequent to find within a design, the identification process should account for variations of the design with respect to the pattern. It assists the designer by showing the pattern elements in terms of the design which can be validated with respect to the classes, attributes, methods and relations of the pattern: the designer can add/remove some elements from the design in order to ensure a good instantiation of the identified pattern. However, not all structural variations of a pattern are tolerated; in fact, some variations may result in non-optimal instantiations of the pattern, *a.k.a.* spoiled patterns. In this case, the identification process can assist the designer by proposing corrections for an acceptable pattern instantiation.

Within this design context, we propose a method that identifies design patterns and spoiled patterns through an XML document retrieval approach. This latter provides for the possibility of tolerating structural variations between the design and the searched pattern. In addition, our pattern identification method can be parameterized in order to delimit the degree of acceptable variations.

Index Terms—**Design pattern identification, pattern instantiation, XML document retrieval.**

I. Introduction

Design patterns [9] are generic solutions for often occurring problems. Being proven solutions, proposed by experts, they promise several reuse benefits, such as high quality software, faster and lower cost software development. However, to attain these benefits, a designer must overcome the difficulties inherent to first understanding and then applying design patterns. In fact, even an experienced designer would spend a considerable time understanding, identifying and instantiating/reusing design patterns pertinent to his/her applications.

A straight forward way to benefit fully from design patterns is to assist an inexperienced designer to improve his/her design by identifying, in the design, instantiations of design patterns. On the other hand, since exact instantiations of a design pattern is less frequent (and is less problematic), an exact pattern identification method is, therefore, of limited use. Instead, pattern identification should look for structures that "resemble" a design pattern. By resembling, we mean structures that vary from a design pattern by adding/removing some elements (classes, attributes, methods, relations). The pattern identification method can, in this case, assist the designer in restructuring his/her design in conformance with the pattern found.

However, while tolerating pattern instantiations with variations, the identification method should watch out for non-tolerated variations, and in particular *spoiled patterns*. A spoiled pattern is a structure that allows to instantiate inadequate solutions for a given problem, where requirements are respected but architecture is improvable [17]. As an example of a spoiled pattern, an observer pattern where observers are subclasses of the subject class, a composite pattern where the leaf classes do not inherit from the composite class but are connected to the composite by a composition relation.

To offer assistance through pattern identification, several approaches propose to determine the potential similarities of the structure, the class names and/or method names between the design and a given pattern. These approaches differ mainly in the pattern concepts they consider (i.e., only the structure, the structure and the methods) and the degree of structural discordance they tolerate: exact match [4] or partial match [6], [13]. All methods that tolerate structural discordance between the design and a pattern treat all pattern elements equally. However, while some elements can be deleted in a design resembling a pattern, others representing the essence of the pattern (its core) should not; otherwise the pattern would be lost and/or spoiled.

In this paper, we present a new pattern identification technique that can: 1) be used to identify the structure, class names and participant roles of the pattern, 2)

doi:10.4304/jetwi.2.3.167-175

identify the spoiled patterns, and 3) take into account the degree of variability of a pattern. In addition, once a similarity is found, the identified design fragment is presented with the pattern roles and variability. This presentation assists the designer in better understanding the pattern through his/her application and in validating its instantiation. For this, we propose to use the P-UML design language [3], a UML profile for patterns.

More specifically, our identification technique reuses an XML document retrieval approach where the pattern is seen as the XML query and the design as the XML document where the query is searched. It relies on the context resemblance function [11] to compute the similarity potential of the design structure and the pattern. One advantage of this approach is that it is applicable to account for both the structure and methods in the pattern. A second advantage is that it accommodates design variability with respect to the pattern structure without losing the pattern essence or spoiling it.

The remainder of this paper is organized as follows. Section 2 overviews currently proposed approaches and tools for pattern and spoiled pattern identification. Section 3 first summarizes the basic concepts of XML document retrieval in general and the P-UML design language. Section 4 presents our approach for pattern identification and illustrates it with the composite pattern and a spoiled composite pattern. Section 5 summarizes the paper and outlines our future work.

II. Current pattern identification approaches

Several works have been interested in pattern identification but for different purposes. For reverse engineering purposes, several proposals addressed the problem of automating the identification of design patterns in source code, *cf.*, [10], [7], [15]. For instance, Lee *et al*., [10] use a static analysis to collect the structural aspect of software and a dynamic analysis to elucidate dynamical aspects of the software during the program execution such as the message passing between objects.

For both reengineering and code improvement purposes, Albin-Amiot et al. [1] present a toolset to help OO software practitioners design, understand, and re-engineer a piece of software using design patterns. Their prototype tool uses a constraint satisfaction technique to detect patterns within a given source code. It has the advantage of taking into account refractoring aspects and identifying distorted versions of the pattern in a source code. In addition, it can transform the source code so that it complies with the detected design pattern.

Besides the code, other works extract design patterns from a design. For example, the work of Tansalis [13] proposes a design pattern detection methodology based on similarity scoring between graph vertices. The graphs of the searched pattern and the examined design are encoded as matrices. These latter are then used to compute a similarity matrix. This matrix is calculated using the similarity scoring algorithm which has been proposed by Blondel et al. [2] The main drawback of the similarity scoring approach is the convergence time which depends on the graph size of the design.

Also within this matrix similarity-based approach, Dong *et al.* [5] use a template matching method to calculate the normalized cross correlation between the pattern matrix and the matrix representing a design segment. A normalized cross correlation shows the degree of similarity between the pattern and the design segment.

On the other hand, Florijin and Meijers [18] proposed a tool capable of detecting all pattern instantiations in an OMT design. The tool implements a graph matching technique. Once a pattern template is detected, the tool associates a set of roles to the classes composing the detected pattern instantiation. This information can assist the designer in validating the correct instantiation of a pattern. However, it does not tolerate any discordance between the design and the pattern.

A second purpose of pattern identification within a design is to improve the quality of the design. Within this context, Bergenti and Poggi [19] propose a tool, called IDEA, to improve UML designs (class and collaboration diagrams) using automatic pattern detection. Their method relies on a knowledge base where each pattern is described in terms of a structure template and a collaboration template (described as PROLOG rules). For example, to detect the Composite pattern, the system searches all triplet classes having the template structure *identical* to the composite. Thus, this method handles only an exact instantiation of the pattern. When IDEA finds a pattern instance, a set of design (PROLOG) rules are verified to test if the design could be improved. Then a set of critiques are proposed as possible design improvements. It is worth noting that the critiques/proposed improvements are pattern specific and they require a high level of understanding of both the design and the pattern.

The work of El Boussaidi and Mili [7] represents the design problem the pattern is meant to solve explicitly. It aims at recognizing occurrences of the modeled problem solved by a design pattern, which is then transformed according to the solution proposed by the design pattern. This work uses a meta model of the pattern problem to identify its instances in a given design. Once a problem is detected, it marks the appropriate entities and finally applies transformations to get the pattern solution. This work relies on graph modeling and transformation. One of its limits is that it focuses only on the pattern structure. However, among the essential constituents of a pattern (problem and/or solution) is the methods used.

Bouhours *et al*. [17] propose a detection approach for "bad smells" in a design that can be remodeled through the use of design patterns. A bad smell is any symptom that possibly indicates a design problem. The proposed approach can identify some spoiled patterns and their alternative model fragments. It uses a generator of OCL queries and a specific profile that encodes structural particularities of spoiled patterns. However, this work allows only exact matches with the spoiled patterns and considers only the structural information.

As summarized in table I, none of the proposed approaches combines the structural and dynamic aspects in their pattern identification. Except for Ka-Yee [21], none of the few works treating the dynamic aspect describes the behavior in terms of scenarios of ordered method invocations and tolerates behavioral variability. In fact, the dynamic aspect treated in the other approaches is limited to method calls between pairs of related classes, independently of the overall temporal behavior.

TABLE I. CURRENT PATTERN IDENTIFICATION APPROACHES

	Technique	Type	Tolerate variation	Aspect
[11]		Pattern, problem, ou spoiled	Yes/no	static & dynamic
[1]	constraint satisfaction	pattern	Yes	static
[13]	similarity scoring between graphs coded as matrices	pattern	Yes	static & partially dynamic (only method calls)
[5]	template matching between graphs coded as matrices	pattern	Yes	static & partially dynamic (only method calls
[7]	Meta-model of the problem and CSP	Pattern problem	Yes	static
[17]	OCL queries	Spoiled pattern	No	static
[21]	dynamic analysis & CSP	pattern	yes	dynamic

In addition, none of the proposed approaches for pattern detection can also detect spoiled patterns. Furthermore, none of these approaches instantiates a detected pattern within the examined design while highlighting the pattern variability and the role of each class in the design. Such information can assist the designer in understanding the pattern (or spoiled pattern) and validating its correct instantiation.

III. XML DOCUMENT RETRIEVAL AND PATTERN NOTATION

Our approach has a two-fold objective. The first objective is to identify correct and spoiled pattern instantiations within an application design, while tolerating certain variability. For this, we adapt an XML document retrieval technique that we overview in Subsection *A*. The second objective is to assist the designer in understanding and validating the instantiation of the identified pattern within the examined design. For this, we propose to use the P-UML notation which we briefly present in Subsection *B*.

A. XML document retrieval

XML document retrieval has been treated in the literature by several researchers. The most complete work has been proposed by Manning *et al.*, [11]. In this work, the authors adapt the vector space formalism for XML retrieval by considering an XML document as an ordered, labeled tree. Each node of the tree represents an *XML element*. The tree is analyzed as a set of paths starting from the root to a leaf. In addition, each query is examined as an *extended* query – that is, there can be an arbitrary number of intermediate nodes in the document for any parent-child node pair in the query. Documents that match the query structure closely by inserting fewer additional nodes are given more preference.

A simple measure of the similarity of a path c_q in a query Q and a path c_d in a document D is the following *context resemblance* function [11]:

$$C_R(c_q, c_d) = \begin{cases} \frac{1+|c_q|}{1+|cd|} & \text{if } c_q \text{ matches } c_d \\ 0 & \text{if } c_q \text{ does not match } c_d \end{cases}$$

Where:

- $|cq|$ and $|cd|$ are the number of nodes in the query path and document path, respectively, and
- c_q *matches* cd if and only if we can transform cq into cd by inserting additional nodes.

Note that the value of $CR(c_q, c_d)$ is 1 if the paths c_q and c_d in Q and D are identical. On the other hand, the more nodes separating the paths of Q and D, the less similar they are considered, *i.e.*, the smaller their context resemblance value will be.

B . P-UML: a design pattern notation

Several UML-based formalisms for pattern representation (*cf.*, [8] , [14])) were proposed. To account for the variability of a pattern, some proposed languages are able to distinguish the "regular" methods from the *hook* and *template* methods in a pattern, *cf.*, [5] [12]. These two types of methods encapsulate the pattern extensibility: template methods define abstract and generic behavior, while hook methods provide their implementation. However, none of the proposed languages both shows the variability and guides potential instantiations of the pattern and identifies the elements,

the structure and the role played by the elements of the pattern.

In response to the above shortages, we have proposed a notation for patterns, called P-UML [3], that is an extension of the UML class diagram. The extensions outline the roles played by both the classes as well as the methods within a design pattern. In addition, they link the pattern elements, and therefore help in visually distinguishing between different patterns used in a system design. Moreover, the extensions identify the pattern hot-spots and meta-patterns (template and hook methods). To illustrate the main concepts of the P-UML language, we will use the Composite pattern (Figure 1). The reader is referred to [3] for a detailed description of this language.

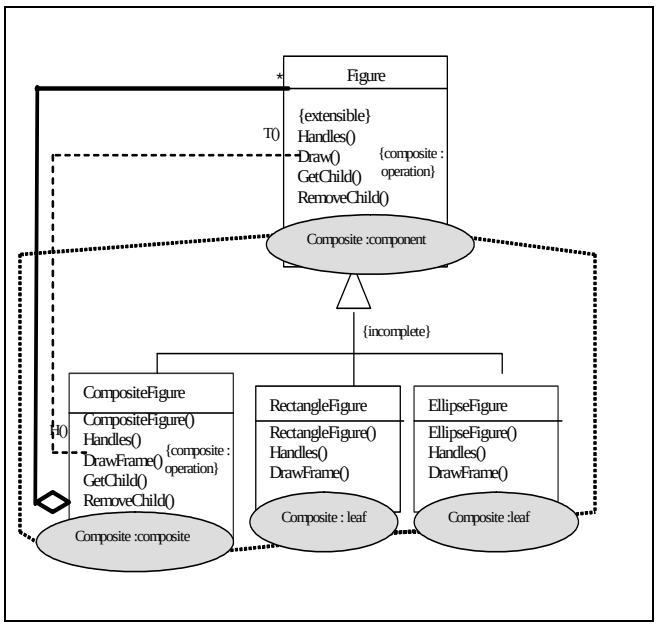


Figure 1. The *composite* pattern in the P-UML notation

Figure 1 shows an instantiation of the Composite pattern for an application in the graphical editor domain. The pattern participant roles (the ellipses) and their relationships are indicated in the instantiation. In addition, P-UML identifies the methods that play essential roles in the pattern: the Draw() method has been identified as a fundamental method in the composite pattern. A dashed line joins the hook Draw() and the template method DrawFrame(). The aggregation relation which is fundamental in the pattern is drawn with a highlight. Finally, P-UML delimits the pattern boundaries, which eliminates any confusion when multiple patterns are composed.

IV. PATTERN DETECTION TECHNIQUE

To detect patterns within a design, we take into account that a given pattern may be represented in various forms that differ from the basic structure without loosing the essence of the pattern. Thus, an exact pattern matching approach is insufficient.

On the other hand, the problem of finding an XML document (query) within a larger document while tolerating structural variations has been treated within the information retrieval domain. Several solutions were proposed to handle the structural differences that may exist between the query and a retrieved document. These solutions motivated us to convert design pattern detection into an XML document retrieval problem. More specifically, we consider a design pattern as an XML query and the design as the target XML document where the pattern is searched. In fact, since we consider the pattern and the examined design as two class diagrams, their transformation into XML documents is straightforward and can be handled by most existing UML editors. Furthermore, by transforming the pattern detection problem into an XML document retrieval problem, our approach can benefit from existing search engines.

```
<!ELEMENT classdef (name, inherit*,  composition*,
          association*,aggregation*, implements*, typdef*, op*)>
<!ELEMENT inherit (type*, incompletetag?) >
<!ELEMENT composition (type*) >
<!ELEMENT aggregation(type*) >
<!ELEMENT  typedef     (name+, type+) >
<!ELEMENT  predcalc    (predcalc*, type *) >
<!ELEMENT op (name, returntype+,param*)>
<!ELEMENT  param     (name+, type +) >
<!ELEMENT  name    (#PCDATA) >
<!ELEMENT type (#PCDATA)   >
<!ELEMENT incompletetag (#PCDATA)   >
```

Figure 2. DTD extract for the UML class diagram

To illustrate our method, we will focus on the structural features of the corresponding class diagram: the classes, generalizations, aggregations, compositions, etc. For this, we will use the DTD shown in Figure 2 to transform the pattern and the design into XML documents. Note that each tree in these XML documents is composed of class nodes interconnected by relation nodes (generalization, association, etc). In addition, each path in a tree contains relation nodes from the same type.

In XML document retrieval in general, the context resemblance function (CR) is calculated based on an exact match between the names of the nodes in the query and the document paths. However, for pattern detection, the nodes representing the classes are often different in the pattern from those in the design. Thus, we first need to calculate the resemblance values for the various matches between the class nodes in the query (pattern) and those in the design. Secondly, we need to take into account: 1) the number of times a given match between two class nodes is used to calculate CR; and 2) the importance of each relation in the pattern.

A. Resemblance determination

The resemblance between a pattern and a design starts by computing the resemblance between each path of the pattern to all the paths in the design. In this computation, we assume that the structural variability should be limited between the pattern and a potential instantiation in the design. That is, we assume that a design path may differ

from a pattern path by adding at most N nodes compared to the longest path of the pattern. The larger the N, the more scattered the pattern instantiation would be in the design, which might loose the pattern essence.

To determine the resemblance between a pattern Q and a document D, we proceed as follows:

1. L := the number of class nodes in the longest path in Q;
2. N := the maximum number of intermediate/additional nodes in the design path;
3. For each path P_q in the pattern Q
3.1 For each path P_d in the document D
3.1.1 If P_d and P_q have different types of relations
3.1.2 then CR(P_q , P_d) := 0
else
//compare P_q with all sub-paths in P_d starting
// from different nodes
3.1.3 For s=1 to | P_d|-1
// tolerate at most w additional nodes
3.1.3.1 For w=1 to min(L+N, | P_d|-1)
3.1.3.1.1 $P'_d := P_d [s .. s+w]$
3.1.3.1.2 Compute CR(P_q , P'_d)
4. Compute the weighed sum of all CR scores for *all* the paths and store them in *CRMatrix(Q,D);*
5. Normalize *CRMatrix(Q,D)* by dividing each entry by the number of classes in D

In step 3.1.3, we consider that the match between the pattern path and the design path may not necessarily start at the root node; for this we need to consider all possible sub-paths of the design. These sub-paths start at different class nodes in P_d. In addition, since the structural difference between the pattern path and the design path is limited, then each sub-path can cover at most L+N class nodes; thus the number of sub-paths to be considered is reduced. This in turn limits the temporal complexity of the algorithm. The tolerated maximal intermediate nodes N can be fixed by the designer.

In step 4, we sum up in *CRMatrix* the resemblance scores (*i.e.,* correspondences) between the classes of the design and the classes of the pattern. This weighted sum accounts for the importance of the relations in the pattern. Finally, in step 5, these scores are normalized with respect the total number of classes in the design; the final matching results are collected in *NormalizedCRMatrix* whose columns are the classes in the pattern and whose rows are the classes of the design. Now given this matrix, we can decide upon which correspondence better represents the pattern instantiation: For each pattern class, its corresponding design class is the one with the maximum resemblance score in *NormalizedCRMatrix*.

On the other hand, given two designs $D_1$and D_2, to decide upon which design better instantiates a pattern P, we first compute their normalized resemblance matrices. Secondly, we compute the sum of the normalized resemblance scores for all the matched pattern classes in $D_1$and D_2; the design with the maximum sum is the one that better instantiates the pattern.

Note that in the worst instantiation, each pattern class must be matched to at least one class in the design; thus, on average, the sum of the normalized resemblance scores of the matched classes should not be less than the number of classes in the pattern divided by the number of classes in the design.

B. *Example: detection of the composite pattern*

Graphics applications like drawing editors let users build complex diagrams out of simple components. The user can group components to form larger components. For instance, we can define classes for Text and Lines and classes that act as containers for these classes. However, when using these classes, the primitive classes such as Text and Line have to be treated differently from the container class: This is the composite pattern problem instantiated in the graphical editor application shown in Figure 1.

The composite pattern (Figure 3.b) applies to compose the objects in a tree structure where individual objects as well as the composed objects behave uniformly. Composed objects delegate the requests to the individual leaf objects.

To detect the composite pattern, we have to identify its structure. Let us try to identify it in the design of Figure 3.a. The XML corresponding paths are illustrated in Figure 4.

Table II shows a sample of the resemblance function scores corresponding to the design and pattern paths of Figure 4. Recall that some concepts are more essential in a pattern than others. In the composite pattern, let us consider that the aggregation relation is twice as important as the inheritance relation. Thus, when collecting the CR scores in the resemblance matrix, the score of the aggregation match is multiplied by two.

The normalized CR matrix identifies the composite design pattern correctly and indicates that the class A matches the *component* class, the class D matches the *composite* and consequently the class C matches the *Leaf* class. Note that the match score of the class C to *Leaf* is equal to the match score of C to *composite* (0.39); however, since D has been identified as *composite* with a greater matching score (0.725), then C is identified as *Leaf*.

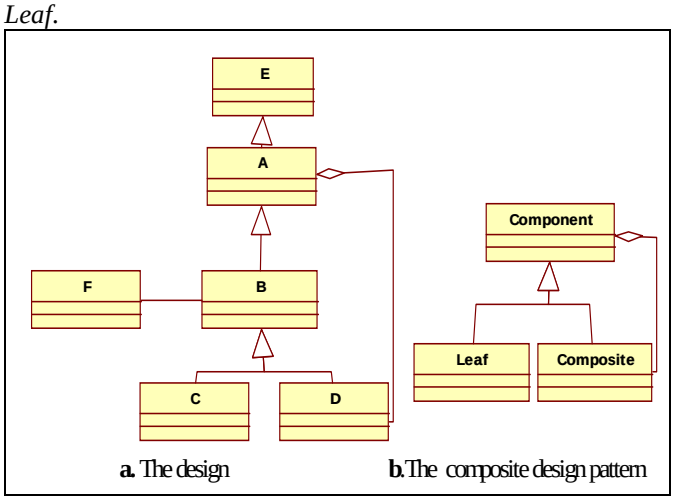


Figure 3. A sample design (a) and the composite pattern (b)

Furthermore, the sum of the maximum normalized CR for the nodes of the pattern (2.365) is greater then the threshold which is equal to 3/6; thus this identification is acceptable.

Given the above matching, we can represent the Composite pattern within the design through the P-UML language as shown in Figure 5. Through this representation, the designer can better understand the roles of his/her design classes as indicated in the identified pattern.

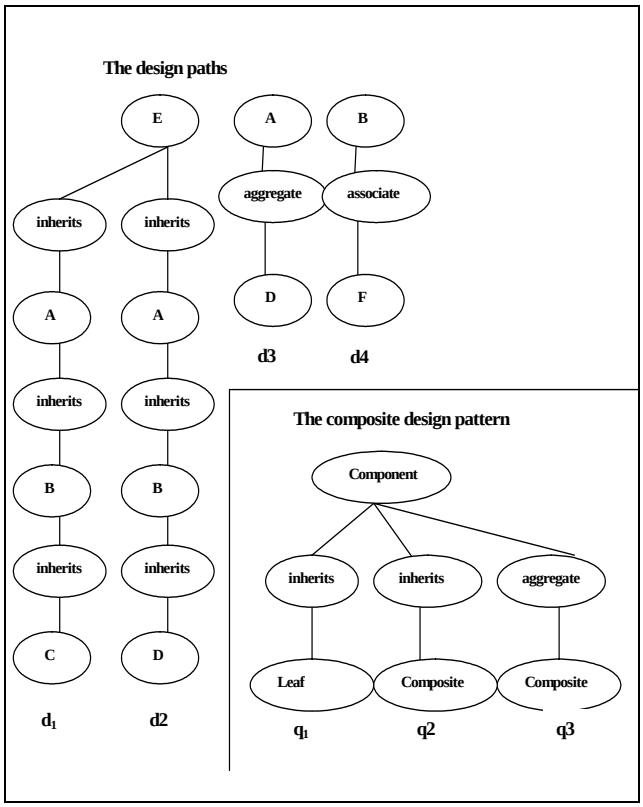


Figure 4. XML document trees for the example of Fig 3.

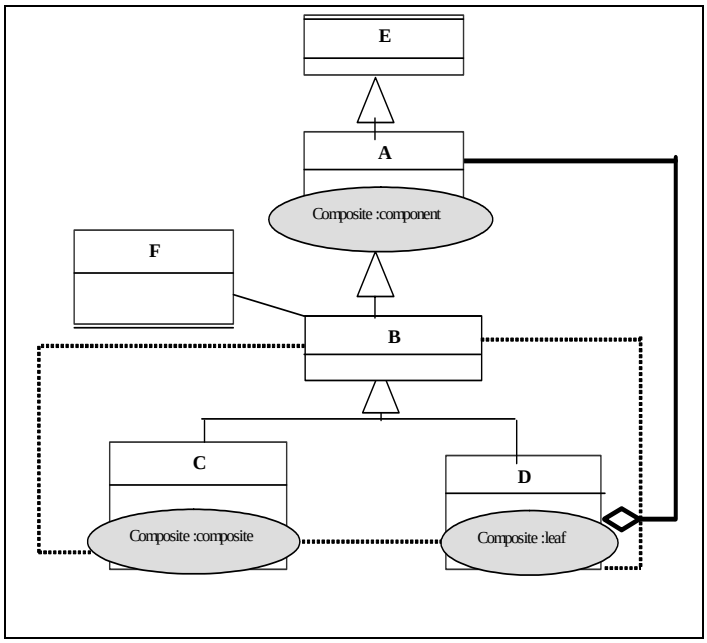


Figure 5. The identified composite pattern with P-UML

TABLE II. SAMPLE CONTEXT RESEMBLANCE SCORES AND NORMALIZED MATRIX

	component —inherits→ leaf	component —inherits→ composite	component —aggregates→ composite
E —inherits→ A	**CR(*cq*1 , *cd*1) =1 If Component=E Leaf=A**	**CR(*cq*1 , *cd*1) =1 if Component=E Composite=A**	**0**
A —inherits→ B	**CR(*cq*1 , *cd*1) =1 If Component=A Leaf=B**	**CR(*cq*1 , *cd*1) =1 if Component=A Composite=B**	**0**
E —inherits→ A —inherits→ B	**CR(*cq*1 , *cd*1) =0.75 If Component=E Leaf=B**	**CR(*cq*1 , *cd*1) =0.75 if Component=E Composite=B**	**0**
A —aggregates→ D	**0**	**0**	**CR(*cq*2, *cd*2) =2 if Component=A Composite=D**
F —associates→ B	**0**	**0**	**0**

$$\text{NormalizedCRMatrix(Design, pattern)} = \begin{array}{c} \\ E \\ A \\ B \\ C \\ D \\ F \end{array} \begin{array}{c} \begin{array}{ccc} \text{component} & \text{composite} & \text{leaf} \end{array} \\ \begin{pmatrix} 5.9 & 0 & 0 \\ 7.5 & 1 & 1 \\ 4 & 1.75 & 1.75 \\ 0 & 2.35 & 2.35 \\ 0 & 4.35 & 2.35 \\ 0 & 0 & 0 \end{pmatrix} \end{array} / 6 = \begin{array}{c} \\ E \\ A \\ B \\ C \\ D \\ F \end{array} \begin{array}{c} \begin{array}{ccc} \text{component} & \text{composite} & \text{leaf} \end{array} \\ \begin{pmatrix} 0.983 & 0 & 0 \\ 1.25 & 0.16 & 0.16 \\ 0.66 & 0.29 & 0.29 \\ 0 & 0.39 & 0.39 \\ 0 & 0.725 & 0.39 \\ 0 & 0 & 0 \end{pmatrix} \end{array}$$

C. *Detection of the spoiled composite pattern*

As we are able to detect design patterns, we are also able to detect spoiled patterns. An example of a spoiled composite pattern is illustrated in Figure 6. It shows a typical composite object structure of recursively composed graphic objects. The Figure class is composed of other Figures which could be composed of lines, texts and rectangles. In this spoiled pattern the pattern quality rules or the decoupling and extensibility properties of the pattern are not respected. Moreover, the fact that *Line*, *Text* and *Rectangle* do not inherit from *Graphic* will cause some duplication of code with excessive use of delegation [17].

A correction of this spoiled pattern is brought by the composite pattern which defines an abstract class that represents both primitive and container classes: the class *Graphic* shown in Figure 8.

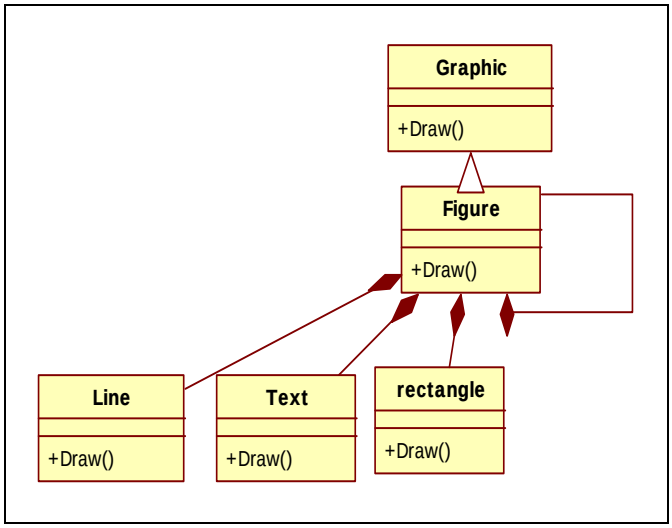


Figure 6. A spoiled composite pattern [17]

To detect the spoiled Composite, an abstraction of the spoiled pattern is necessary. The abstraction is shown in Figure 7. Now, let us consider the design fragment illustrated in Figure 9 and determine if it is similar to the spoiled composite abstraction illustrated in Figure 7.

Similar to pattern detection, the design is converted into XML trees as illustrated in a graphical format in Figure 10. Some of the context similarity function scores are illustrated in table III and the resulting Normalized matrix is shown below.

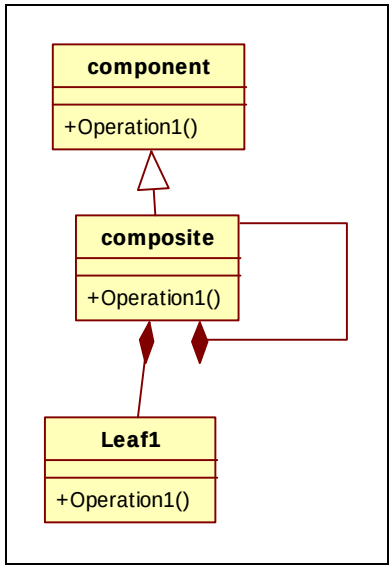


Figure 7: An abstraction of a spoiled composite pattern

The normalized CR matrix identifies the spoiled composite design pattern correctly. Thus, the class *Graphic* is identified as the *component* class, the class *Figure* is identified as the *composite* class. The classes *TextFigure*, *TriangleFigure* and *EllipseFigure* match the *Leaf* class. Note that the match scores of these classes to *Leaf* is equal to their match scores to *composite* (1/6); however, since the *composite* has been identified with a greater matching score (10/6), then they are identified as *Leaf*.

Given the above matching, we can substitute the spoiled Composite pattern with a correct instantiation of a composite pattern.

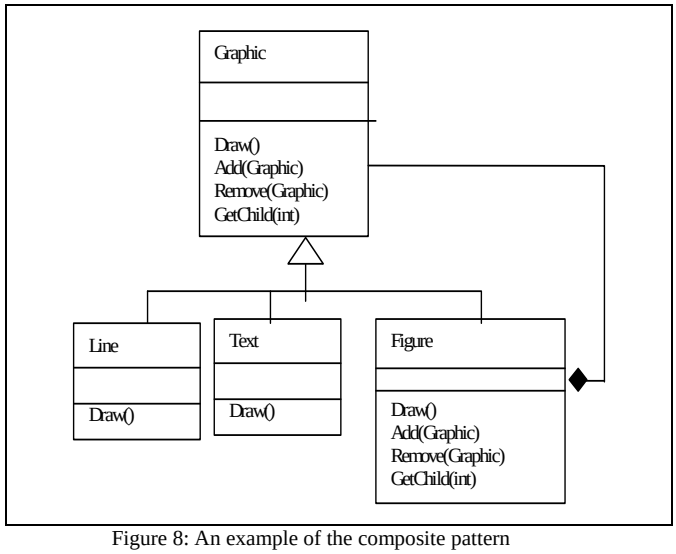


Figure 8: An example of the composite pattern

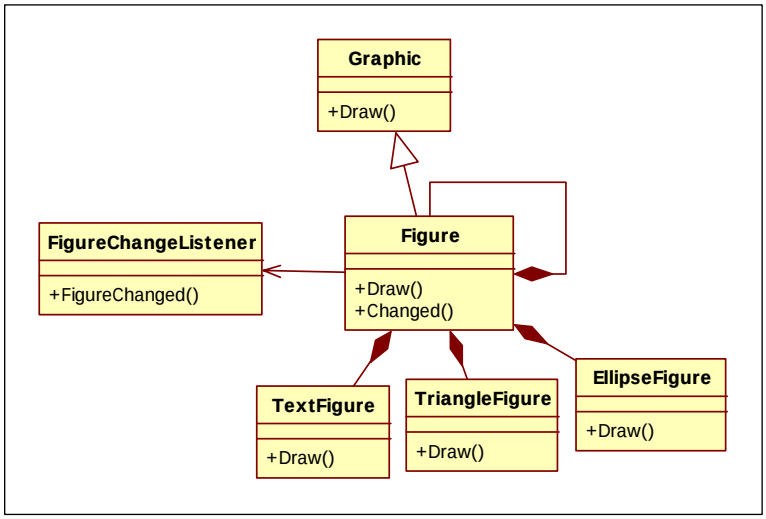


Figure 9: a fragment of a design

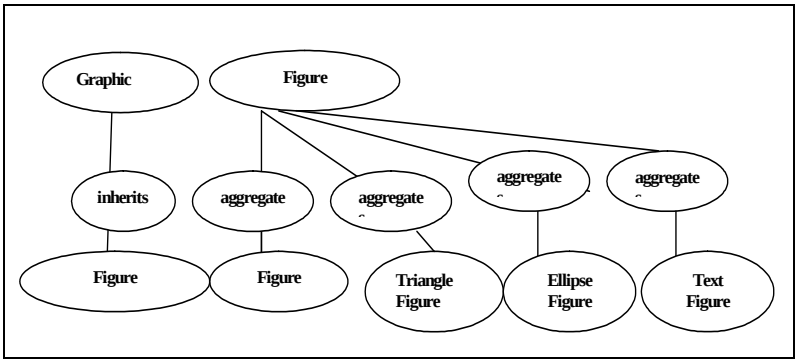


Figure 10: XML document trees for the spoiled composite pattern

TABLE III. CONTEXT RESEMBLANCE SCORES FOR SPOILED COMPOSITE PATTERN DETECTION

	Component $\xrightarrow{\text{inherits}}$ ***Composite***	***Composite*** $\xrightarrow{\text{aggregates}}$ ***Composite***	***Composite*** $\xrightarrow{\text{aggregates}}$ ***Leaf***
Graphic $\xrightarrow{\text{inherits}}$ ***Figure***	$CR(c_q, c_d) = 1$ *if Component=Graphic Composite=Figure*	*0*	*0*
Figure $\xrightarrow{\text{aggregates}}$ ***Figure***	*0*	$CR(c_q, c_d) = 1$ *if Composite=Figure Composite=Figure*	$CR(c_q, c_d)) = 1$ *if Composite=Figure Leaf=Figure*
Figure $\xrightarrow{\text{aggregates}}$ ***EllipseFigure***	*0*	$CR(c_q, c_d) = 1$ *if Composite=Figure Composite= EllipseFigure*	$CR(c_q, c_d) = 1$ *if Composite=Figure Leaf=EllipseFigure*
Figure $\xrightarrow{\text{aggregates}}$ ***TextFigure***	*0*	$CR(c_q, c_d) = 1$ *if Composite=Figure Composite=TextFigure*	$CR(c_q, c_d) = 1$ *if composite=Figure Leaf=TextFigure*
Figure $\xrightarrow{\text{aggregates}}$ ***TriangleFigure***	*0*	$CR(c_q, c_d) = 1$ *if composite=Figure composite=TriangleFigure*	$CR(c_q, c_d) = 1$ *if composite=Figure Leaf= TriangleFigure*
Figure $\xrightarrow{\text{asociates}}$ ***FigureChangeListener***	*0*	*0*	*0*

$$\text{NormalizedCRMatrix(PatternPb, Design)} = \begin{array}{c} \\ \textit{Graphic} \\ \textit{Figure} \\ \textit{EllipseFigure} \\ \textit{TextFigure} \\ \textit{TriangleFigure} \\ \textit{FigureChangeListener} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{Component} & \textit{Composite} & \textit{Leaf} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 10 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array} /6$$

V. CONCLUSION

Design patterns ensure an improvement of design quality, traceability and a better documentation [9]. However, the difficulty of their detection and instantiation reinforces the need for a technique that automates these tasks. This paper proposes a new approach for pattern and spoiled pattern detection and instantiation.

The proposed approach adapts an XML document retrieval technique. That is, it considers a design pattern (or spoiled pattern) as an XML query to be found in an XML document representing a design. It uses the context similarity function [11] to determine the most probable correspondences between the classes of the design and those in the pattern (or spoiled pattern). It has the advantage of tolerating certain structural differences in the design compared to the (spoiled) pattern; the designer can fix a threshold below which the differences are un-tolerated. In addition, our approach can be applied for both structure and method correspondences. Furthermore, once a (spoiled) pattern instantiation is detected, the correspondence information produced by our approach can be exploited to represent the design fragment with the (spoiled) pattern elements. This representation assists the designer in understanding the found (spoiled) pattern within the context of his/her application. Moreover, it allows him/her to validate (correct) the instantiation of the (spoiled) pattern.

Our future works deal with three axes. In the first, we are examining how to add more intelligence in our assistance for the recognition of pattern problems inside a design: how to alleviate the search task by adding priorities. In the second axe, we are seeking to exploit the information of the collaboration diagrams to best handle the (spoiled) pattern detection. In the third axe, we are looking into the formalization of design patterns. This will provide us with two benefits: 1) precise definition of patterns, and 2) analysis facilities to validate a pattern instantiation.

REFERENCES

[1] H. Albin Amiot, P. Cointe, Y. G. Guéhéneuc, "Un meta-modele pour coupler application et détection des design patterns", *L'objet*, N° 8, 2002, pp1-18.

[2] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren, "A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching", *SIAM Review*, Vol. 46, N°. 4, 2004, pp. 647-666.

[3] N. Bouassida and H. Ben-Abdallah, "Extending UML to guide design pattern reuse", *Sixth Arab International Conference On Computer Science Applications*, March 2006.

[4] K. Brown, "Design reverse-engineering and automated design pattern detection in Smalltalk". Technical Report TR-96-07, University of Illinois at Urbana-Champaign, 1996.

[5] J. Dong, "UML extensions for design pattern compositions", *Journal of object technology*, Vol. 1, N° 5, 2005, pp 149-161.

[6] J. Dong , Y. Sun and Y. Zhao, "Design pattern detection by template matching". SAC'08, Ceara, Brazil, March 16-20, 2008 .

[7] G. El Boussaidi and H.Mili, "Detecting patterns of poor design solutions by using constraint propagation", In MODELS, *Proceedings of the 11th international conference on model driven engineering languages and systems*, 2008.

[8] M.F. Fontoura, W. Pree and B. Rumpe, "Extending UML to improve the representation of design patterns", *JOOP*, Vol. 13, N°11, 2001, pp. 12-19.

[9] E. Gamma, R. Helm, R. Johnson and J. Vlissides, 1995, *Design patterns: Elements of reusable Object Oriented Software*, Addisson-Wesley, Reading, MA, 1995.

[10] H. Lee, H. Youn, E. Lee, "A design pattern detection technique that aids reverse engineering". In the *International Journal of security and applications*. Vol 2, N°1 January, 2008.

[11] C.D. Manning, P. Raghavan and H. Schütze, *An introduction to information retrieval*, Cambridge University Press, England, 2008.

[12] W. Pree, "Meta-patterns: a means for capturing the essentials of object-oriented designs", *Proceedings of the 8th European Conference on Object Oriented Programming*, Bologna, Italy, 1994.

[13] N. Tsantalis, A. Chatzigeorgiou, G. Stephanides and S. T. Halkidis, "Design pattern detection using similarity scoring" In IEEE transactions on software engineering, vol 32, N°11, san franciso,USA,2006.

[14] Y. Sanada and R. Adams, "Representing Design Patterns and Frameworks in UML-Towards a Comprehensive Approach", *Journal of Object Technology*, Vol. 1, N°2, July-August, 2002.

[15] N. Shi and R.A. Olsson. "Reverse Engineering of Design Patterns from Java Source Code", P*roceedings 21st IEEE International Conference on Automated Software Engineering*, 2006.

[16] XML Metadata Interchange: OMG Document ad/98-07-03, July 6, 1998

[17] C. Bouhours, H. Leblanc and C. Percebois, "Bad smells in design and design patterns". In *Journal of Object Technology*, Vol 8, N°3, May-June, 2009.

[18] G. Florijin, M. Meijers and P. Van Winsen, "Tool support for object oriented patterns", In *Proceedings of the European conference on object oriented programming: ECOOP*, 1997.

[19] F. Bergenti and A. Poggi, "Improving UML design pattern detection", In p*roceedings of the 12th international conference on software engineering and knowledge engineering SEKE*, 2000.

[20] N. Bouassida and H. Ben-Abdallah, "Structural and behavioral detection of design patterns". In: *International Conference on Advanced Software Engineering & Its Applications (ASEA)*, December 10-12, Jeju Island, Korea, LNCS Proceedings, Springer, 2009.

[21] J. Ka-Yee Ng and Y. G. Gueheneuc, "Identification of behavioural and creational design patterns through dynamic analysis", *Proceedings of the 3rd International Workshop on Program Comprehension through Dynamic Analysis (PCODA)*, October 2007, pp 34-42.

On-line Retrieval and Management System based on Wireless Communications for Schools

Adnan A. Tala'a
Ministry of Education, Private Education Sector, Irbid, Jordan
malrousan@gmail.com

***Abstract*—Information keeping and manipulation is vital for any society. This paper aimed at combining the power of wireless communications with the strength of database information keeping through Internet as a media to make it possible to implement an automated record keeping system. The system not only facilitates remote record keeping, but also utilizes very popular devices to do so. The minimal cost and friendly environment of those devices adds to the advantages of the system. The way the system was implemented is by using a wireless portable device that sends a number of requests to Java servlets residing on a Web Server Which will manipulate the data, store it on the database and respond to the wireless devices with required information. The paper uses University Attendance taking as a case study that can be modified to be applicable for any of the record-keeping applications.**

***Index Terms*—Automation, Record Keeping and Retrieval, Wireless, GSM, Attendance.**

I. Introduction

Due to the rapid development in computer and network technology, the use of the Internet has been expanding exponentially. It is now extensively used as a reference tool for personal, educational, commercial, and industrial use. For many years the Internet has been used extensively in browsing homepages, searching for information, chatting, downloading and uploading information. The rapid development of new technologies such as Java, made the Internet an efficient medium that allows monitoring, control, and interaction with machine and devices.

In recent years and due to the revolution in information technologies there has been a great interest in rerecord keeping and retrieval. An record keeping is an electronic system in which records are collected, organized, and categorized to facilitate their preservation, retrieval, use, and disposition may be either, a distinct system designed specifically to provide recordkeeping functionality, or part of another system. A distinct electronic record keeping system will comprise an application program which provides recordkeeping functionality, data and metadata needed for management of the records controlled by the system, and any electronic records managed by the system. An electronic recordkeeping system may be part of another system, such as an application system or an electronic document management system, when the design of that system includes record keeping functionality [1,2].

This work proposes the design and implementation of a distinct electronic record keeping system. The system utilizes the state-of-the art technologies in wireless communications using advanced software engineering approaches.

Record keeping has been used in the core of many management systems. The work presented by Safran [3] implements an electronic record keeping system for patients across all medical hospitals in Singapore. The system utilizes the power of the Internet and database systems. Similar systems are presented for medical applications in [4-8]. The above systems propose solution of record keeping providing records are created and entered manually to the computerized system. Our system automates the whole process; from data entry to report generation.

Among many applications that require record keeping, attendance taking for class session at schools is, probably, the most demanding one. Almost, in all course syllabi instructors highlight the statement "*class attendance is mandatory for all students; multiple absences will affect a student's final grade for the course.*" Currently, most instructors take attendance in a classical way during class session. This task is time-consuming and inefficient, especially for class with large number of students attending it. For online courses, this is an easy task since class attendance is automatically logged whenever a student performs a task on the course page. In general, as the mass of information for a class session accumulates, the methods of organization, storage, and retrieval of that information become extremely important. Florida Department of Education, for example, has made the use of automated record keeping a mandatory on school [9]. The statute in [9] further requires that the enrollment register shows the absence or attendance of each student enrolled for each school day of the year in a manner prescribed by the State Board of Education. Although the system automatically generates various types of attendance and absence reports, yet, instructors enter daily data manually, consuming time and efforts.

Thus, the need fully automated record keeping will be soon becoming a must for several applications [10]. We are not aware of a published record keeping system that is fully automated. In this paper we present a complete implementation for record keeping and retrieval system. We take class attendance as case study for such system. However, it easily can be mapped to any application with similar needs.

doi:10.4304/jetwi.2.3.176-181

This paper is organized as follows. Section 2 presents the system requirements including functional and sites requirements, followed by an overview of the system in Section 3. The software design and development is presented in Section 4. Then a discussion for the implementations is made in Section 5. Results and conclusion is presented in Section 6, followed by the conclusion in Section 6.

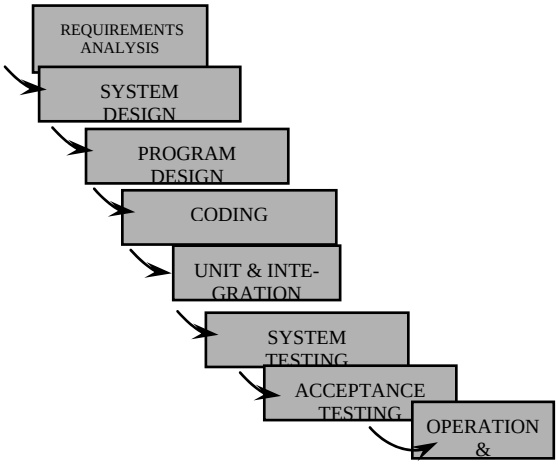


Figure 1. Waterfall model.

II. Functional and Performance Requirements

As shown in Figure 1, as the design methodology, we chose to use the software engineered waterfall model because it helps to understand the nature of the system we build. This section specifies the software requirements for the proposed system. The specifications of the system described here are based on the needs of and constraints recommended by experts in the fields [9, 10]. These requirements are as follows.

A. Class Room Site Requirements

- Each student should have an identification (ID) card number that carries a barcode.
- A barcode reader shall capture students' IDs at class room during a class session.
- The reader should work with the existing portable wireless devices that have the capability to connect to the Internet.
- The readings (IDs) shall be transferred to the wireless device.
- The readings shall be sent to the server immediately after every scan, or after a group of scans.
- A reading is sent to the server via a wireless connection.
- In case of remote class, readings are sent via the GSM/GPRS network as a message.
- The instructor should be able to access the attendance system during and after class sessions.
- All clients including instructors and students should be able to submit their requests to the server via Internet browsers.
- Requests issued by instructors shall be able to run on a handheld device.

B. The Server Site Requirements

- The attendance server should stay functional and up all the time regardless of wireless (GSM) network availability.
- The server should continuously receive the SMS messages sent by all wireless devices from class rooms sites
- The server should extract the request details from the SMS and do the necessary modifications and analysis on the attendance data base.
- The server should continuously update each class profile accordingly.
- The server should automatically generate reports up a request from the client.
- The system shall send the requested information to the client's device via SMS messages.
- The system shall receive messages from clients (students) to inquire about their relevant information.

C. Attendance Database Requirements

- The database should keep track of all users allowed to access the database and their password for authentication purposes.
- The database should keep track of the personal information of each student such as student name, ID, mobile number, region, telephone and P.O. Box.
- The database should keep track of all students attendance for predefined period of time (e.g. one semester/year).
- The class attendance can be accessed at anytime by all authorized instructors.
- The number of class attendance for a student can be accessed at anytime by the associated instructor and by the authorized student.
- The database should be able to support dynamic attendance policy where needed.

Based on the above requirements, the sequence diagrams for the systems are depicted in Figure 2 through 4.

III. System Overview

The automated record keeping system provides an easy and handy way to keep track of those students who miss classes during an academic semester or year. It is aimed at making the attendance taking process an easy and fast one via the use of the Internet and wireless technologies. As shown in Figure 5, all what is needed for implementing such a system includes a barcode reader, a wireless portable device, and a host server. The barcode reader is used for scanning the IDs of students while entering the class room. The wireless device (A PDA in our case) is sued as an intermediate point which establishes a communication link between the class room and the attendance server. Hence, the instructor is

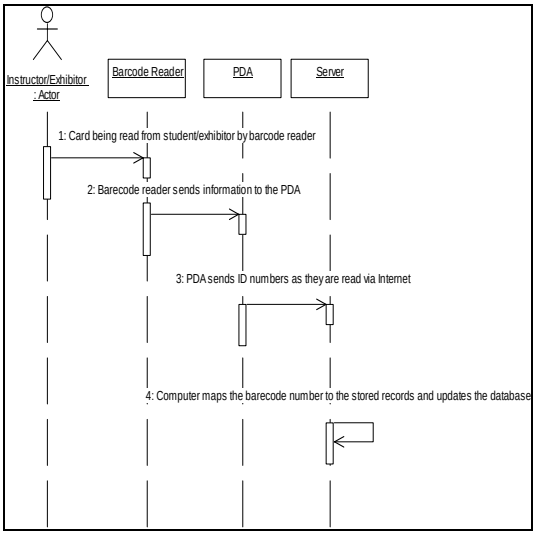


Figure 2. Operational Sequence Diagram.

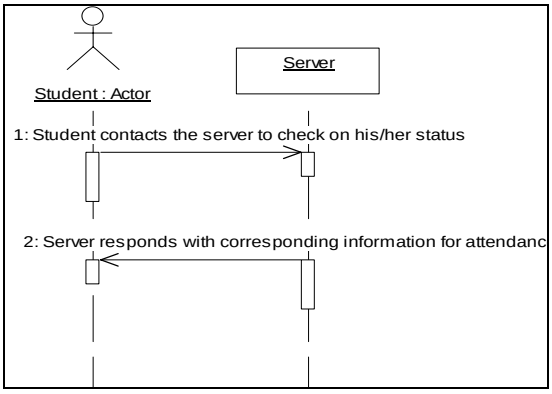


Figure 3: Student's Sequence Diagram

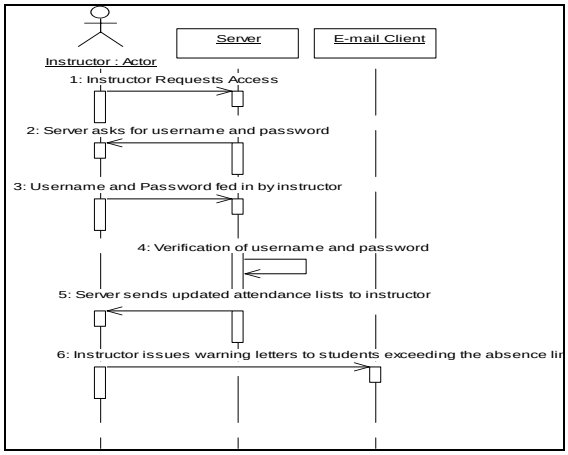


Figure 4: Instructor's Sequence Diagram

required to bring his/her PDA to the class room. It receives the reading from the barcode and, thereafter, each ID read will be immediately written into the main database on the server which is running remotely. Connection to the Internet is established via the card phone installed on the PDA using the PCMCIA jacket or via wireless access points that are present. The mobile phone shown in Figure 5 is used in case the PDA has no mechanisms for mobile communications. The instructor may check attendance status in-class, or later using a nice graphical interfaced JAVA application that runs on his/her desktop, as shown in Figure 6.

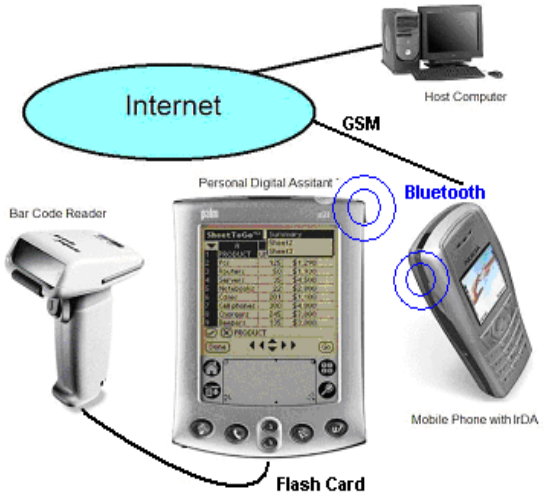


Figure 5. System Components

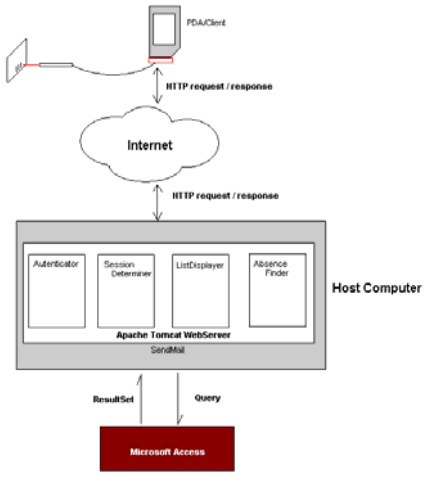


Figure 6. Client Java-based Applications.

IV. Java Servlets for Software Development

Many considerations were kept in mind while building the system. The initial thought was to have a Java program running on the PDA which requires either Java 2 Micro Edition [11, 12]. Java has been proven as a powerful and efficient tool for real time systems [13-15]. However, since the goal was to make it easy to implement the system on any PDA, the idea was modified. The entire work is done on the server and the PDA exchanges HTTP requests and responses with the host computer. Having done this, the overhead of programming on the PDA was eliminated.

Table 1. Functions of Java components.

Java Component	Function
Authenticator. java	A servlet page to enable user authentication through a login-ID and a password.
SessionDeterminer .java	If the user is granted access to the system, this servlet will ask him or her for the session number.
ListDisplaye.java	This servlet provides the means for inputting the student ID and displaying the list of entered ID's. It also validates whether the entered ID is among the registered ID's of the system.
AbsenceFinde.java	Once ID's are entered, the list is passed to the absence finder that issues the required SQL queries using JDBC environment to determine absent students. It also updates the student absence records and asks the instructor if he or she wishes to send warning letters to students exceeding a certain number of absences. Makes communication through get Data and User State beans
get Data & ***UserSate***	Directly connected to the database and the actual performers for the Absence Finder requests.
SendMail.java	A normal java class that utilizes Simple Mail Transfer Protocol (SMTP) to send the warning letters to the selected students.

The host computer runs a number of servlets and other Java components that reside on Apache Tomcat Web Server [12] to handle the HTTP requests and responses and perform the tasks of authenticating, querying the database and returning back the result to the browser (See Figure 5). There are five different components running on the application server. These are namely, *Authenticator*, *Session Determiner*, *List Displayer*, *Absence Finder*, and *Send Mail*. The first four components are written as servlets, whereas the last component is a normal java class. The function of each component is shown in Table 1.

The interactions taken place between different software components in the system are shown in the following steps and illustrated in Figures 7 and 8.

(1) The Client requests a pass
(2) Authenticator confirms identity from the database and forwards if pass to session determiner.
(3) Session determiner returns an HTML page that asks for session information.
(4) Users selects the required sessions and return it to session determiner for integrity verification i.e. attendance for this particular session was already taken.
(5) Session determiner forwards the control to List Displayer.
(6) List Displayer asks for IDs' of available students.
(7) List Displayer makes a list of those IDs' and shows the list to the client.
(8) The list is passed to the Absence Finder that will get the student who are absent by comparing the list of attending students and the class list.
(9) The Absence Finder will send back to the client the IDs of absent students. And will ask the client if a warning letter is to be sent to absent students.
(10) The response of the client will be returned to server to be executes.

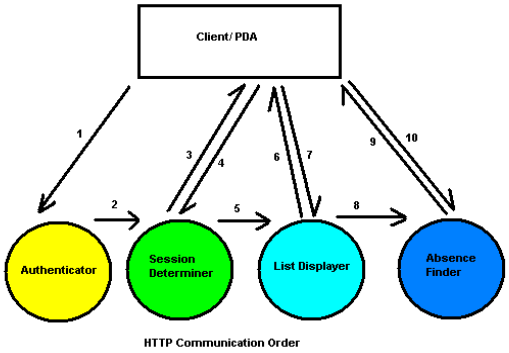


Figure 7. Software Interactions

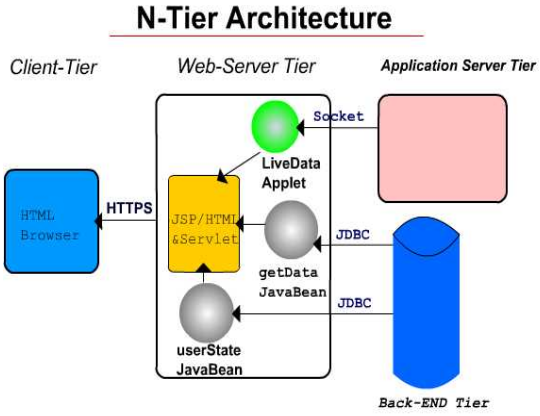


Figure 8. Architecture of the Software Engine

V. IMPLEMENTATIONS

Moreover, our system was supposed to connect to the internet via a mobile using Bluetooth technology. The initial implementation has included such option, Then we thought of provide an alternative implantation. Thus we decided to use a PCMCIA [16] jacket in which we can either insert a WIFI wireless card that connects to the internet using the available access points or a PC phone

card that connects to the internet using the GSM network. Our approach will not only reduce the hardware tools involved in the system, i.e., mobile device, but will also increase the range that the PDA works on, thus making it a better system for future expansions. However, the use of Bluetooth will make it cheaper to implement the system for small scale applications.

Although GSM was proven to an efficient communication technology in many applications [17-20], the use of GPRS technology reduces the cost of mobile communications. This is mainly because GPRS has higher data rate and carries more information within one packet [21].

For the web server we have used the Tomcat Apache server which is a free and open source developed by Sun Corporation [12]. It is easy to use and provide easy mechanisms for interacting with most of database server including Microsoft access. After dealing with system, we recommend the use of MySQl server for the database since it is an open source and compatible with the Apache Server [12]

VI. Results and Discussion

Based on the above development process we have built the system with all requirements in minds. The system has been tested and proven to a working tool for school learning process. Figure 9 shows the system waiting for the user commands. Recall that there are two types of users: instructors and students. The instructor can interact with the system using a PDA during the class session, or from the office using a desktop. Figure 10 shows a list of students' names who are absent for a given session.

Another interesting scenario is shown in Figure 11 in which the system displays a list of names who have exceeded the allowed number of sessions that a student can be absent. The system here triggers the instructor to send warning letters to these students in form of email messages. At any time the instructor can view the entire class attendance record for all her/his sections, as seen in Figure 12. However, a student can only check his attendance in the courses he or she is registered in, as see Figure 13

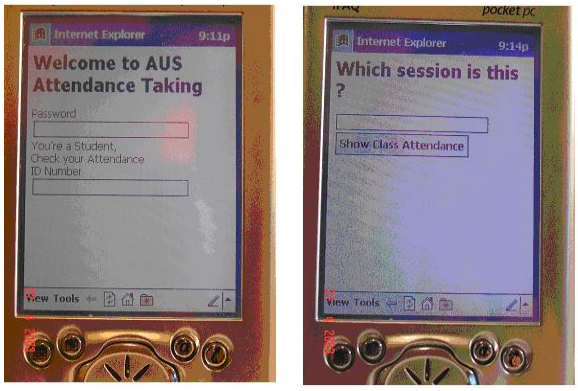


Figure 9. Record keeping system on the PDA.

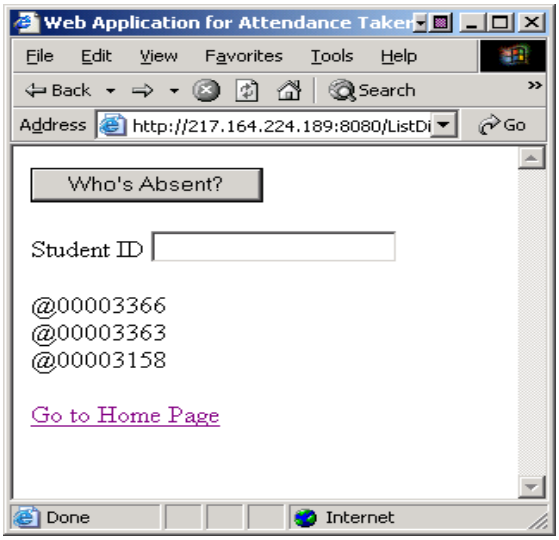


Figure 10. Absent Students for a Session.

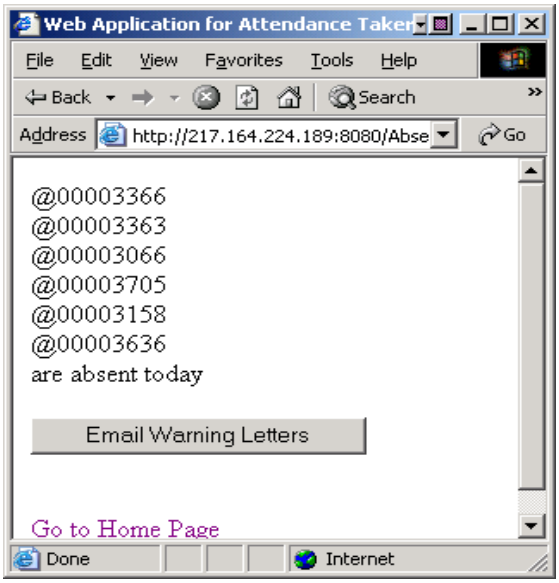


Figure 11. Issuing Warning for students exceeding limits.

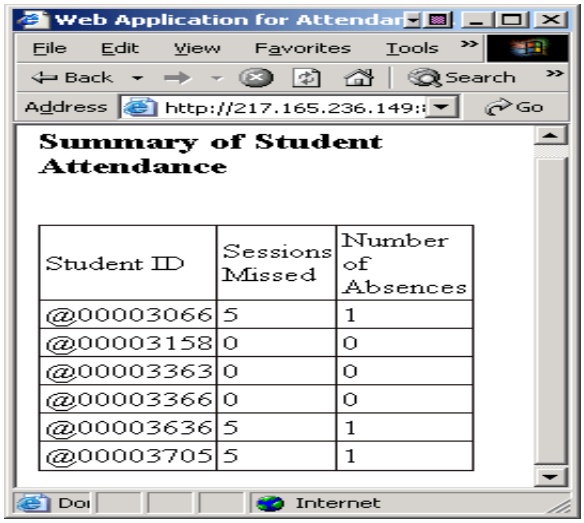


Figure 12. Summary for Instructor.

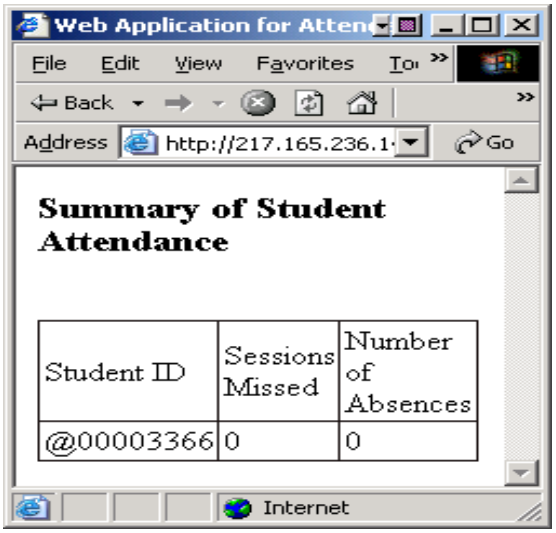


Figure 13. A report requested by a student.

VII. Conclusion

The system designed and implemented in this work fulfills society needs in many fields in instant information retrieval and management. Out objective in this work was after a solution that is easy to install and use, yet powerful and efficient. The main goal was to utilize off-the-shelf hand held devices that are already popular without the need for any external complex circuitry in order to make the system available to society.

The system components include a PDA, a bar code reader (or any other input device that can be interfaced with the PDA), a mobile phone and the host computer running our tool. The software engine, which is a number of communicating servlets and beans, handles HTTP requests and responses from and to the client PDA to retrieve information.

One of the advantages of the system is the optional use of mobile phones and the GSM network to connect the PDA to the Internet which gives it an ultimate wireless power. After the connection is established any database can be accessible by the implemented server. The use of Bluetooth technology makes the connection between the mobile and the PDA very convenient to the user since no line-of-sight is required as in IrDA nor an extra jacket that adds up to the weight and size of the PDA. Moreover, this option allows connectivity even though there is no wireless infrastructure around the classrooms.

The Attendance Taking was used as a complete case study. However, the system can be modified easily to fit a various number of applications such as Business Card Keeping, Traffic Violation Report and E-Supermarket.

References

[1] Margaret Hedstrom, "Descriptive Practices for Electronic Records: Deciding What is Essential and Imagining What is Possible," *Archivaria* vol. 36. (Autumn 1993), pp. 53-63.

[2] David Bearman, "Item Level Control and Electronic Recordkeeping", Archives & Museum Informatics, Vol. 10, No.3 1996, pp. 195-245.

[3] C. Safran and H. Goldberg, " Electronic patient records and the impact of the Interent," International Journal of Medical Informatics, Vol. 60, No. 2, 2000, pp. 77-83.

[4] Andrea Jones, Flis Henwood, Angie Hart, "Factors facilitating effective use of electronic patient record systems for clinical audit and research in the UK maternity services", Inyernational Journal of Clinical Governance, Vol. 10, No. 2, pp. 126-138.

[5] Van Bemmel JH, van Ginneken AM, van der Lei J. "The computer-based patient record: an essential technology for health care," Technical Report, Rev. ed. Washington DC: National Academy Press, 1997, pp. 21-43.

[6] M. Abraham, Siikaa, K. Joseph "An electronic medical record system for ambulatory care of HIV-infected patients in Kenya," International Journal of Medical Informatics, Vol. 74, No. 5 , 2005, pp. 345-355

[7] Yen-Liang Lee, " Development and deployment of a web-based physical order entry system,", International Journal of Medical Informatics, Vol. 62, No. 1, 2001, pp. 135-142.

[8] National Institute of Health, website:"http://oma.od.nih.gov/ms/records/rkeeping.html"

[9] Florida Department of Education, "Automated Student Attendance Record Keeping System,", Handbook, 2006

[10] Mohsen Attaran, Ilja VanLaar, " Managing the use of school technology: an eight step guide for administrators," Journal of Management Development, 2001 Vol. 20, No. 5, pp. 393 – 401.

[11] H. Deitel, and P. Deitel, " Java How to program,", 6th edition, prentice Hall, 2005.

[12] Sun site, "http://java.sun.com/

[13] A. Al-li and M. AL-Rousan "Java-based remote," IEEE Transacrion on Consumer Electronics, Vol. , No., 2004, pp. 498-504.

[14] M.Chetty, S. Hu, and J. Bennett, "An interactive Java-based educational module in electromagnetic," International Journal of Electrical Engineering Education, Jan 2003.

[15] S. Deftereos1, C. Lambrinoudakis, P. Andriopoulos, D. Farmakis, and A Aessopos, "A Java-based Electronic Healthcare Record Software for Beta-thalassaemia," Journal of Medical Internet Research, Vol. 3, No. 4, 2001.

[16] Compaq/Hewlett Packard Wireless Solutions: Ericsson T68i/T68 and iPAQ 3870/3970 Bluetooth configuration guide, 2004.

[17] B. Woodward, R S. H. Istepanian and C. I. Richards, "Design of a Telemedicine System Using A mobile Telephone," IEEE Transactions on Information Technology, Vol, 5, No. 1, March 2001. pp. 13-15.

[18] P. Bauer, M. Sichitiu, R. Istepanian, and K. Premaratne, "The mobile patient: wireless distributed sensor networks for patient monitoring and care", Proceedings of the Information Technology Applications in Biomedicine-IEEE EMBS International Conference, 2000, pp: 17 -21.

[19] A. T. S. Chen, "WWW + Smart card: towards a mobile health care management system", International Journal of Medical Informatics", Vol., 57, Issue 2-3, July 2000, pp. 127-137.

[20] S. Pattiches, E. Kyriacou, et., "Wireless Telemedicine Systems: An Overview", IEEE Antenna's& Propagation Magazine, Vol. 44, No., April 2002, pp 143-153.

Nokia 30 GSM Connectivity Terminal User's Guide for modem use, 2002 Nokia Corporation, www.noka.com

Commercial Activities through Mobile Phone Distributed Processing Integrated With Mobile Agents

Meisam Hejazinia
Amirkabir University of Technology/Dept. of Computer and IT Engineering, Tehran, Iran
Email: hejazinia@aut.ac.ir

Mohammadreza Razzazi
Amirkabir University of Technology/Dept. of Computer and IT Engineering, Tehran, Iran
Email: Razzazi@aut.ac.ir

***Abstract*— The number of mobile phones is ten times more than the number of PCs today. They are ubiquitous, and provide the same processing power and memory capacity as PCs in early 90s. The mobile platform is a huge opportunity for future commercial transactions, and also many services could be conducted on them. The main challenge of this platform that hinders the development of commercial applications over it is heterogeneity in both wireless technology and the underlying hardware and software platform. Although J2ME facilitates developing application over this platform, developing application is still complex. Mobile phones should be equipped with a proper middleware that not only provides flexibility, and optimization, and usability, but also facilitates common modules for conducting commerce over them. By having such a middleware, traditional commercial applications could be redefined for this new platform. On the other hand different surveys conducted on mobile agent issues could not be applied on mobile phones due to their specific characteristics. We analyzed the mobile phone platform based commercial applications requirements together with mobile phone based mobile agents and proposed an architecture style to fill the gap, and excess the development of commercial applications integrated with mobile agents on mobile phones. We also described the existing challenges on mobile phones and the opportunities that mobile agents can provide with this new commercial activity enabled middleware on mobile phones. Different challenges are described. We provide solutions to these challenges in the form of software system including the architecture. By using the underlying society network, we will show how commercial services could be provided. Our architecture is based on mobile phone and the previous work on developing P2P computer supported collaborative work. Our contribution is providing the required mechanism to have mobile commercial applications Integrated with mobile agents on mobile phones. Our approach is a new break through to the new world of ubiquitous commercial services provided by mobile applications and mobile agents on mobile phones.**

***Index Terms* – mobile phone, m-finance, m-retailing, J2ME, architecture, peer-to-peer, RMI, mobile agent, grid computing.**

I. INTRODUCTION

The number of mobile phone devices has already exceeded the number of personal computers. 4.2 billion People worldwide have mobile phones. Mobile phones have the same capability in terms of process power, and memory size as the personal computers of early 90s, yet there are not enough applications developed over this platform [2].

Current applications developed over mobile phones, follow the following architecture: The interactive mode between mobile information device and server adopts the multi-layer architecture. Considering the full sharing of enterprise information and the future extension (based on J2EE), the system adopts J2ME+J2EE architecture. It develops the server program under J2EE platform and the server end uses Servlet, JSP and EJB, etc, and connects JDBC to the backend database. As a client end, the application program on the mobile phone end is composed of various user interfaces. Additionally, some frequently used and simple datasheets with smaller data size are stored in the client database, which are connected with Web server through wireless network [8].

What we propose is, to change the perspective, and look at mobile phones as servers. Most of the work done on mobile phones until now was looking at mobile phone as terminals. The current infrastructure is used as a backbone, and mobile phones are added to this backbone. This idea is dominant in both heterogeneous wireless networks and hardware platforms, and it makes using mobile phones with limited capabilities, which always depend on the previous network and hardware/software platforms [1].

There are driving factors for using mobile phones in commerce: first, there are enormous mobile users around the world, second, most mobile phones at least support SMS and Bluetooth. In addition, there are some opportunities for 3G networks. Mobile phones are the best devises which could be used for personalization purposes.

On the other hand mobile agents are used in many applications today. We have mobile agents in grid environments, which are leveraged in load balancing [12]. In a grid environment [14], the concept of mobile agents, applied with web services, is leveraged to integrated grid

doi:10.4304/jetwi.2.3.182-190

service systems. Mobile agents in a grid environment are also used in migration of services.

Mobile phones have the same characteristics of grid components, which make them different from distributed systems. Distributed systems are under the same domain policy. On the other hand, grids have different domains, with different policies. The same is true for mobile phones. They are personalized and full of personalized data. Management of this environment could not be conducted based on a centralized manager. Each mobile phone could have different policies, so it needs the same concept of virtual organizations like what we have in grids.

This mobile phone platform has some limitations, in terms of power consumption, supporting network, processing power, memory size, and input/output capacity, which should be taken into consideration when designing applications for this specific platform. There are not so many applications developed over this platform, and the main barrier was complexity. On the other hand as we previously mentioned the number of mobile phones is three times more than the number of PCs. Thus, mobile phones together could provide a huge processing power, and memory capacity. In order to leverage these capacities, distributed mechanisms should be developed. These mechanisms could not be the same as what we have in a grid environment. Grid computing has a high processing power, and a high memory capacity, but here mobile phones do not have these facilities. Moreover, input and screen capabilities are heterogeneous.

While grids consist of heterogeneous devices, this challenge becomes more severe on mobile phones. We not only have device heterogeneity, but we also have network heterogeneity, that should be handled. Mobile phones mostly support Bluetooth, and some support WLAN. So we do not have a seamless network like grids. Since mobile network is heterogeneous they are imposed by high network failure.

Mobile agents are autonomous processes, which collaborate with each other, and change the dynamic environment. Mobile agents are used in resource management, workflow management, and performance optimization in grids [12]. Mobile agents have four characteristics. Firstly, they have autonomy and computing capabilities such as learning and reasoning which are determined by the agent itself. Secondly, they are adaptable, which means they adapt their behavior according to the dynamic environment. Thirdly, they have interactions, which mean they affect other agents, and ask them to do actions via messages. Fourthly, agents are interoperatable, which means they can perform their own jobs and cooperate with other agents to achieve their goals.

Mobile agents improve grids' performance, by grid resource learning and determination of interactive agents. The main agents' advantage is that they provide flexibility while they are autonomous. Different approaches for having mobile agents on mobile grids are proposed [12][14]. These approaches mainly use a block called agent management system, which monitors and schedules agents, and provides them with management mechanisms such as create and delete. These approaches also use a hierarchical directory facilitator, which act like yellow pages for agents. These approaches could not be used on mobile phone platforms, since they have different characteristics from grids, as we mentioned previously.

Any software system, on mobile phones, for their unique characteristics, should have different nonfunctional requirements. This system should be lightweight and optimized for a lower memory capacity and a lower processing power, and battery power that are the most critical attribute and it must be preserved. Since mobile phones are heterogeneous, adaptability is a very important nonfunctional requirement. Flexibility is also desired for systems on mobile phones. These entire nonfunctional requirements are considered in our design.

By a quick look at these non-functional requirements, we find them contrasting, but if we look into details we find out that this problem is the same as other optimization problems. So we should change our perspective and look globally and as a result, the problem could be solved, if we just provide flexibility in terms of distributed processing, and provide optimization locally for CPU and memory usage, and globally by reducing the number of messages transferred between mobile phones, and roam between different choices of networks available. For robustness, we need to have good exception handling, and proper protocols for distribution handling.

Because mobile phone devices lack standard application platforms and operating systems, it brings the application programs developed for mobile communication devices, higher risks of interoperability. To solve this problem, J2ME provides a proper software layer, but still it is complex, and as a result of that there are not many commercial mobile applications over it. The second problem is that still network programming is complex, especially for Bluetooth, with the odd concept of master and slave. The third problem is the lack of RMI support; you cannot invoke a method of another object on another mobile phone, which is the key point in commercial applications development.

J2ME adopts the modular architecture composed of JAVA virtual machine layer, configuration layer, profile layer and MIDP layer to meet the demand of developers. J2ME consists of four layers, the first layer is JVM, which is customized for the host specific J2ME configuration, while applications rarely see this layer. The second layer over JVM, is the configuration layer, which defines the function of JVM and smallest set of Java Class library available on specific class devices. Developers can support these functional units and libraries available on all devices belonging to a certain specific class.

J2ME has two configurations defined by Java Specification: CLDC (Connected Limited Device Configuration) and CDC (Connected Device Configuration). The first one aims at the limited devices and the size of memory is often between 128KB and 512KB. The second aimed at the devices with higher

configuration and these devices may have 32-bit or 64-bit processors, and at least have the storage space of 512K. CDC uses virtual machine called CVM. While the system based on CLDC uses virtual machine KVM. Users rarely see this layer but it is very important for the realization of profile. The third layer over Configuration is the Profile layer, which defines the smallest set of API available on specific series devices. MIDP is the mobile information device profile layer. It is a Java API set, which mainly deals with the problems, such as user interface, persistent storage, networking, and etc [8].

The implication [3] shows us these complexities are common in most commercial applications over mobile phones; also these complexities could not be solved by normal developers. There are two main challenges for mobile phone platforms: first, is the lack of a suitable business model, and second, is the lack of a suitable middleware for supporting these key issues. These challenges made the mobile platform unsuitable for development, and we cannot see enough commercial applications purely on mobile phones.

The solution for these problems we proposed is an architecture style together with software mechanisms which makes this platform suitable for commercial application development integrated with agents. A mobile platform is a huge opportunity for future commercial transactions, but this platform should be equipped with a proper middleware, and then traditional commercial applications could be redefined for this new platform.

The paper is organized as follows: Section 2, describes different proposed commercial scenarios that could be redefined for this new perspective of mobile phone commerce, also it discusses scenarios and requirement of mobile agents on mobile phones which could be integrated with this commercial scenarios. Section 3 provides a definition for society network and explains why it is suitable for mobile platforms, it also describes different network infrastructures, for developing commercial applications on mobile phones; and proposes a solution for heterogeneity. Section 4, provides proposes a middleware architecture style for commercial application development over mobile phones together with our solution for mobile agents on mobile phone and provides details on how differently the challenges should be overcome, and Section 5, provides the concept of process granularity of commercial applications over mobile phones it also discusses the evaluation issue related to our proposed solution, and section 6, concludes the paper with highlighting our contribution.

II. Commercial Scenarios

Let us start with a scenario; Suppose John has an application on his mobile phone recommending him to choose the portfolio for the stock market. This application is The Intelligent Risk Management application that not only needs the global risk information, which is real-time and is in different places, but also it needs complex calculations. His mobile phone has some limitations in the dimension of the processing power, and memory size. But John is with his friends, and his friends are located in different social networks. John cannot open his laptop and wait for its start up and then connects to the internet, while he is in the street. Thus, if he could use these social networks and his friends let him use the unused mobile phone processing power, memory, network, and battery capacity, he is able to run his application, by just distributing it over his friends' mobile phones. Each mobile phone in this scenario has its own tasks that is locally initiated or initiated by other mobile phones in the same social network with it. In this way, life becomes easy.

This scenario has one important requirement, which is the underlying middleware. This middleware should provide security, and support RMI for invoking methods on other mobile phones. It should support protocols to distribute tasks; it should break a task into parts and assign it to other mobile phones. In this way, mobile phones together will be converted to a super computer. Moreover, it will transparently handle the job that is assigned by a special application on the software layer.

Mobile phones are on most of the time, while PCs are sometimes off. Mobile phones are most of the time in the proximity of each other in a social network, but PCs are connected to each other through the network infrastructure. The main point is that, most mobile phones do not support WLAN and 3G due to being expensive for people. Thus, it is not easy to connect to PCs to assign jobs to. With these specifications, the solution is to use other mobile phones' processing power and memory capacity, which are always free, and wasted.

This scenario could be extended to scenarios for portfolio optimization, mobile payment, and broader to mobile retailing [4].

The second scenario is about mobile retailing [3]. We could have different people having different products, with their specifications and prices, there are also different people having interest in buying products. Then by broadcasting the request over mobile phones, using experience of people with different credits, the person could find proper retailer to buy the product. These scenarios could be extended by personalization, and location based service, which are the unique attributes of the mobile phone platform. The whole scenario could be done during a day, when a person is going to work, during the work hours and even when he is coming back home. Even we could go further and have deals and contracts signed over mobile phones, and this entire scenario is done by simple data transfer over the underlying middleware, which is recommended by our paper.

The third scenario is about credit financing [9]. Effective information sharing, risk mitigation and cost reduction are the most critical challenges to encourage financial institutions to finance individuals and fairs. According to an extensive study, underlying reasons included for problems in financing is that there is an information asymmetry between financial institution and individuals, high risk scoring of SMB induced by lacking good credit history and qualified deposit, high transaction cost to profit ratio. Financing requests arising from various phases differing in the degree of risks. If financers

could have comprehensive information about the purchase orders, account receivables, invoices, and individual's social network, then credit financing could be done easily. Many people are not engaged in e-commerce transactions, so information from e-commerce systems is incomplete. Mobile phones have a big advantage that is portability; they are with you anywhere and anytime. Thus, their simplicity makes more people to use them in order to do their commercial transactions, if there are suitable commercial applications developed over them. In this case, the information related to people's transaction's history, could be comprehensively stored over their mobile phones. There is also the other more important possibility, which is using people's social network. Mobile phones contain people's calls duration, and number, so with this, the financer could use this information with the user permission, to know people's credits related to the social network. Risk control is the core of any financial services and business models, and our middleware could provide the risk control. Our model provides a platform to develop innovative financial services to address the issue. Now let us discuss different scenarios which mobile agents could be integrated in this platform.

Grids are used for scientific, financial, commercial application which needs high capacity, and large processing power. But such applications could not be used on mobile phones. Mobile phones are portable and personalized, that is the key for providing mobile agents on mobile phones.

The first question is why should we have mobile agents on mobile phones, while we could have static services? The second key to answer this question is the battery power, which is critical in mobile phones. If we have a consulting application, which will guide you to a decision, through answering some questions, we have two approaches to handle it. First, we could use services on mobile phones, which are static. Then, you may have 10 questions, and after answering each one, you may ask different kinds of questions according to the answers. This means the next question depends on the answer of the previous question. This is the same as the 20 questions game. If we use this approach of static services, 20 messages are sent and received by the provider and consumer of service. This leads to a high battery usage, which is not desired on mobile phones.

The solution to the previous scenario is mobile agents. We could describe each mobile agent as an autonomous process; they are simple codes, which could be executed in a container. These codes have logic; we call that logic the decision graph (like decision tree). By sending a consultant agent to the consumer, our mobile phone does not need to send and receive 20 messages; instead it only sends 5 messages and which means 75% save in the battery usage of mobile phones.

The supporting scenario to use mobile agents on mobile phones is not limited to a simple consulting. However, the consulting scenario could be defined more broadly. For example, you may be a financial consultant, and want to recommend your customer which of eight options of stocks to use. In this scenario, you do not want to let the customer know your logic of decision making, since it is your core competency.

We could have other scenarios. For example, you are a seller of eight brands that are for one type of product. What you do is you go to a person, and ask them the requirements, and suggest which product to choose. All the attributes of the previous scenario are the same and only the context has changed. We could have many different scenarios related to doctors and travel advisors. Anything that contains decision making and choosing between options is included in this scenario, and could be provided by mobile agents on mobile phones. Agents should not always interact with people themselves. People could have their own agents on their mobile phones. In this way, the agent who migrates to their mobile phones will communicate, with the local agent, and this will make communication automatic. The main advantage that supports this scenario only on mobile phones is that mobile phones are personal. People put their personal profile on their mobile phone, and their local agent could use this personal data to negotiate and communicate to answer question of mobile agents, which act as advisors.

In our scenario, communication means exchanging the information that the mobile agent needs. Perceptions and reasoning is done by getting the required information from the local agent. Using agents on mobile phones provides advantages of load balancing. In this way agents could roam on different mobile phones on the society network, which we will explain in the next section. Using agents on mobile phone will reduce the battery usage, which is very significant in mobile phone networks. Let us look into what a society network is. In section 4 we will discuss thoroughly the advantages of providing mobile agents on mobile phones.

III. Network Infrastructure

There are four options available for providing a middleware, over mobile phones: Bluetooth, SMS, WLAN, and GPRS. GPRS is based on GSM. In the middle of these two technologies is the improved version of GSM, which cellular operators enhance their systems, to improve to GPRS, and SMS is built over GSM, it is simple, fast highly flexible, scalable, wide spread and user friendly. The specification of each of these infrastructures could be seen in figure 1.

To make decisions over these infrastructures, Bluetooth shows up, for it is free but has some

TABLE I.
NETWORK MEDIUM COMPARISION

	Bluetooth	WLAN	GPRS
Bandwidth	1 Mbps	11Mbps	115-117 kbps
Power	1-10mw	50-70mw	200-800mw
Range	10-100m	100-200m	1KM
Cost	None	Low	High
Frequency	2.4GHz	2.4GHz	900/1800MHz

shortcomings upon its limited range. Bluetooth and WLAN support the concepts of same time and same place. At first sight, we may think that this would not be a suitable infrastructure to conduct commerce. But when we look thoroughly, we find out that there is another thing, which is society network. Society network means people we meet every day, in buses or bank queues, waiting for our flights. We are not staying in a place day and night, and we are moving, and this is a great opportunity for providing exchange with many people. Although we meet and talk with limited people, our device is not limited like us. It could automatically interact with many people we may not see, but they are in our proximity, and it could match profiles, use their processing power, or memory. Our device could even trigger and guide us to the right person. With this mentality, Bluetooth is not limited anymore, while we are not Robinson Crusoe who don't have any people around us in an isolated island, even when we are in an island there are many people around us, and this means a huge opportunity for computing and commerce. The main advantage of Bluetooth is that, we are not controlled by central agents anymore, there is no operator, and we could conduct commerce without the middleman (operators).

After that we have SMS, which is accessible from anywhere, but has a shortcoming in terms of cost and middleman controls and regulations. The third suitable choice is WLAN which is also free and suitable, with the ability to connect to the internet, and the fourth one is GPRS, due to having a high cost of transfer; it makes it difficult for a normal application to communicate.

To overcome the shortcomings of each of these options there is a solution, which is a software layer over them for roaming [6] transparently, so that the application over them wouldn't be notified that the underlying layer has changed the network medium. Thus, the first underlying layer searches for the mobile phones in the proximity to check whether they have the special service. If they do not, the software layer under it does the roaming and goes to WLAN or SMS according to the user preference between cost and time. In addition, if it is not found, then automatically it connects via GPRS to another mobile phone, or sends an SMS to a default mobile phone that he knows it contains the required service.

IV. Architecture for Conducting Commerce over Mobile Phones

So far, we have discussed main issues for the enhancement of commerce on mobile phones. The main challenge today for developing applications over mobile phones is that, there are not enough software infrastructures, for distributed processing, specially RMI support, over mobile phones, fulfilling four qualities, usefulness, performance, flexibility, and robustness simultaneously. For solving this problem you can see our proposed architecture in figure 2. We are going to describe the components.

This Architecture is an enhancement of the previous architecture of Peer2me [7] framework, which was developed over mobile phones for collaborative work. We enhanced it so that it could support the distributed processing and commercial basic functionalities required for conducting commerce over mobile phones. The Domain layer contains an abstraction of concepts for network modules, it contains, node, group classes.

The Network Module layer is an implementation of network specific classes and methods. The Network Interface layer separates the implementation layer from the application layer, to decouple application from the underlying network implementation.

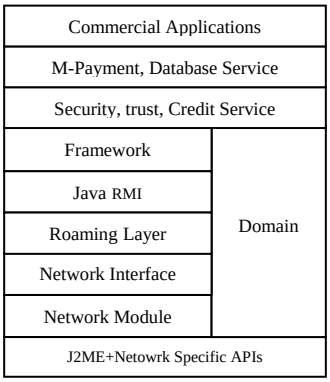


Figure 1. Proposed Architecture for mobile phone commerce

The Roaming layers, provide transparency for applications, so if the user was moving, and he exits the coverage of one network medium, then automatically this layer switches the network medium, according to the user's preference, without interrupting the application.

By this architecture, we have different peers for supporting commerce, and each peer provides different services. We should have bank peers, the mobile phones which support the storing account information, and do transactions. Also, we could have some directory peers, the mobile phones that provide addresses of the bank peers. We could have authentication peers, which are mobile phones that provide certification of the special buyer peer, or seller peer, and this service could be provided with SMS, for the security attribute.

The Security, trust, Credit service layer are the basics for any commercial applications. Even for payments, in this layer, people rate each other, and specific credits. Trust algorithm could be deployed. For example, it could be a simple rating or a complex one, which incrementally corrects itself by multiplying the credits of different people and chooses between them. The point is that, all these interactions are done using the underlying layer of network, which is decoupled from the network implementation, and even we could define a policy in our Roaming Layer, in which this service is being done by SMS and using a specific database or by Bluetooth, using special hash codes.

There are different types of payment, (1) Petty Cash Payment: The amount of money paid below 2 Euros. (2) Small Payment: The amount of money paid between 2 Euros and 25 Euros. (3) Large Payment: The amount of money paid above 25 Euros, it can be regarded as Acer Payment as well. By this architecture even Large

Payments could be supported. If different peers provide credits for each other, according to the game theory we will not do coalition with everyone. In this way, if and only if we do successful transactions with many people, we get credits from them. If we restrict payments to small payments for each individual, we get credits from. Then, we could have Large Payment if and only if we get many credits from many peers, and this will solve the problem of large payment over mobile phones.

The main advantage of our architecture is that we decoupled network from other higher layers, and one can easily develop network for new network technologies. The most important point is that, one could even have payment without the existence of a middleman like an operator. Instead, you are dependent on the society, for they provide you the credits. By the credit you get from the society you could have commercial transactions using your mobile phone.

We propose also database services, in which we recommend new developing concepts of micro databases that data are distributed over different peers. Each peer, itself does optimization on itself according to its policy, and utilization function and what information to put on the current device. Although the general recommendation policy, as a default could be developed based on security, trust, and credit service layer, still each mobile phone is autonomous, and could decide whether to put some data on its storage or not. Each mobile phone would provide database services according to their own databases (e.g. providing free experience about a specific product).

To support this architecture, we need to have the unique address of each mobile phone, which could be the phone number. Also for providing a more dynamic atmosphere, we could provide roaming on the upper layer, which means if someone changed their SIMcard, then we could have some peers, for providing the directory service, and this service could be provided on the roaming layer. Now it is time to discuss about how to integrate mobile agents into this platform.

Mobile agents in our software system contain two parts: the first part is the logic of the agent. This part could be roamed in society network; it could be broadcasted to the other mobile phones in approximation as the service provider is walking. Moreover, mobile phones, who wish to advertise this service, could have this part of mobile agent on their mobile phones to advertise it for people who want it. The second part of mobile agents is the container that is built in the middleware layer; each logic of the agent could be executed using this container. Each mobile phone has a local agent to negotiate with other mobile agents. To communicate between agents, we need some protocols. We could have different protocols and conventions. For each convention we have a CID that is defined as metadata of each agent. To communicate through a special convention we have a convention service. Convention could be loaded from the mobile phone that transfers the agent, or it could be the response of a request of convention from other mobile phones that inputs CID.

Each mobile phone has a special policy. Providing data from the profile is done by the local agent according to that policy. Also this policy defines whether a person is willing to host special mobile agents or not. Each mobile phone has a queuing service. Mobile phones put the incoming preferred agent in the queue. If the person is willing to host a mobile agent, the agent stands in the queue. How to put agent in queues is defined in local policy of user. Each agent has a source address, which is defined as UID. Additionally, it has a unified identifier which contains the code of the agent, which is AID concatenated with UID. Each mobile phone could have a cache and put the favorite agent's identifier in it, and ask from the agent in the future, when they will want it again. Each agent has a specific life time, and it will die upon the finishing time. This is useful for agents who may update. Each agent has a specific credit, which is calculated using people's rate put on it. This credit is updated when the agent serves a specific person, and it is calculated by the average rate, people have so far put on them. This credit could be used by the person, whom this agent is offered to. A person could define a local policy so that, it would not accept agents with a rate lower than 0.6. This policy is executed by the door keeper module in the mobile phone. Thus, the credit and rating of each mobile phone are done through a number between zero and one.

Each agent could have a cost. Any time a person wants to use it, could pay for the agent's use by m-payment. They could advertise an agent, so it could earn money by that. Each time a person uses a mobile agent, contacts to the center to pay. The center could know who has advertised this service and pays for the advertisement in the same model as the model per click on search engines is paid. One user may set special policies relative to the lifetime, cost, and credit of the agent to let it in. When agents are hosted on a mobile phone, the mobile phone could do three things: it could prevent agents to come in, it could advertise agents, or it could use agents, so agents will be put in the queue. User could define special policies to kill the agents, for example it could be by time, credit or other policies.

The main advantage of this model is that, if service that one agent has proposed, is not valuable, it will be killed, and so we wouldn't have congestion in broadcasting an agent. User may also define policies to broadcast this agent to an only person who is their friend.

Each agent could have a special policy to prevent people to hack its logic. For example, it could be defined to not roam over more than 5 levels, or to not let local agents to communicate more than a certain time. In addition, different pricing policies could be adopted for agents inside it. This pricing policy could be used by the pricing container on the host mobile phone. Each agent could have a special logic for learning. Learning of an agent could be divided into two parts: the first one is local learning, in which logic will change according to the user's input, in other word tree of choices changes, according to user input, by agent. Secondly, we have central learning; in this case, agent will send back the

feedback it receives to the UID, who creates it. In this learning, the creator will change logic, and provide a new version of the mobile agent. This feedback provided by the agent to the creator could be prevented by the mobile phone's door keeper according to the set policy.

The convention, which defines how local agents speak to mobile agents could also be cached locally on mobile phones according to the ordinary convention is used. All pervious activities are conducted in the middleware layer. We discussed cache previously in this section. But the point is that capacity of cache could be configured according to the local policy. This functionality will help us provide adaptability on mobile phones. The point we should consider is that agents should be lightweight that is very important. So consultation over a limited number of choices should not be implemented in the agent any more. If a person wants to consult someone on more than one choice, they could have different agents hosted on their mobile phone. In this way, it will preserve the result of consultation with each agent, and compare them with each other. Thus, he, himself, could choose between them. This decision is done, while processing on a single machine consumes much less than the network battery power. So the user may have local applications, or local agents, which aggregate the result of a number of agents' consultation, and will provide users with the result. This lightweight agent could also be transferred over a limited number of SMSs.

What we discussed about method of agent distribution was all push models, but if someone wants to pull an agent for the first time what should they do? The search service is not a big deal for grids, but when it comes to mobile phone platforms, it is really challenging, while mobile phones do not have enough processing power, and they do not access the huge metadata related to the agents and conventions. Even if we want to have a simple search over mobile phones, the input capability is very limited, and it makes difficult to search over them. But what is the solution for mobile phone? If we cannot have search, then there will not be any data grid over mobile phones.

We presume we should use the PC platform to provide search, but on mobile phones, each agent has a specific AID concatenated with UID and each conventions has specific a CID as we explained before. We leverage the PC platform to find these identifiers. Since agents and conventions we are searching for could be repeated over time, then simple map of domain, especially ontology these identifiers, could solve our problem. This mapping is between metadata to identifiers by using hash functions. If a person does not have enough time to go to PC, what should they do? We could have two solutions: firstly, a centralized service, which you can contribute by SMS, then that center, there could be an automatic machine, which according to your history of preferences helps you, or it could be a person behind SMS system, or you can use the call center to find out identifiers. The point is that using web applications configured for mobile phones, is not desirable on mobile phones, since it has limited input abilities, and a small screen, so we recommend using our two solutions. Figure 2, shows the architecture of our software system to provide mobile agents on mobile phones, the relations between components is explained in detail in the text.

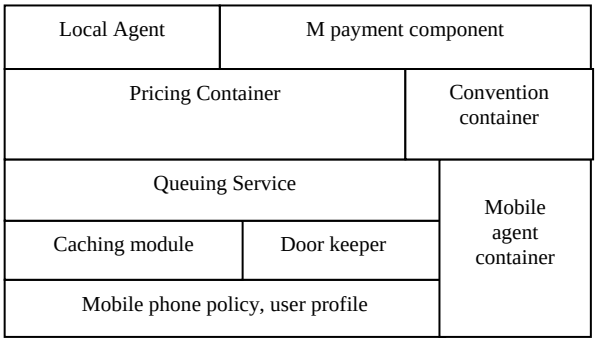


Figure 2. Software system proposed for providing mobile agents on mobile phones

Our proposed software system helps load balancing, which is very critical in mobile phones. While unlike servers, they could not handle a high number of requests for a limited memory capacity and processing power. In agent based services, which we propose anyone needs services, will host the agent to serve him/her, and we do not have such load on special mobile phone. The other advantage of having mobile agents on mobile phones like our system is that by sending the credit to the center each time the agent is used, we will find out which service is popular in a special place. It is somehow a marketing tool to find customers and expand markets. Agents could be stateful but they should be lightweight, and being stateful helps them to learn. Our approach does not have a central agent management system, which exists in grid, and we have no single point of failure, of course we could not have them on mobile phones.

V. The Process Granularity Concept

Previously we discussed having two simultaneous non-functional requirements, which were flexibility, and optimization. At first glance, we find it contrasting, because, if we want to have a flexible application, it would have some overhead, which contradicts optimization.

Mobile phones for their capacity constraint need optimized applications. But if we look globally, there are many mobile phones that do not use their infinite capacities.

These specifications, guide us to the concept of process granularity. Previously, the concept of granularity was developed in databases, but today, for this feature of mobile phones, we should inject it to the concept of process(literature). If we decouple data from logic and break the processes to many simpler processes and distribute them, each of these simpler processes also could be broken into pieces and distributed over other mobile phones. After that, we could collect the results, and make the required results in each level.

The main point of this approach is to decouple the workflow of the special commercial application, from each process in every level of breakdown.

With this approach there wouldn't be any limitations on mobile phone platforms. Moreover, we should use some optimization algorithms with the required time, available battery power, and estimated level of breakdown, network, processing, and memory power usage, to find out whether we should do this sub process locally or send it to other mobile phones. There are situations, where we cannot do it locally, while we do not have that method, we should use RMI capabilities that are supported by our proposed middleware.

New protocols should be developed for these concepts that are configured according to the special requirements of this platform. Also economical system should be developed over it, accordingly, and the costs are input of the optimization algorithm.

To evaluate our system, we should first have other systems to compare with. Unfortunately there is no such system to compare with, and also previous works on grid were not suitable for this platform, for unique characteristics of this platform. But we compare the proposed system with nonfunctional requirement of this specific platform.

First being lightweight, is the main nonfunctional requirement, it is supported by our layered architecture. Modularity of proposed system provides flexibility, and adaptability. This modular system, could provide even lighteweighness, which means someone who has a mobile phone with a higher memory capacity and processing, can install more modules, and enjoy optimized applications, and less battery power usage. Our proposed system also provides adaptability, that means the threshold of caches, and other parameters could be set according to special devices requirements.

Our mechanisms are optimized and less computational, and more distribution makes it more suitable for mobile phone platforms. Also our platform is scalable, since our mechanisms are configured for distribution, and do not depend on the central component. The main characteristic as we talked before is flexibility that mobile agents provide the system with. The other characteristic is load balancing, which is also provided by our proposed system, as we explained before.

VI. Conclusion

Mobile phones provide a very good opportunity, for commercial application development, for its ubiquity, which means you can use it real time in any places, accessibility, which means you have the freedom to use them anywhere, anytime, security, SIMcard are smart cards that contain confidential users' information, convenience, for its hand held size, and light weight, and personalization, which means it is not shared between users and could be adjusted for users' needs.

Unfortunately, there is no suitable middleware to support commercial activities purely on mobile phones that addresses the special constraints of limited computing power, limited information display for navigation, limited power and small memory size. Also there is no software system to provide mobile agents on mobile phones. Most previous works were focused on looking at mobile phone as terminal device. We changed our view, and look at mobile phone as server, we defined concept of society network, which could be leveraged to enhance mobile phone mobile agent capability. Our contribution was redefining mechanisms of mobile agents for specific characteristics of mobile phone. Due to unique characteristics of mobile phone, previous solutions were not suitable, and we defined the mechanism ourselves. Our publish agents services was not done globally, like grid, but we used locally publishing, we used lightweight agent concept. Our approach was adaptable, and provides services according to local policies. We used caches, to improve performance.

We also proposed an architecture style for the required middleware, with the goal of usefulness, performance, robustness and flexibility. Additionally, we introduced the concept of process granularity over this new platform that could support scenarios we described in section 2. Our architecture provides flexibility, adaptability, and performance for mobile phones. We identified new scenarios for using mobile phone for mobile agents, which was focused on consultancy services. We talked about mechanism needed for communication, which are configured for specific characteristics of mobile phones, and we abstracted protocol from agent, and agent from its container.

There are still challenges related to optimization of algorithms, and economical systems which are properly configured for this platform, and could be the subject of future research. In addition our work could be enhancing by investigating threshold for different mobile phones for cache. Also simulating our framework and investigate how granular each agent should be, could be issued for future investigation.

Acknowledgment

This research was created through a financial grant offered by the Iran Telecommunication Research Center.

References

[1] W. Niu, L. Bai, "Business-based Enterprise M-Commerce Models", IEEE 2008.

[2] D. J. Xu, S.S. Liao, Q. Li, "Combining empirical experimentation and modeling techniques: A design research approach for personalized mobile advertising applications", *Decision Support Systems 44.pp.* 710–724, 2008.

[3] C.K. Georgiadis, S.H. Stergiopoulou," Mobile Commerce Application Development:Implementing Personalized Services", *7th International Conference on Mobile Business*, IEEE2008.

[4] J.Hu,N.Zhong, "Developing Mining-grid Centric e-Finance Portal", *Proc. of the IEEE/WIC/ACM International Conference*, 2006.

[5] X.Yang, Z.Qiang, X.Zhao, Zh.Ling, "Research on Distributed E-Commerce System Architecture", IEEE 2007.

[6] C.K.Chen, C.W.Chen, C.T. Ko, J.Lee, J.C.Chen, "Mobile Java RMI support over heterogeneous wireless networks: A

case study", *Journal of Parallel and Distributed Computing*, 2008.
[7] A.I.Wang, M.S.Norum, "A Peer-to-Peer Framework for mobile collaboration". *International Symposium of Collaborative Technologies and Systems*, 2007.
[8] X.Zhang ,"Design of mobile electronic commerce system based on J2ME", *proc. International Conference on Electronic Computer Technology, IEEE*. 2009.
[9] W.SUN, J.LI, D.LIU, G.FENG, P.LI, "Business Models and Solution Architectures for SMB Financing in a Supply Chain Ecosystem", *Proc. IEEE International Conference on E-Commerce Technology for Dynamic E-Business*, 2004.
[10] X.Wang,F.Liu, R.Yang, F.Xie, "Social Computing-Based Trust Establishment in E-Commerce", *Proc. International Conference on Internet Computing in Science and Engineering, IEEE*, 2008.
[11] L.Yunhong, L.Siwen, "Research on Mobile Payment in the E-Commerce", *Proc. International Conference on Management of e-Commerce and e-Government, IEEE*, 2008.
[12] M.Bunruangses, W.Poompattanapong, P. Banyatnoparat, B.Piyatamrong , QoS Multi-Agent Applied for Grid Service Management, Proc. ACM, 2004.
[13] T.E. Athanaileas, N.D. Tselikas G.V. Tsoulos, D.I. Kaklamani, "An Agent-Based Framework for Integrating Mobility into Grid Services", Proc.ACM, 2008.
[14] F.Magoules, J.Pan, K.A.Tan, A.Kumar, "Introduction to grid computing", Taylor & Francis Group, 2009.
[15] C. Blundo, E.D.Cistofaro, A Bluetooth-based JXME infrastructure, Proceedings of the 9th International Symposium on Distributed Objects, Middleware, and Applications (DOA'07), LNCS. 2007.

Reuse and Remix: Content Recomposition System based on Automatic Draft Generation

Kosuke Numa, Katsuaki Tanaka, Mina Akaishi and Koichi Hori
Research Center for Advanced Science and Technology, the University of Tokyo, Tokyo, Japan
Email: numa@ai.rcast.u-tokyo.ac.jp

***Abstract*—In this paper, we propose a framework for content circulation. Reusing and remixing contents are keys to expressing activities. In our framework, a system supports decomposing and recomposing by automatic draft content generation. As a field test, we implemented the framework on a support system for a workshop, in which participants created contents based on a format of expression named *photo-attached acrostics*. Through observation of the practice, we concluded that our framework could help content recomposition. We think our framework is applicable to various manners of expressions like Web content creation.**

***Index Terms*—content recomposition, creative activity support, content circulation, automatic content generation, participatory workshop**

I. Introduction

We live in a world of rapidly increasing information and face difficulties in managing such information. We need a way to find and select wanted information, to arrange them, to think over them, and to create and publish new information.

Information retrieval and recommendation technologies will help to find and select information, and arrangement and consideration will be supported by content management and visualization techniques. Creation of information, however, is difficult to be directly supported. Truly "new" information is rarely produced but usually new combinations of information are devised. In this research, we propose a framework for recomposition of stored contents from a creativity support research perspective.

In our framework, a system shows draft contents, which are automatically generated by remixing the user's and the others contents, when a user produces a new content. A user finishes her content by selecting and modifying a draft content. Through such process, we aim to develop an iteratively growing loop of expressions. In the loop, others contents are taken in a user's newly produced content, and the content are used in others contents again.

Contributions of this research are as follows: For database application field, presenting a new model for using/reusing others contents. For expert system field, presenting a new model for leading users to better results even when there are no answers.

This paper is organized as follows: After describing related works, we present our proposed framework in section 3. We implemented the framework on a support system for a workshop. We introduce the workshop in section 4 and the developed system in section 5. We show the results and discussions in section 6.

II. Related Works

A. Creativity Support

In the beginnings of 1990s, research area called creativity support was raised. In the area, problems like how computers can support human creative activity and what kind of creative activity can be supported were discussed.

Boden distinguished two sorts of creativity: H-creativity, which indicates historically new idea/concept formation, and P-creativity, psychologically new idea/concept formation in human minds [1]. In our research, we aim P-creativity support rather than H-creativity support. For ordinary people, our target users, what they express — externalization of internal nebulous thoughts — is more important than how they express — surficial originality of expressing techniques.

In psychology field, Guilford made the distinction between convergent and divergent thinking [2]. Our approach emphasizes neither of them specially, but if daring to say, it matches divergent one. One of our aims is to support expressing, which seems to be a convergent process; but widening users' views and unsticking users' stuck thinkings are more important.

Many and many creative methods have been proposed, including KJ method [3] and brainstorming [4], and many systems to help creative methods using computer systems have been developed [5].

B. Automatic Content Generation

Our research aims to stimulate users by presenting draft expressions. It doesn't mean that the system takes a user's place to "create" expressions; it just presents candidates. Users decide to insert the draft into their expression or not, and if so, they select which candidate is added and modify it according to their will. Letting users place the generated candidates into their own content affords them to think deeply about it. Automatic content generation techniques, however, are useful for our purpose.

doi:10.4304/jetwi.2.3.191-196

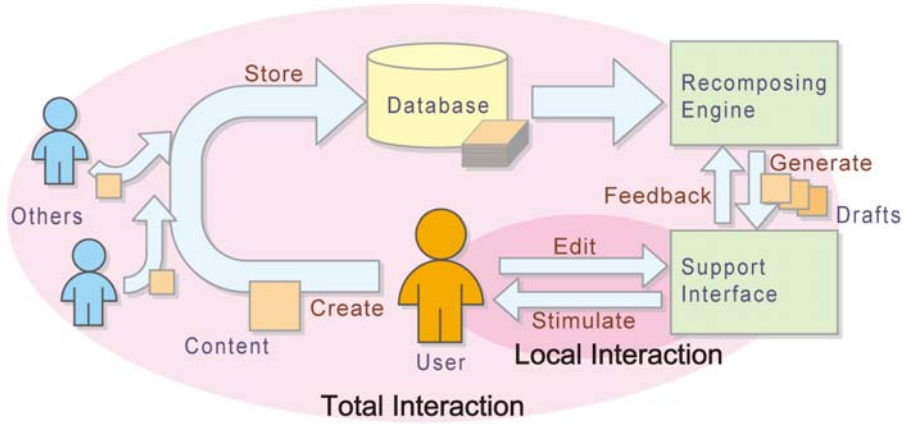


Figure 1. Proposed framework.

Bringsjord and his group developed a system called Brutus1, which generates literary stories [6]. Knowledge bases and grammar rules are programmed in advance, and it generates quite readable and natural stories. When sufficient knowledge and enough rules are provided, machines can generate high-quality unexpected expressions.

AARON programmed by painter Harold Cohen is known as a painting software [7]. AARON generates paintings according to parameters given by Cohen. There is an interesting story: Someone asked him "who is the 'creator' of the paintings?" Cohen claimed that AARON does not paint, but Cohen paints using AARON. This is the very what we emphasize: A system is a tool for creation. An output of the system can be an expression only after evaluated and accepted by the user as her expression. If she is insufficient, she can modify parameters or edit the output, then "create" her work. Here an output is a stimulation for a user.

Multiple document summarization [8, 9] is technically related to our research. We have not implemented these techniques, but these will be helpful.

C. Knowledge Models in Knowledge Management

Our research aims to design a new loop of content circulation. As for knowledge management area, SECI model [10] is widely known. SECI is the abbreviation for Socialization, Externalization, Combination, and Internalization, which are the processes of knowledge cycle. Shneiderman categorized creative activities into following four activities: "collect," "relate," "create," "donate" [11] And Ohmukai *et al.* expanded Shneiderman's model to distinguish information activity layer and communication activity layer [12]. In their model called ICA model — Information and Communication Activities model, two layers of information activities ("collect," "create" and "donate") and communication activities ("relate," "collaborate" and "present") form cycles related to each other.

Hori and his group developed a cycle model which consists of the knowledge *liquidization* and *crystallization* processes [13]. They called decomposition of expressions into units in proper granularity with every possible connection among each as *liquidization*. And as *crystallization*, they called new expression formation from decomposed partial units based on new relationships within the context. Our research is based on this concept [14]. In our proposed framework, a system decomposes and recomposes collected users expressions.

III. PROPOSED FRAMEWORK

Fig. 1 illustrates our proposed framework for content circulation. Contents created by multiple users are stored into the database. The recomposing engine decomposes stored contents and generates draft contents. Here we aim not to create complete contents but to stimulate users. The support interface shows drafts to a user and she edits and finishes her content. These operations spread to the recomposing engine and it shows other drafts.

We show two levels of interaction loops here. Direct and local interaction between users and the support system is shown in editing and stimulating loop. Remixing and reusing stored contents form indirect and total interaction loop.

This model is applicable for various manners of creation and publication of contents. For example, writing process of papers or blog entries include information collection phase and editing phase. Of course authors need to add their own original opinions, but candidate combinations of related information will help their considerations. Format of expression can vary and is not limited to text expressions. While we expect the framework can be applied to any types of contents, we dare focus on text content in this research. Decomposition and recomposition are realized by usual text processing techniques. We wanted to focus not on techniques for implementation but on the content circulation framework itself. For that reason, we held a participatory workshop where participants created their contents and recomposed them into new contents.

IV. PHOTO-ATTACHED ACROSTIC WORKSHOP

We designed and organized a workshop as a field practice [15]. In the workshop, participants create contents based on rules. We designed a new format of expression called *photo-attached acrostics* to highlight the process of decomposing and recomposing. Acrostic is

Figure 2. Example of photo-attached acrostics

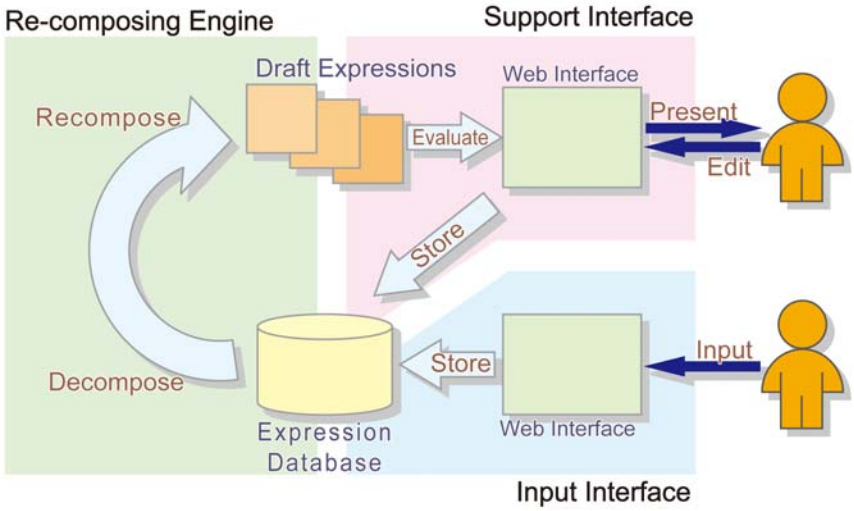


Figure 3. Architecture of developed system

"a poem or other writing in an alphabetic script, in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out another message[1]." We modified it to include pictures for each sentence. Participants take and select photos, write sentences whose first letters match a message given. Here a pair of sentence and photo should correspond and both photos and sentences should be along a theme given. An example of photo-attached acrostic is shown in Fig. 2. The message of the example is "ABCDE."

In the workshop, participants create an acrostic using their own photos at first. Then next, they are divided into groups and collaborate to create new expressions by remixing their expressions. Collaboration with others will raise new context and stimulate participants. In the third step, they create expressions by themselves again, using all pictures used in the former steps. Participants are requested to place others' (partial) expressions in their new expressions. We aim that participants form new opinions/ideas stimulated by others. At the same time, the workshop facilitator shows other new remixed acrostics using the developed information system described below.

V. Support System for the Workshop

The system consists of four parts (Fig. 3): expression database, expression input interface, expression recomposing engine, and expressing support interface. It has the same structure with the framework illustrated in Fig. 1, but is modified to highlight its dataflow.

Users input their works, which are created in manual and analog manner in the workshop. The expressing support interface shows draft expressions, which are generated from the expression recomposing engine (see Fig. 5).

The expression recomposing processes are as follows:

- Decomposition phase
  1) Analyze morphological structures of text.
  2) Calculate term relation weights and term weights. We use term dependency for term relation weights and term attractiveness for term weights [16]. Term dependency $td(t,t')$ from term t to t' is given by:

$$td(t,t') = \frac{sentences(t \cap t')}{sentences(t)} \quad (1)$$

Here $sentences(t)$ indicates the number of sentences in which term t appears, and $sentences(t \cap t')$ is the number of sentences term t and t' appear at the same time.

Term attractiveness $attr(t)$ of term t is a total of incoming term dependencies. T is the set of all appearing terms.

$$attr(t) = \sum_{t' \in T | t' \neq t} td(t,t') \quad (2)$$

[1] Acrostic - Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Acrostic

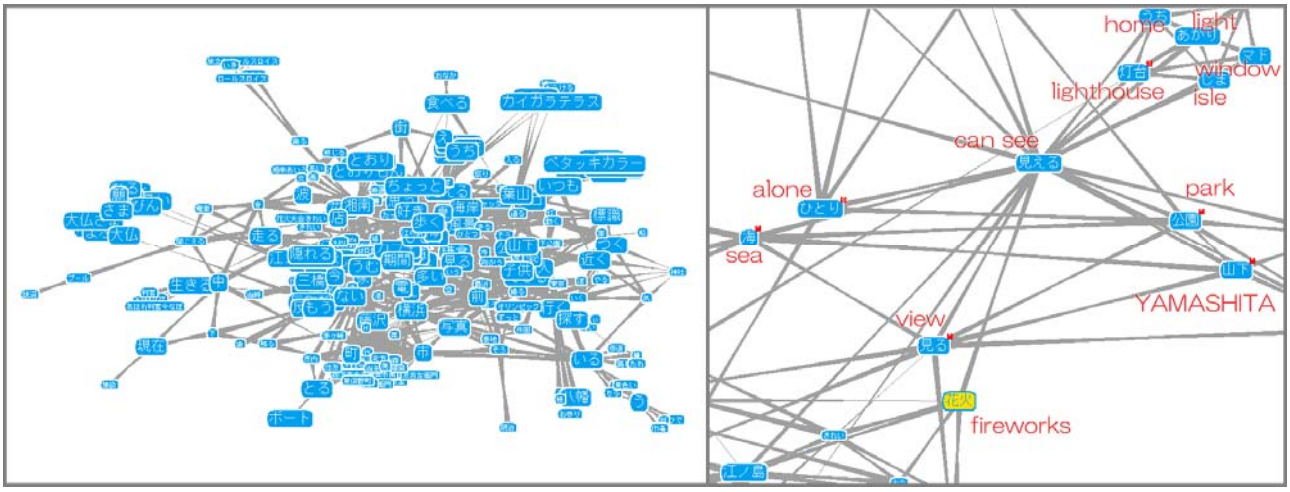


Figure 4. Extracted network in the decomposition phase. Whole structure (left) and partial enlargement (right).

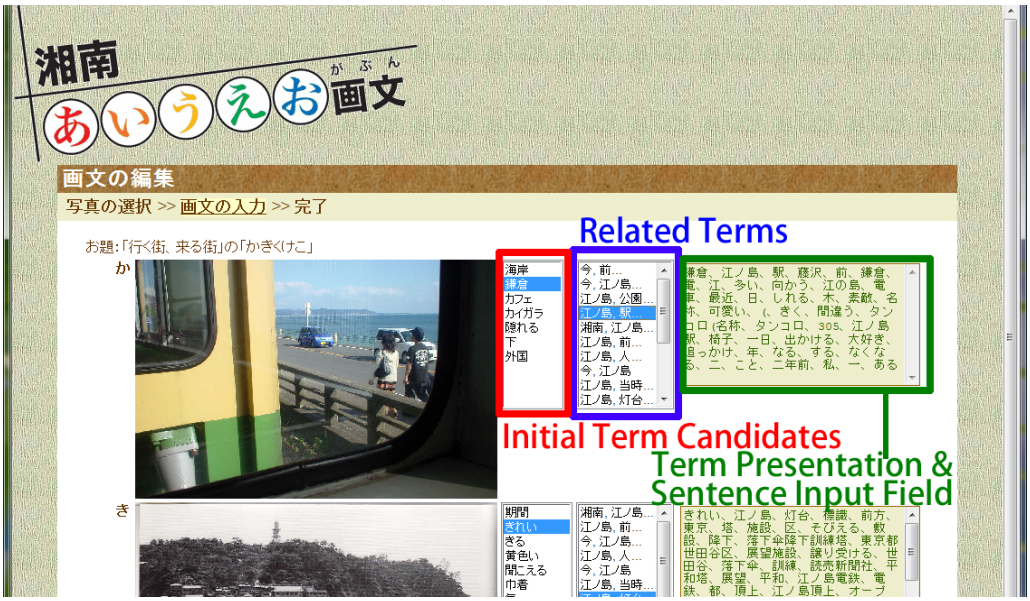


Figure 5. Screen image of photo-attached acrostic creation support interface.

- Recomposition phase
  1) Extract candidate terms according to their initial letters.
  2) Extract photos which include each term in 1.
  3) Evaluate photos.

  We define the weight $w_t(p)$ of a photo p for term t as follows:

$$w_t(p) = \sum_{t' \in T_p | t' \neq t} td(t,t') \cdot attr(t') \quad (3)$$

  For each initial letter, the term candidates, their related terms, and attached photos are structured.

Fig. 4 illustrates an example of word network extracted in the decomposing phase. Sizes of the nodes indicate their weights (td) and the distances between two connected nodes shows their strength of the link ($attr$). This figure just visualizes word relations which are calculated internally in the system; we didn't show users the figure directly.

Based on this network, we generate candidates of expressions and present them in creation support interface (Fig. 5). Choose one candidate of initial terms in the left textbox, then other related terms are presented in middle textbox. Choose one again, a photo combined to selected terms and the rest of related terms are shown. User can overview candidates presented on the system, select one she likes, and edit it.

In the workshop, a facilitator shows semi-automatically generated expressions, which are edited in certain rules like choosing photos with the highest weights or the lowest. With these expressions, we aim to stimulate the participants by machinery generated context. Through this step, we observe the effects of the expression recomposing engine. After the workshop, we asked the participants to try the expressing support interface. We evaluate the interface from this test.

Figure 6. Photo-attached acrostic workshop

VI. Results and Discussions

The theme of our first practice was "Shonan" — the name of a region along a coast in central Japan. We called for participation to the people related to — e.g., living around, working around, or was born around — Shonan area. Through the workshop, participants are expected to discuss together and get new opinions about the area.

The workshop was held at 8th and 16th December 2007 in Fujisawa city, the center of Shonan area, with nine participants. Most of their occupations were related to media activities or media literacy: information media-major students, an elementary school teacher, an art university professor, members of citizens' television at Shonan, and so on. While the youngest was an undergraduate student, a retired person was also included. Three were female, and six were male. The participants were divided into three groups and finally they made 30 photo-attached acrostics from 259 photos. Fig. 6 shows the scenes in the workshop. The analyses of the workshop itself are discussed in the other paper [15], so we discuss the effects of the system in this paper.

In the first step, the facilitator selected photos and terms based on certain rules so that we could observe the effects of the recomposing engine. As a result, a content, which was created by choosing photos and terms with the highest weights in the expression (2) and the expression (3) (shown on the top of each list of candidates in the support interface) happened to have a similar story structure to a participant's one. We aimed to form a different context, but made a similar story. The facilitator, however, could create much more expressions in much less time. The outputs of the system were not always new, but the number of outputs was large enough to stimulate the participants.

As the second step, we asked the participants to try the support interface after the workshop and conducted interviews with them. While positive comments like "I could easily create new acrostics" were heard, a problem was pointed. The system shows candidates for each sentence separately; connecting sentences — making story — is not supported enough.

VII. Conclusion

In this paper, we proposed the framework for content circulation. We implemented the framework on the support system for the workshop. Through the observation, we found that recomposing activities and support framework were pretty effective, but algorithms to generate drafts should be revised.

Generated drafts vary depending on a corpus — "database" in the framework. A system which recomposes all other expressions and another system which utilizes only expressions made by a user herself generate different draft by all means. Without purposes or goals, we cannot decide which implementation is the better. It totally depends on the contexts. In our first workshop, our aim was to widen participants' views, so we designed the workshop program to contain remixing process of others contents and our system also needed to remix other participants' expressions.

Our first trial dealt with a peculiar type of expression. But as we already mentioned, our framework is applicable to other types of expressions. Especially it suits on Web content creation. The Web can be a database from which a system draw others expressions, and can be a place where people present their created contents. Reusing and remixing loop of circulation is natively on the Web. We are planning to develop an application of our framework for blogging.

Acknowledgment

This work has been supported by a grant from the Japan Science & Technology Agency under CREST Project.

References

[1] M. A. Boden, *The Creative Mind: Myths and Mechanisms*, New York: Basic Books, 1991.

[2] J. P. Guilford, *The Nature of Human Intelligence*, New York: McGraw-Hill, 1967.

[3] J. Kawakita, "The KJ method: a scientific approach to problem solving," Technical report, *Kawakita Research Institute*, Tokyo, 1975.

[4] A. F. Osborn, *Applied Imagination: Principles and Procedures of Creative Problem-solving*, Scribner New York, 1957.

[5] J. Munemori and Y. Nagasawa, "Gungen: groupware for a new idea generation support system," in *Information and Software Technology*, Vol. 38, No. 3, pp. 213-220, 1996.
[6] S. Bringsjord and D. A. Ferrucci, *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus a Storytelling Machine*, Lawrence Erlbaum Assoc Inc., 1999.
[7] P. McCorduck, *Aaron's code*, WH Freeman & Co. New York, 1991.
[8] I. Mani and E. Bloedorn, "Multi-document summarization by graph search and matching," in *Proceedings of Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pp. 622-628, 1997.
[9] R. Barzilay, K. R. McKeown, and Michael Elhadad, "Information fusion in the context of multi-document summarization," in *Proceedings of the 37th Association for Computational Linguistics*, pp. 550-557, 1999.
[10] I. Nonaka and H. Takeuchi, *The Knowledge Creating Company*, Oxford University Press, 1995.
[11] B. Shneiderman, *Leonardo's Laptop: Human Needs and the New Computing Technologies*, MIT Press, 2002.
[12] I. Ohmukai, H. Takeda, M. Hamasaki, K. Numa, and S. Adachi, "Metadata-driven personal knowledge publishing," in *Proceedings of 3rd International Semantic Web Conference 2004*, pp. 591-604, 2004.
[13] K. Hori, K. Nakakoji, Y. Yamamoto, and J. Ostwald, "Organic perspectives of knowledge management: Knowledge evolution through a cycle of knowledge liquidization and crystallization," In *Journal of Universal Computer Science*, Vol. 10, No. 3, 2004.
[14] K. Numa, K. Tanaka, M. Akaishi, and K. Hori, "Activating expression life cycle by automatic draft generation and interactive creation," in *International Workshop on Recommendation and Collaboration (ReColl'08)*, 2008.
[15] K. Numa, K. Toriumi, K. Tanaka, M. Akaishi, and K. Hori, "Participatory Workshop as a Creativity Support System," in *12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2008)*, 2008.
[16] M. Akaishi, K. Satoh, and Y. Tanaka, "An associative information retrieval based on the dependency of term co-occurrence," in *Proceedings of the 7th International Conference on Discovery Science (DS2004)*, pp. 195-206, 2004.

Kosuke Numa received his B.A, M.Inf. degrees from Yokohama National University, Japan in 2004, 2002 and Ph.D. degree in informatics from the Graduate University for Advanced Studies, Tokyo, Japan in 2007 respectively.

Since 2007, he has been a post-doc researcher at Artificial Intelligence Laboratory, Research Center for Advanced Science and Technology (RCAST), the University of Tokyo, Japan. His research interest include applications of AI, Web and mobile technologies in social practices, creativity support systems, and community based information sharing systems.

Dr. Numa is a member of IEICE, IPSJ, JSAI, and DBSJ.

Katsuaki Tanaka received his MS degree in informatics from Waseda University, Japan, in 2000.

He is currently an assistant professor at the University of Tokyo. His research interests include natural language processing and knowledge management systems.

Mr. Tanaka is a member of JSAI and IPSJ.

Mina Akaishi received B.Eng, M.Eng, and Dr.Eng. degrees in electrical engineering from Hokkaido University, Japan, in 1990, 1992, and 1995, respectively.

In 1995, she took work as a associate researcher at Shizuoka University, Japan and then she moved to Hokkaido University, Japan. In this period she developed several Media Database systems and Visualization systems based on the concept of Context. Since 2004, she has been with the University of Tokyo, Japan.She is currently a associate professor with School of Engineering, University of Tokyo.Her current research interest includes AI technology, Information Visualyzation, Media Database and Story Generation. She have been developing a Narrativity based Information Access Framework with the support of JST.

She is a member of IEICE, IPSJ and JSAI.

Koichi Hori received B.Eng, M.Eng, and Dr.Eng. degrees in electronic engineering from the University of Tokyo, Japan, in 1979, 1981, and 1984, respectively.

In 1984, he joined National Institute of Japanese Literature, Japan, where he developed AI systems for literature studies. Since 1988, he has been with the University of Tokyo, Japan. He is currently a professor with School of Engineering, University of Tokyo. From September 1989 to January 1990, he also held a visiting position at University of Compiegne, France. His current research interest includes AI technology for supporting human creative activities, cognitive engineering, and Intelligent CAD systems.

Prof. Hori is a member of IEEE, ACM, IEICE, IPSJ, JSAI, JSSST, and JCSS.

Development of a Recommender System based on Extending Contexts of Content and Personal History

Katsuaki Tanaka
RCAST, the University of Tokyo, Tokyo, Japan
Email: katsuaki@ailab.t.u-tokyo.ac.jp

Koichi Hori
School of Engineering, the University of Tokyo, Tokyo, Japan
Email: hori@computer.org

Masato Yamamoto
NTT DOCOMO, Inc., Tokyo, Japan
Email: yamamotoma@nttdocomo.co.jp

***Abstract*— The flood of information on the Internet makes a person who approaches it without some strong intention feel overwhelmed. One way to redress the balance between a person and the flood is a computer-based recommender system, and many Web sites use such systems. These systems on a Web site work for similar items. However, the field of personal activity is not limited to handling one kind of knowledge or one Web site, but also involves off-line activities in the real world. To handle personal off-line activity, LifeLog was proposed as a method to record it, but the main purpose of LifeLog is to record a personal history. The uses of such a history are still being studied.**

We have developed a recommender system that captures personal context from a history of personal online and off-line activities, treats information on Web sites as a large set of context, and discovers and extends the overlaps of personal activities and Web sites, then recommends information located in the Web sites. The aim of the system is to allow a person to enjoy waves of information again.
The system was implemented as part of the *My Life Assist Service* for mobile phones provided by NTT DOCOMO, Inc. as a field experiment from December 2007 to February 2008.

***Index Terms*—recommender system, mobile phone, context awareness, LifeLog**

I. Introduction

There is too much content on the Internet, so people who consult it can lose sight of their own objective without strong intention. A similar thing can also occur on a Web site. To solve this issue, some Web sites have employed recommender systems such as collaborative filtering [1] to help a user browsing the Web site.

Thus, using a recommender system for computer-based content selection and presentation is a way to strike a balance between content generated by other people and a user [2]. Usually, the target of a recommender system is a single Web site.

Such recommenders are sufficient for a person whose activity is limited to a Web site. However, most human activity occurs in the real world. To capture such human activity, the LifeLog was proposed [3]. The main topic of LifeLog research is how to capture activity. The present focus of study is the reuse of captured records.

We developed a context-based recommender system that derives content recommendations from a user's activities as captured by a mobile phone, as part of the *My Life Assist Service*. The service was a field experiment performed from December 2007 to February 2008 by NTT DOCOMO, Inc. and other collaborative partners as a part of the *Information Grand Voyage Project* promoted by the Ministry of Economy, Trade and Industry in Japan. The service was a platform that included specialized services for content providers and consumers. The service for consumers was named the *Preview channel*.

Typically, a recommender system recommends pages from the Web site where the user has been browsing, so the user's needs would be clear, and the purpose of a recommender system is that a user accepts the recommendation and makes an action such as buying something. Therefore, precision enhancement of the recommended content acceptance is the first goal.

On the other hand, the Preview channel was quite different from other services such as searching or recommending content for a user based on the user's location as identified by the user's mobile phone. The genres of recommendation were not given and the recommended content was always displayed on a phone screen, as shown in Figure 2 and was updated only four

This paper is based on "A Recommender System Based on Context Extending of Content and Personal History," by K. Tanaka, K. Hori, and M. Yamamoto, which appeared in the Proceedings of the International Conference on Information and Communication Systems (ICICS), Amman, Jordan, December 2009.

doi:10.4304/jetwi.2.3.197-203

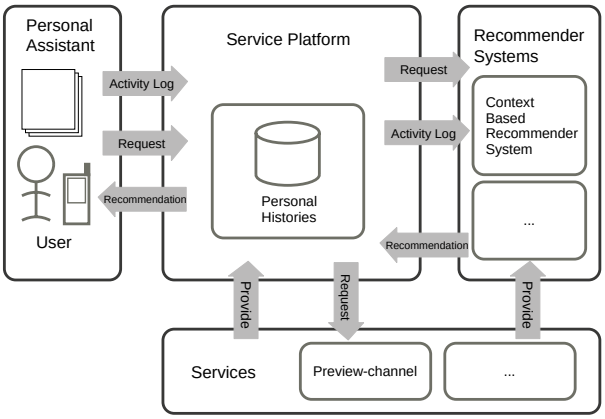


Figure 1. Components of the My Life Assist Service

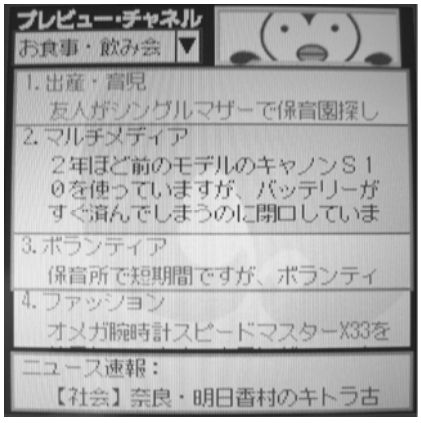


Figure 2. Screenshot of the Preview channel

times a day. Therefore users became bored with precise but unchanging recommendations [4]. The context-based recommender system therefore did not aim for precise recommendations, but recommendations that shifted with the user's interest. We felt that content expresses part of its creator's intentions, and should be located in a context. For example, a promotion for a shop has contexts not only of things that the shop wants to sell but also of the intent in opening the shop, reasons for selling the item, and so on. The context-based recommender system shifted with the user's interests by finding overlaps between the context of a message and the context of the user, and extending them.

II. The My Life Assist Service

A. Components of the My Life Assist Service

The My Life Assist Service consisted of the components shown in Figure 1. The Personal Assistant Application on a mobile phone collected user activities and sent them to the Service Platform. The Recommender System scraped personal history collected from the user's activities, and selected content for recommendations. The Service Platform sent details of requested content to the Recommender System, and sent the answer to the Personal Assistant Application.

The Personal Assistant Application collected locations from the global positioning system (GPS) receiver installed in the mobile phone, content that the user had read on the Preview channel, and texts that the user wrote to other services in the My Life Assist Service, as personal history.

B. Preview Channel

The My Life Assist Service was a platform that included specialized services for content providers and consumers. The Preview channel was a service for consumers designed by NTT DOCOMO, Inc., and the content for the service was provided by *Oshiete! Goo*[1] by NTT Resonant Inc. and *Walkerplus*[2] by Kadokawa Cross Media Co., Ltd.

The Preview channel was provided as an application for NTT DOCOMO's mobile phones. The application (Figure 2) was always displayed on the screen. The application displayed four categories; each category included four brief summaries of content. The user could select a summary and access the relevant content via the phone's Internet access function. The recommender system selected four messages for each of the four categories, for a total of 16 messages.

Because the service always displayed recommended content on the screen, it was considered successful when at least one message attracted the user's attention.

C. Content for Recommendation

Sources of the content were Oshiete! goo and Walkerplus. Oshiete! goo has many Q&A dialogs across many categories; the recommender system used 107 of the categories. Walkerplus consists of *Gourmet Walker*, *Hotel Walker*, and *Wedding Walker*; the recommender system treated each of them as a category. The number of categories that the system used was therefore 110. On the Preview channel, making a recommendation meant selecting four categories from the 110, and choosing items from each of the selected categories.

For example, an item in Walkerplus described the detail of a restaurant. It contained the latitude and longitude of the location. On the other hand, an item in Oshiete! goo is generated when a user of the Oshiete! goo Web site wrote a question and other users answered it. Items did not contain latitudes or longitudes originally, but the service platform added locations to each item that mentioned places or addresses. The recommender system was required to recommend as many items with no locations as with locations.

D. Timings to Update Recommendations

The recommender system was required to update recommendations four times a day. The recommender system can specify the timing of updates. Personal histories are stored in the platform that is updated every three hours.

Recommendations did not change when the personal history did not change because the recommender system

[1] http://oshiete.goo.en.jp/
[2] http://www.walkerplus.com/

Figure 3. Outline of the context-based recommender system

makes recommendations based on personal history. The system should update recommendations at least at three-hourly intervals. There is little flexibility to choose timings to match four updates with three-hour intervals in the daytime. Therefore, the recommender system did not infer timings of updates for each user, but fixed the update times to 8, 11, 15, and 19 hours each day.

E. Design of the Recommender System

The Service platform stored the personal history of each user. The personal history was collected by the Personal Assistant Application and consisted of locations, items that the user had read on the Preview channel, and text items that the user wrote to other services in the My Life Assist Service.

When the My Life Assist Service was designed, it was assumed that most of the personal history would be locations in the real world that were collected automatically every 30 minutes.

We could not make any assumptions about location history, because the history would be quite different for each user, the history items could be grouped for some reason, and we could not find relations within these groups and so on in advance because we had no previous location histories of many persons. Therefore we could not design a user model or create a learning mechanism based on user histories.

We therefore designed a context-based recommender system that did not use generalized statistical models determined from many users' histories, but made personalized models.

III. CONTEXT-BASED RECOMMENDER SYSTEM

The context-based recommender system aimed at nonconverged recommendations to shift with the user's interest.

First, it articulated (see Section III-B) the personal history of activities of the user and the content for the recommendation source. It extracted contexts from them and generated a graph for each personal history and Web page. Then, it found overlaps between the graphs, extended the overlap together with the context of a Web page and recommended content from this extended area.

Several methods have been proposed to provide variation to recommender systems based on attributes of recommended targets or relations in described texts [5]. The feature of the context-based recommender system was that it extended the recommendation area together with the context of the content by inserting two graphs that represented the contexts of the personal history and the content (Figure 3).

A. Inputs to the System

The recommender system used the personal history of activities of each user and the content for recommendation.

1) Personal History: The personal history of each user consisted of two parts: on the Web and in the real world.

The history of Web use consisted of content that the user had read on the Preview channel and texts that the user had written to other services in the My Life Assist Service. The history was expressed as a combination of literal texts and time stamps.

History from the real world consisted of a combination of locations and time stamps obtained by the mobile phone. A location was obtained by built-in GPS receiver or locator in the mobile phone that used radio waves from base stations.

2) Content for Recommendations: Content was recommended to each user in text form. It was desirable that each text, such as a description of a shop, was long enough to allow differentiation and to articulate groups of messages. It was also desirable that each text had corresponding locations in the real world so content could be selected based on location.

3) Converting Personal History to Text: The personal history on the Internet was acquired as text, while the real-world history was acquired as locations in the form of combinations of latitude and longitude. Most of the personal history consisted of location data, because texts were only obtained when the user used the service, while locations were obtained automatically.

Therefore, the system should discover the user's habits from a history of locations. It was presumed that the character of a location was described in terms of the content located in the area; the system converted a location to text from the associated content. In addition, the context-based recommender system did not aim to provide content around the user's location like existing location-based services. The system did not record locations, and did not recommend content based on the locations that the user often visited.

B. Articulation

As mentioned in Section III, the context-based recommender system generated graphs for both the personal history and the content by clustering based on similarities of expressions and linking based on context in the data. This graph generation was called articulation. In this paper, `articulation' means the decomposition of text

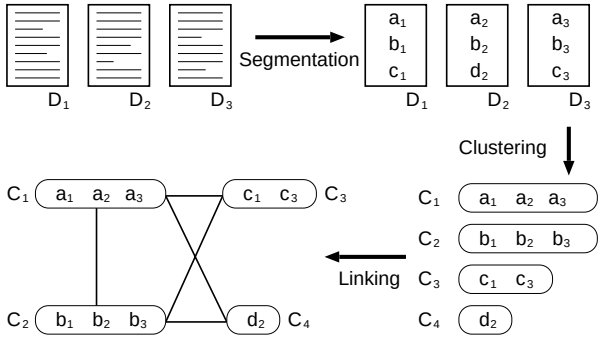


Figure 4. Example of articulation

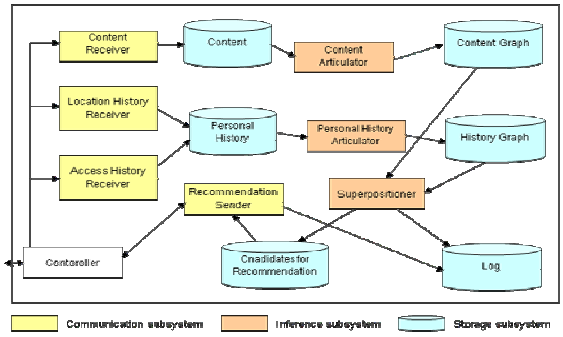


Figure 5. Components of the context-based recommender system

information into fragments by cutting and organizing the information into chunks connected with the writer's context.

In the articulation phase, the recommender system segmented each text of the personal history and the content. Next, the system grouped these text fragments into clusters. Then, it linked clusters that included text fragments from the same text.

In this way, graphs were generated from each personal history and the content. Each graph was organized from clusters that were found by a computer, based on similarities of text expressions and links based on the context of the expressions.

1) Example of Articulation: As shown in Figure 4, text fragments a_n, b_n, c_n, d_n were obtained from text D_1, D_2, D_3:

$D_1 = \{a_1, b_1, c_1\}$
$D_2 = \{a_2, b_2, d_2\}$
$D_3 = \{a_3, b_3, c_3\}$,

and they were clustered as follows:

$C_1 = \{a_1, a_2, a_3\}$
$C_2 = \{b_1, b_2, b_3\}$
$C_3 = \{c_1, c_3\}$
$C_4 = \{d_2\}$.

C_1 and C_2, C_1 and C_3, C_1 and C_4, C_2 and C_3, C_2 and C_4 included text fragments that were generated from the same text, so links were set between them (Figure 4). These links indicated contexts in original texts.

C. Superposition

The graphs for each user's personal history and the content were generated in the articulation phase. In the superposition phase, the recommender system looked for an overlap between the graph of a user's personal history and the graph of the content, extended the overlap together with links in the graph of the content, and selected items to recommend from the extended area.

Nodes of the graphs could be represented as word vectors because the graph was a group of text fragments. First, the system searched for similar nodes based on similarities between nodes represented as word vectors. The nodes found in this step in the content graph were named the primary nodes. Items in the primary nodes corresponded to the personal history. Next, the system followed links from the primary nodes, selected nodes that were connected to the primary nodes in the graph of the content, and called them secondary nodes. The secondary nodes were close to the personal history. Thus the system could extend overlaps of personal history and content. Then, the system selected items corresponding to the fragments as candidates for recommendation.

IV. Implementation

A. Components of the Recommender System

Figure 5 shows the components of the context-based recommender system and their relationships. The components of the system are the system controller, the communication subsystem (content receiver, personal location history receiver, personal access history receiver, recommendations sender), the inference subsystem (content articulator, personal history articulator, superpositioner), and the data storage subsystem (content storage, personal history storage, content graph storage, personal history graph storage, recommendation candidates storage, log storage). The communication subsystem and the inference subsystem both reference information stored in the data storage subsystem and work in parallel.

B. Capturing and Storing Personal History

Personal history was stored in the Service platform, which organized the histories into the desired format every three hours. The receiver subsystem in the recommender system received organized personal history from the platform at intervals of three hours or more.

Locations in the real world included in the personal history were converted into text that represented nearby content as mentioned in Section III-A.3. Items that had locations within a selected distance around a user's history of locations were considered content related to the user.

The history of Internet use, which consisted of items that the user had read on the Preview channel and text that the user wrote to other services in the My Life Assist Service, was considered to reflect a user's interests more directly than items related to the user. Therefore the receiver subsystem could give some weight to each history, and stored them in the personal history storage.

C. Articulation of Personal History and Content

The context-based recommender system aimed to extend the user's interests by providing enhanced context to the user's personal activities and augmented the content by discovering and extending the overlaps of personal activities and the content. First, the system articulated personal history and content for recommendations.

In the articulation phase, a graph structure of information was extracted, with these chunks as nodes and links as connections. An articulation process has three steps: segmentation that decomposes information into small fragments; clustering that groups fragments into clusters; and linking that connects clusters.

The system iterated articulation for 110 categories with 30,000 items selected from each category using the most recent first strategy.

1) Segmentation: In the segmentation phase, the recommender system segmented each text of the personal history and the content by windowing with a window size of L and the overlap $L/3$. Smaller values of L increased the number of branches in the graph; larger values reduced the number of branches [6].

There are text segmentation methods that analyze text expressions [7], [8]. However, the texts in the Preview channel were created without any prescribed rules by many people, and because the system applied clustering to text fragments in the next step, windowing was employed to split text.

2) Clustering: In this phase, the system converted fragments into vectors of words using a morphological analysis system. It used MeCab[3] as a morphological analyzer and collected nouns and unknown words. It generated vectors of words with the likelihood ratio[9], [10] between a fragment and each word in a fragment. Then it clustered them using hierarchical Bayesian clustering (HBC)[11].

Tfidf is generally used to generate vectors of words from documents, but the effect of word frequency is large and high-frequency words are not excluded [12]. Thus we introduced the likelihood ratio to generate word vectors. The common clustering methods for text data such as Ward's method tend to generate a few very large clusters that include half of all data, but HBC tends to generate clusters of similar size.

If a very large cluster is generated, almost all clusters have links to that cluster. It means that there is less variation of context in the graph structure of information [6]. Therefore we introduced HBC for the articulation.

3) Linking: Generated clusters are groups of text fragments, which are segmented from a text. Fragments from the same text can be distributed to several clusters. Thus, the system made links between clusters that included text fragments from the same text.

The strength of a link $l(i, j)$ from cluster C_i to cluster C_j was calculated using Jaccard's coefficient. That is, if $|C_k|$ is the number of original text fragments included in C_k:

$$l(i, j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}.$$

In this way, graphs for both the personal history and the content were generated.

Clustering was performed based on similarity between text fragments, so clusters were groups with similar information generated by the system. On the other hand, linking was performed based on the number of fragments from the same text, so links indicated the context of the text creator's intention.

D. Superposition of Personal History and Content

The recommender system looked for an overlap between the graph of a user's personal history and the graph of the content, extended the overlap together with links in the graph of the content, and selected candidates for recommendation from the extended area.

1) Selecting Categories: The content was articulated for each of the 110 categories; the system then selected categories to recommend. It generated word vectors of all fragments of a user's personal history by the method mentioned in IV-C.2. It also generated the same kind of vector for each category of the content, and calculated the similarities the user's personal history and each category based on the similarities of these vectors.

The real-world part of the personal history was part of the content, converted from locations that related to items. Categories that consisted of many items with locations would appear frequently in personal histories. On the other hand, categories that had few locations would not appear often. Therefore, categories that included many locations would tend to be chosen as candidates for recommendation.

However, the Preview channel required the recommender system to recommend all categories. We therefore set weights for similarity between personal history and categories. The weight was N_A/N_P, where N_A was the number of items in a category and N_P was the number of items with locations in the category. The system selected categories as candidates based on weighted similarities.

2) Selecting Clusters and Content: The system generated word vectors from each cluster in the graph of each category that was selected as a candidate for recommendation. It also generated vectors from each cluster in the user's personal history.

It calculated similarities between each cluster in the graph of personal history and each cluster in the graph of selected categories of the content. It selected the combination of most similar clusters in the personal history and the content. The personal history cluster was named the base cluster, while the category cluster was named the main target cluster.

Next, the system selected three clusters that linked to a main target cluster based on the strength of the link $l(i, j)$ for a category. These clusters were named subtarget clusters, and formed the extended area of recommendation along with the context of the content.

[3] http://mecab.sourceforge.net/

The system generated word vectors of text fragments from the base cluster, the main target cluster, and the subtarget clusters. It calculated similarities between the base cluster and the fragments included in the main and subtarget clusters, to find K similar fragments. Then it stored items that had been segmented to these fragments in the recommendation candidate storage. It gave the recommendation candidate rank $RC_{c,k}$ to an item when it was included in the cth candidate category and had the kth largest similarity to the base cluster:

$$RC_{c,k} = K(c\text{-}1) + k.$$

E. Choosing Recommendations

Before sending recommendations to a user, the system selected them from the available candidates based on the location of the user. When a candidate had a location, the recommendation order $R_{c,k}$ was determined by (1); d was the distance in the real world between the user's location and the item and D_R was a constant that indicated how the distance in the real world affected the rank. On the other hand, if a candidate did not have a location, $R_{c,k}$ was determined by (2) with a constant value D_F :

$$R_{c,k} = RC_{c,k} + d\,D_R \quad (1)$$

$$R_{c,k} = RC_{c,k} + D_F \quad (2)$$

First, the system selected four categories to recommend with high recommendation ranks $R_{c,k}$. Next, it selected recommendations from each category based on $R_{c,k}$. In this way, four recommendations for each of the four categories were chosen and they were sent to the user's mobile phone via the My Life Assist Platform.

V. Experiments

A. Outline of Service Operation

The My Life Assist Service that included our context-based recommender system operated as a field experiment from December 2007 to February 2008. Figure 6 shows the weekly total of users and captured activities in the My Life Assist Service. In the last week of the experiment, the context-based recommender system obtained 196 users' personal histories. Activities obtained as personal histories over the same period totaled 601,773, including 360 accesses to the content via the Internet-access function of a mobile phone. Thus, 99.94% of activities were locations of users.

B. Effect of Extending Context

Table I shows the number of Web pages recommended and the number accessed that were selected from the primary nodes and the secondary nodes by 50 users in the last week of the field experiment. The access ratio for content selected from the secondary nodes was higher than that from the primary nodes by about 20%. Thus, we can say that the aim of the context-based recommender system, to recommend nonconverged contents and shift with the user's interest, was almost achieved.

VI. Summary

We have described a recommender system that recommended content based on personal histories. These

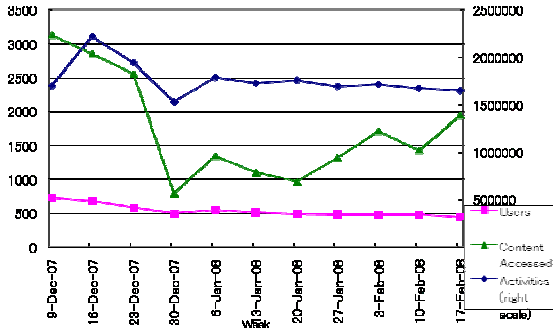


Figure 6. Weekly total of users and logs

TABLE I.
NUMBER OF CONTENTS RECOMMENDED AND ACCESSED (FEB.23-20, 50USERS)

Source	Recommended	Accessed	Ratio
Primary	9,071	72	0.79%
Secondary	5,898	58	0.98%
Total	14,969	130	0.87%

personal histories were collected by personal mobile phones, and consisted of a user's locations, a log of accesses to the Internet, and the user's contributions to the Internet. The system operated from December 2007 to February 2008 as a part of the *My Life Assist Service* by NTT DOCOMO, Inc. It recommended content from a user-generated Q&A Web site and a restaurant information Web site.

Because the service supported ordinary personal life via a mobile phone that was always with the user, the recommender system aimed to select content that was not converged, but varied.

We thought that the content expressed part of its creator's intention, and was located in a context. The recommender system found overlaps between the context of a Web page and the context of the user, extended the overlap along with the context of the content, and recommended Web pages from this extended area. In this way, the system recommended content from different aspects of the user's personal history.

In the experiment described in this paper, the recommender system extracted contexts of a Web page from the material recommended to users. However, it would be preferable for content to belong to various contexts. Clearly, content should consist not only of simple advertisements, but also background information that would interest a user. For example, if a Web page for a shop was selected using multiple contexts, not just suggestions from the shop but also information about the thoughts that the owner of the shop had had, the recommender system could make recommendations that may allow the user to enjoy visiting the shop more.

Content on the Internet is often created to attract public attention, sometimes to improve ranking in a search engine listing. However, to move beyond connecting a person to information by searching the Internet using key

words, a system must create content in the context of the intentions of the creator that accords with the context of the user's personal history. A recommender system that finds the overlap between them, extends that overlap in the context of the content and presents the extended content, will enhance the user's experience. Then people will again be able to enjoy waves of information.

ACKNOWLEDGMENT

This research was a part of the *My Life Assist Service* by NTT DOCOMO, Inc. as a part of the *Information Grand Voyage Project* promoted by the Ministry of Economy, Trade and Industry in Japan. The authors would like to thank Dr. Koji Abe, AdIn Research, Inc., TAMON Co., Inc., NEC Corporation, Kanshin!, Inc., Kadokawa Cross Media Co., Ltd., and NTT Resonant Inc.

REFERENCES

[1] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[2] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[3] M. Hirose, "Lifelog and virtual timemachine," *Journal of the Japan Society of Mechanical Engineers*, vol. 109, no. 1056, pp. 892–895, 2006.

[4] S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough: how accuracy metrics have hurt recommender systems," in *Proc. SIGCHI '06 Conference on Human Factors in Computing Systems*, 2006, pp. 1097–1101.

[5] C. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proc. of the 14th International Conference on World Wide Web*, 2005, pp. 22-32.

[6] K. Tanaka and A. Takasu, "Topic change extraction from problem solving records," in *Proc. of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2004.

[7] J. Allan, T*opic Detection and Tracking: Event-based Information Organization.* Kluwer Academic Publishers, 2002.

[8] M. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, pp. 33–64, 1997.

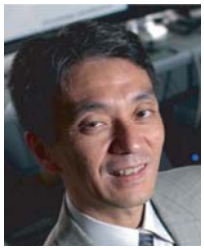
[9] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, pp. 61–74, 1993.

[10] M. Harada, S. Sato, and K. Kazama, "Finding key people and visualizing their relationships on the web," in *Information Processing Society of Japan SIG Notes, DBS-130/FI-71*, 2003, pp. 17–24.

[11] M. Iwayama and T. Tokunaga, "Hierarchical bayesian clustering for automatic text classification," in *Proc. 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1322–1327.

[12] A. Takano and S. Nishioka et al., "Development of the generic association engine for processing large corpora," in *19th IPA Technical Forum*, 2000.

Katsuaki Tanaka received his MS degree in informatics from Waseda University, Japan, in 2000. He is currently an assistant professor at the University of Tokyo. His research interests include natural language processing and knowledge management systems.
Mr. Tanaka is a member of JSAI and IPSJ.



Koichi Hori received B.Eng, M.Eng, and Dr.Eng. degrees in electronic engineering from the University of Tokyo, Japan, in 1979, 1981, and 1984, respectively.

In 1984, he joined National Institute of Japanese Literature, Japan, where he developed AI systems for literature studies. Since 1988, he has been with the University of Tokyo, Japan. He is currently a professor with School of Engineering, University of Tokyo. From September 1989 to January 1990, he also held a visiting position at University of Compiegne, France. His current research interest includes AI technology for supporting human creative activities, cognitive engineering, and Intelligent CAD systems.
Professor Hori is a member of IEEE, ACM, IEICE, IPSJ, JSAI, JSSST, and JCSS.

Masato Yamamoto received his MS degree in interfaculty information studies from the University of Tokyo, Japan, in 2004. He is currently a member of corporate sales division at NTT DOCOMO Inc. His research interests include natural language processing, knowledge management systems, and information retrieval.

Drivers of E-Government and E-Business in Jordan

Rand A. Obeidat
Jordan University of Science & Technology, Department of Computer Information Systems, Jordan
Email: raobeidat@just.edu.jo

Emad A. Abu-Shanab
Yarmouk University, Department of Management Information Systems, Jordan
Email: abushanab@yu.edu.jo

Abstract— **With the global emergence of e-government and its potential benefits to citizens in all its endeavors, there has been a growing need for research on drivers influencing the adoption of e-government services. This paper focuses on drivers influencing the adoption of e- government services among business organizations; hoping to have a better delivery of government services, the increased transparency and availability of information, and the improved interaction with businesses.**

Jordan is currently striving to move forward in e-government. However, figures reported in the Economist Intelligence Unit's (EIU) E- Readiness Ranking Report for the year 2008, Jordan ranked 53 out of 70 among countries with respect to its business environment. Also, Jordan ranked 50 among 192 countries according to the UN Global survey of e- government readiness in 2008. This paper aims to review the relationship between e-business and e-government in general, as well as e- government and e-business readiness indicators particularly in Jordan. In addition, it examines the motivators and barriers for adopting e- government among business organizations. Jordan needs to overcome barriers for adopting e-government among businesses, and reduce the gap between e-government and e-business with a mutual effort from both parties.

Index Terms— **Jordan, G2B, Motivators and Barriers, E-Government, E-Business**

I. Introduction

With the vast growth of the Internet and the huge options that it gives for conducting business with customers electronically, and the low costs and great convenience it gives to customers and businesses, it is becoming inevitable to avoid engaging in e-business. This issue is not limited to businesses seeking profit, but also to all types of business and even public firms. Park [22] states that electronic commerce can be summarized in three words: paperless, timeless and borderless. On the other hand, e-government is often referred to as "The E-Business of the State." Governments are requested to provide convenient services with affordable price.

They are responsible for a large proportion of the gross national product, and it is important for the government to use electronic media in order to save cost and make processes like procurement run smoothly and more transparently.

The Internet serves as an interface between the government and citizens in all its endeavors, and if it was utilized properly it will serve the objectives of firms in reaching a competitive advantage. Increasingly, CEOs are looking for improvements in their businesses; improvements in products, services and business processes. Researchers are asserting that the Internet has yielded new business models, where others claim that the Internet also changed existing business models to drive growth and improve overall business value. Based on that argument, and through the innovative use of technology, business transformation provides a solution for businesses to achieve their goals. Government transformation in the way government does business and serves its citizens through the review of its missions, shaping local industries and making sound recommendations as to how to shape processes to maximize the existing and new IT investments as well as to reengineer processes to enhance productivity and efficiency.

On the other hand, Jordan has been investing heavily in developing its ICT sector, aiming at enhancing the performance of its public and private sector organizations in terms of service provision, efficiency, accuracy, time and satisfaction [17]. However, businesses are also an important element of the information society, and the diffusion of ICT into businesses is essential for the achievement of productivity, growth, and the efficient functioning of the information society [12].

In this context, this paper is organized as follows: the following two sections will describe the relationship between e-business and e- government, and examine e-business and e- government readiness in Jordan. Also, this paper will identify the motivators and barriers for adopting electronic government among business organizations in Jordan. Section four will explore the government to business initiatives in Jordan, followed by conclusions and findings.

II. Relationship between E-Business and E-Government

This section describes the relationship between e-business and e-government with definitions of terms in

doi:10.4304/jetwi.2.3.204-211

the e-business arena. As described by research [24], e-government started as a subset from e-commerce or e-business, and its importance is attracting many researchers to work and try to understand the relationship between this force (e-government) and e- business.

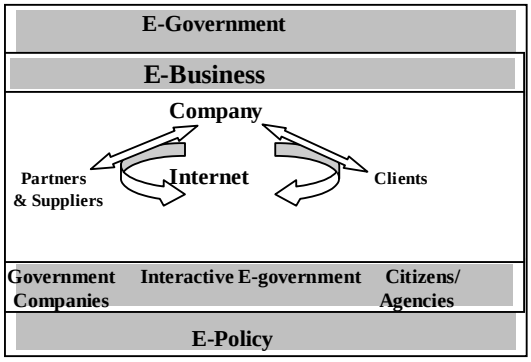


Figure 1: E-Business / E-Government

Beginning with the center of the Figure 1; e- business is defined as a business model and focuses on the support of processes, and relationships between business partners, employees andcustomers by means of electronic media.

The Government is confronted by new technologies in two different aspects; e-policy and interactive e-government. E-policy covers the legal framework governing the use of IT in the respective country (e.g. recognition of the digital signature), and the subsidies awarding or other means of support for the government operations. The second aspect is the interactive e-government that has to deal with defining its role as a market layer in the virtual environment, moreover, providing services and making public information available to companies and citizens over the Internet. As a result, e-government has two dimensions: (1) Endowing the economy with the necessary legal framework (making e- business possible), and (2) performing its operations and tasks in a cost effective manner. Finally, these related terms span the field of e- government in its purest sense: e-government as the governmental counterpart to e-business; the use of IT infrastructure for procurement, distribution of services and internal organizational functioning [24].

III. E-Business and E-Government Readiness in Jordan

The World Bank concluded to a set of international indices that can describe the level of readiness in the area of e-government. The set of indices are: Network Readiness Index (NRI) by the World Economic Forum (WEF), Technology Opportunity Index (ICT-OI) by the International Telecommunication Union (ITU), the E-government Readiness Index by the United Nation Public Administration Network Methodology by the World Bank Institute (WBI) and the E-Readiness Index by the Economist Intelligence Unit (EIU). This paper discusses the details of three indices: WEF's NRI, UNPAN's e-government index and EIU's E-Readiness report.

A. WEF's NRI

The Network Readiness Index (NRI), covering a total of 127 economies in 2008-2009, measures the degree of preparation of a nation or community to participate in and benefit from ICT developments. The NRI (shown in Figure 2) has three premises: a) there are three important stakeholders to consider in the development of ICT: Individuals, businesses and governments; b) There is a general macroeconomic and regulatory environment for ICT in which the stakeholders play out their respective roles; c) The degree of usage of ICT by the three stakeholders is linked to their degrees of readiness to use and benefit from ICT (Source: http://www.weforum.org/gitr). Table 1 lists some measures of the NRI that are related to Jordan.

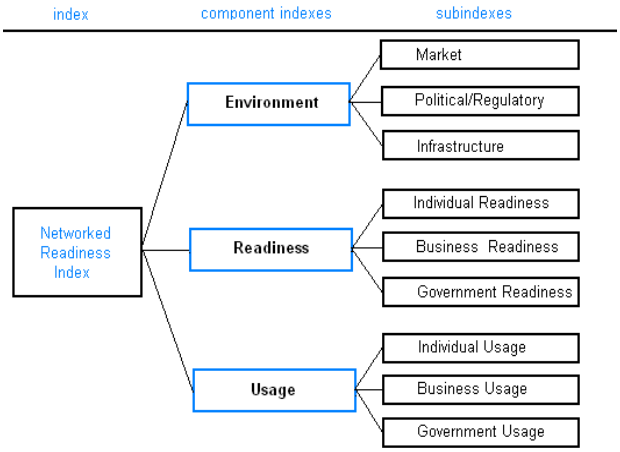


Figure 2: NRI premises (*Source: http://www.weforum.org/gitr)*

TABLE I. Related Measures of the NRI to Jordan

NRI 2008-2009	Jordan rank (out of 127)
Environment component	48
Market	51
Political	38
Infrastructure	59
Readiness Component	45
Individual readiness	53
Business readiness	76
Government readiness	29
Usage Component	45
Individual usage	73
Business usage	33
Government usage	33

We can see that Jordan scores best on government readiness and government usage, and scores least on business readiness and individual usage. Table 2 shows the top 10 strengths of Jordan position when measured in the NRI.

TABLE II. JORDAN'S TOP 10 STRENGTHS IN THE NRI

Indicator Name	Ranking
E-Participation index	15
Government prioritization of ICT	17
Burden of government regulation	18
Importance of ICT to government vision of the future	18
Government success in ICT promotion	18
Availability of new telephone lines	19
Quality of competition in the ISP sector	22
Property rights	23
Education expenditure	23
Intensity of local competition	24

The NRI index provides the ability to drill down from the overall index to component indices, sub-indices and eventually to individual indicators, to locate areas of comparative strengths or weaknesses in a country's ICT performance. It includes survey-based indicators that are unavailable in other indices or measures and have only one year lag time. However, firms surveyed are too few in number, too large in size and international.

Jordan had the greatest move upwards from the 90th position in 2005 to the 15th in 2008. The government of Jordan has put in place enhanced national portals which include features that increase citizen's engagement. The sites have a formal online consultation section, where the government receives feedback from its citizens on government policies and services.

B. EIU's E-Readiness Report

The Economist Intelligence Unit E-Readiness Ranking Index (EIU index) is a country's measure of its e-business environment. The ranking allows governments to gauge the success of their technology inhibitors against those of other countries, the ranking is produced in co-operation with the IBM Institute for Business Value and published yearly
(Source:http://www.eiu.com/siteinfo.asp?infoname).

The EIU e-Readiness Rankings consist of six categories: 1) Connectivity and technology infrastructure (allocated weight is 20% from the overall score), which measures the extent to which individuals and businesses can access mobile networks and the Internet, and their ability to access digital services through means such as digital identity cards. 2) Business environment (allocated weight is 15% of the overall score), evaluates the general business climate, the EIU screens 70 indicators to provide a comprehensive view of each country's attractiveness as a trading economy and as a destination for business investment. 3) Social and cultural environment (allocated weight is 15% of the overall score), which measures the population's "e-literacy"; its experience using the Internet and the technical skills of the workforce. 4) Legal environment (allocated weight is 10% of the overall score), which reflects the country's overall legal framework that e-business development depends on and directly affect the use of digital technology to inform, communicate and transact business. 5) Government policy and vision (allocated weight is 15% of the overall score), which assesses the activities of governments to adopt the technology, and their ability to lead their countries toward a digital future. 6) Consumer and business adoption (allocated weight is 25% of the overall score), which measures the amount that businesses and consumers spend on accessing ICT services and their adoption levels of e-commerce.

The EIU 2008 e-Readiness report found that Jordan in particular, is one of several Middle Eastern countries that have committed significant resources to e-government and e- business. However, Jordan ranked 53 out of 70 countries [26]. EIU 2008 e-Readiness report does not provide the exact indicators and data for assessment for each country. Such issue calls for extra work on this measure to pin point the weaknesses and strengths that lead to such ranking.

C. The UNPAN E-Government Index

The UNPAN E-Government Index is a composite measurement of the capacity and willingness of countries to use e-government for ICT-led development (Source: http://www.unpan.org/egovernment5.asp).

The E-government Readiness Index is composed of three indices: Web measure index, telecommunication and infrastructure index, and the human capital index. The web measure index mainly reflects the advanced services provided by the websites.

The telecommunication and infrastructure index reflects the technological infrastructure readiness for e-government application. Finally, the human capital index: reflects the degree to which citizens are prepared to participate in the networked world.

Jordan has scored a web measure index of (0.6054 out of 1.0000), a telecommunication infrastructure index of (0.1693 out of 1.0000); and a human capital index of (0.8677 out of 1.0000), which means that Jordanian citizens' readiness would not be an issue of debate. However, Jordan was ranked 50 among 192 countries according to the UN Global survey of e-government readiness in 2008.

IV. THE MOTIVATORS AND BARRIERS FOR ADOPTING GOVERNMENT TO BUSINESS E-SERVICES

Governments can play a dual role to enhance business activities within an economy. The first role is to facilitate e-business initiatives, in terms of providing the relevant infrastructure, talent/skills development initiatives, favorable policies (i.e. taxation, rules and regulations). Governments can also become active participants in e-business initiatives. For example, governments can use e-procurement solutions, when dealing with its suppliers [21].

This section explores the importance of government to small and medium enterprises (SMEs) under the umbrella of government-to-business (G2B) e-services. G2B transactions include various services exchanged between the government and the business community, including dissemination of policies, memos, rules and regulations. Business services offered include obtaining current business information, downloading application forms, renewing licenses, registering businesses, obtaining permits, payment of taxes and e-procurement [27]. In addition to services offered through G2B, transactions assist in business development, especially development of SMEs [23].

Eurochambres [10] found that the optimal framework that businesses need is result-oriented and business-centered, and should be market based. G2B services should be business centered means that they should motivate both supply and demand, as well as, should be proactive and responsive. And businesses need result-oriented e-government service, which means that services should neither overload technology, nor be bureaucracy-centered. Furthermore, it should actively promote innovation.

A. G2B E-Services Motivators

The literature proposed a large set of motivators, where they ranged from cost and tangible measures to more qualitative intangible measures [3][7]. The following is a list of the main motivators to a G2B service paradigm.

- Cost cutting of government and business operations.
- More efficient procurement process.
- Streamline and improve the consistency of personnel intensive tasks, such as: processing licenses, renewals or employee benefit changes.
- Reduction in cost associated with registration and submission of forms.
- Reduction in errors and redundant data entry at the point of submission of forms.
- Reduction in cost associated with registration and submission of returns for business and government.
- Consistent, accurate and up-to-date information for businesses on legislative requirements.
- Reduced time spent searching for forthcoming government procurement.
- Automatic notification of forthcoming procurement.

B. G2B E-Services Barriers

Lam [23] has classified the barriers to G2B services into four categories from the government point of view: strategy barriers, technology barriers, policy barriers and organizational barriers. Two other categories were added to cover the major obstacles facing governments in developing a good relationship and service towards businesses. The following is a description of the six categories.

Strategy Barriers:

The lack of common goals and objectives is an issue, where lacking collective thoughts and aims creates confusion among governmental agencies, and also becomes a part of conflict in responsibilities; it is necessary to have common thoughts for e-government initiatives between government agencies. The Jordanian government is aware of this issue and considers the e-government policy as an integral part of overall national socio-economic development and government transformation. It is not in isolation of the national set of priorities and is built around the service-oriented e-government initiatives facade.

It is imperative to setup a vision for the e-government project and spread it all through the organization. One great example of e-government vision is Dubai's e-government vision, where they emphasize serving citizens and businesses on one hand and providing the notion of becoming a global economic center on the other; and most importantly with few words to remember [8].

On the other hand, the Jordanian e-government project carries a more detailed vision that states: "E-Government in Jordan is dedicated to delivering services to people across society, irrespective of location, economic status, education or ICT ability. With its commitment to a customer-centric approach, e-Government will transform government and contribute to the Kingdom's economic and social development"[25]. It is important to realize such vision, but also to state explicitly the objectives of such huge transformation initiatives. Objectives need to be specific and comprehensive; the following are the objectives of the Jordanian e-government project:

1. Improve service delivery
2. Increase transparency
3. Improve responsiveness to citizens and businesses' needs and requirements
4. Save time and money of both the government and citizens and businesses.
5. Create positive effects on the society.

It is important to realize that businesses are not only the government customers, but also they are its partners and providers [14].

The second factor in the strategic dimension is the lack of ownership and authority: This issue regarding ownership and governance, program management requires solving this matter. Formal project responsibility or the strength of accountability is the major reason for the lack of ownership and authority. The third factor is the deficiency of implementation guidance: whereas the central government set up a vision for the e-government project, agencies and other public management parties require directions on how to transform such vision into reality. Without the proper guidance it is difficult to establish a good e-government project.

Finally, the financial issues are important, where large investment is needed to complete a successful functional e-government project. Due to the lack of financial support, e-government projects can't be successful. Financial support is needed at two stages: the first is the

facilitation of a capable infrastructure that has the capacity to serve the purposes of the project and provide support for all functionalities required. The second stage is the direct cost associated with the project which includes three major aspects: hardware needed like servers, computers, networks and other peripheral devices; software needed in the form of databases, systems and interfaces; and finally, training required that would enable the public sector staff to accommodate with this new technological area.

It is important to see that the dimensions related to the strategic category are not of a static type as they change with time and impose financial burdens on the government because of the cost of updating/upgrading and the continuous training effort.

Technology Barriers

The lack of architecture integration and capable infrastructure yields to different technological and policy problems. The under utilization of proper technologies, lack of quality application interface, and differences in frameworks applied create a barriers to architectural integration. ICT infrastructure does not consist of telecommunication and equipment only, but requires e-readiness and ICT literacy.

Having the basic needed knowledge related to IT is necessary to accept and use e-government services. The second aspect is the deficiency of data standards. The primary function in e-government is to transfer data into information and then knowledge needed in the sustained developmental efforts. The lack of data standards is a major technical barrier. Third, different security models implemented, which is described as one of the success factors in e-government adoption as it yields faith and confidence between users and governments. However, security model is identified as a major barrier in technical combination of e-government systems. The forth aspect is the lack of resources: in most cases, government does not have all the resources to complete e-government projects. Therefore, most of e-government initiatives fail to provide healthy outcomes. Finally, the technology needed to compete in e-government arena might not be available to all countries of the world, specially the developing countries, which raises digital divide between countries as an important aspect in e-government domain. One example of such issue is the differential status of the 3G technology as some developed countries lagged behind against others, where the USA implemented the technology after Japan.

Policy Barriers

Sharing of data between government agencies should be done in a controlled and proper way to protect sensitive information related to citizens. Second, data possession: many governments consider themselves as an owner of particular records and they are very concerned about sharing these records with others. The right or ownership to data must be clear in order to achieve e-government goals. Finally, e-government should be built around a national e-government policy, where policy-execution is a major cause of conflicts and incompatibilities in implementation of the overall policy. Lack of comprehensive course of action and inappropriate step of development may delay the process of e-government program.

Researchers in the Arab world concluded to four main administrative barriers related to e-government: the complexity of the policy execution in some countries and the bureaucratic procedures adopted; the lack of flexibility in policies and procedures, the lack of proper planning, and the lack of support of top management [18].

Organizational Barriers

Lack of organizational motivation is an important issue as many agencies are not yet ready for e-government challenges, and they are not well prepared for the initiatives of e-government because many agencies are not accepting this change. Bhatnager [11] emphasized the importance of managing change when implementing e-government initiatives. Second, the slow speed of government reform; some agencies have found it difficult to run with the speed of reform. It is necessary to change the state of mind rather than Governments' focus to move towards more technology oriented environment. Third, the lack of internal management and technical ability: lack of proper training within organizations is considered as a barrier for e-government implementation. As we discussed in the technical side of this section, training on software implemented is a costly and tedious process. Several agencies don't have trained and skilled people to execute e-government projects and also lack IT training programs. Finally, change management approach within organizations and governments is considered a vital issue. Change management approach includes the procedures established within organizations towards changing existing cultures. This is considered a key issue of organizational change and a big step towards a high capacity to change.

Legal Barriers

Most research in Arab world focused on the legal side of the equation especially in a developing countries [12] [4] [1]. Abu-Shanab, Abu-Al-Rub and Md Nor [2] recommended the following in the case of Jordan and in relation to the regulations and legislations; first, one of the most difficulties that encounter the Jordanian government in its experience is the absence of needed legislations that regulate electronic payments and its related issues. There is no doubt that e-government projects can't success unless a legal framework governing electronic transactions, which secures the rights of all parties and regulates the domain. Second, government employees need some special competencies when dealing with e-government applications especially the legal side of it. Finally, transparency and fairness are vital when engaging in e-government activities, thus employees in the e-government stages need training and special programs in change management that help facilitate the management and control of the process without any legal

contradictions. The authors recommended importing other countries' experience to easily facilitate the implementation of a comprehensive legal framework.

Human Barriers

The human factor in the area of e-government is a crucial one that appears in more than one dimension. As discussed in the organization and technology barriers, change management is crucial in the process of adopting e-government initiatives. It is related to human aspects and their power distribution and their personal and professional stake. The human aspect also is vital when estimating the training cost, which is considered the largest after the initial investment. Also, it is important when switching or adopting e-government initiatives to train staff and even citizens and businesses on new skills and competencies needed for this new era. Finally, needed human capital might not be available even if we have available financial support for them. When dealing with an issue like e-government, we talk about a wide spectrum of countries where needed skills and competencies are not available.

On the other hand, it is important to take into consideration the resistance of employees for such project as it change the power status and put some jobs at risk. The cases in the literature indicate such resistance whenever an e-government initiative is considered.

Suda [6] examined the G2B challenges from the government point of view and concluded to the following list:

- Deficiency in recruiting, retraining and training of in demand technical and business skills.
- Digital divide will require service delivery via multiple channels indefinitely.
- Extremely large pool of heterogeneous data.
- Traditional budget process makes funding of cross agency projects difficult.
- Requires substantial investments and is competing for funds with "visible" missions.
- Leveraging existing IT investments by integration with legacy systems.

Mizza [13], examined the G2B e-services barriers from the business point of view and concluded that to a long list of barriers and issues like: cross border legal issues, cultural differences, language, lack of payment vehicles (like PayPal), lack of resources, cost of telecommunications, total cost of ownership (TCO) is too high, IT illiteracy among decision maker, shortage of IT skills, intellectual property rights, fear of channel conflict, resistance to change, fear of fraud, fear of identity theft, mistrust of electronic payments, and local customer base too small (critical mass threshold).

V. Government to business initiatives in Jordan

To provide proper services and facilitate the G2B transactions, a set of fast-track projects were implemented in Jordan [8]:

- Business regulations: this service will enable businesses to register and pay fees online.
- Telecommunications licensing and regulation: this service will allow ISPs and telecommunications providers to apply for and obtain licenses online.
- Taxation and social security: would enable citizens and businesses to electronically submit tax returns and make payments to different revenue and social security institutions.
- Selling to government: this project will digitize government procurement.

However, these fast-track projects are still under implementation, but the most important issue is how to assess G2B practices. G2B Good practices are evaluated and assessed according to specific "Quality Control" check, consisting of a set of assessment criteria [15]. The assessment criteria are:

- Use of ICT: to verify successful, innovative and value-for-money use of ICT (open standards, multiple access platforms for users, open access tools, interoperability within and between government agencies, technologies and systems enabling data protection and security).
- Innovativeness: to see if there is something new or different, at least going beyond the average situation to become a forerunner in its field, thus providing new ways of successfully implementing eG2B services.
- Managing e-government implementation: the efficient and innovative coordination between initiatives and projects; the coordination and decision-making between agencies and/or different levels of government and between government and private sector and/or non-profit sector partners; and between government and private sector and/or non-profit sector partners; and management of the different aspects of the changes.
- Real practical results: evidence of impact can be given by documented economic results (i.e. time and cost savings) or other qualified results which document the extent and type of impact.
- Functionality (for users and for government agencies): for users they are enjoying upgraded services which support/improve the functioning of their business; for government agencies, they constitute significant improvements in fulfilling its functional requirements.
- Visibility: the case is visible and distinctive, i.e. it has a clear identity and is recognized, at least in the region or country of location, by users and/or government agencies as making a beneficial contribution to government, society and economy.

VI. New Architecture for the E-government Program in Jordan

The Strategy for the e-government program was to deliver government information, services, and processes, using information and communication technologies (ICT) to transform the way government engage with people and businesses [20].

According to [9], the new Strategy reaffirms E-government to the goal of delivering high-quality customer centric and performance-driven services to E-government customers. The new aims are summarized below as :

- Improve service delivery, quality and speed of government's interaction with clients, businesses and organizations.
- Provide a single point of contact for all government entities.
- Facilitate sharing and exchange of data, information and knowledge between government entities.
- Increase e-business and e-commerce activities in Jordan. This will be reached by passing manymaturity stages, and after overcoming barriers to adopt e-business and e-commerce services.

The new strategy of the e-government program in Jordan implies the need for new architecture. The well suited architecture for achieving the aims of the new strategy; such as integrating distributed systems, delivering high-quality customer centric and performance-driven services, is planned to be Service Oriented Architecture (SOA). SOA is considered to be one of the best ways to provide interoperability and integration between various range of services, implemented by different software applications, running on a variety of platforms in government organizations.

VII. Conclusions

E-government is growing to a size that requires from governments full attention and demand for the collaboration and facilitation with all parties involved (i.e. private sector and NGOs in society). This requires an adequate infrastructure in all aspects: technical, human and procedural. According to the NRI index, Jordan would need to focus policy attention to the individual usage and business readiness for ICT. Macroeconomic and business environment poses several challenges for Jordan, more so than other dimensions of e-readiness. So for policy makers, there is a vital need to know e-readiness indicators for the government as well as for businesses in order to apply needed strategies that reflect realistic objectives. Also, the government needs to monitor and evaluate the activities that have been lunched to assure that their implementation is tracked with accountability and transparency.

Jordan needs to overcome barriers for adopting e-government among businesses with a mutual effort from both parties. Also, there is a vital need for more information regarding G2B case in Jordan in order to be able to apply the quality control assessment criteria. A good step would be to establish a higher committee for the monitoring e-government performance, where indicators of e-readiness are monitored (covering the first theme of this paper), and watch for barriers that hinder the advancement of e-government initiatives. As well as, measuring the percentage of strategy, technology, policy and organizational barriers of G2B e-services to identify correct actions that overcome these barriers and to improve efficiency and organizational effectiveness.

VIII. Future Work

Future work is encouraged to explore other e-government readiness measures and try to integrate those into the process of improving e- government performance to assure quality results. As we mentioned, some indices did state the measures that were used to built the index, which means that more research is needed to improve the fit of the index to the country and the accuracy of the rankings.

Future research is needed to review the Service Oriented Architecture (SOA) principles, benefits and best practices that helped countries to satisfy the goal of delivering high-quality customer centric and performance-driven services in the field of e-government.

References

[1] Aboud, F. (2005). An Arabic Book entitled: “E-government, Between Implementation and Planning”, King Fahad National Library, Reyadh, Saudi Arabia, 2005.
[2] Abu-Shanab, E., Abu Al-Rub, S. & Md Nor, K. (2010). Obstacles Facing the Adoption of E-government Services in Jordan. Journal of E-Governance, Vol. 33(1), 2010
[3] Alomari H., ' E-Government architecture in Jordan: A Comparative Analysis'. Journal of computer science, 2(11): 846-852, 2006 .
[4] Baz, B. (2009). An Arabic Book entitled: “E-government Role in Managerial Decision-Making and Electronic Voting”, Dar Al-Kotob Al-Qanonieh’, Egypt, 2009.
[5] Bhatnager, S. (2004). “E-government From Vision to Implementation”, Sage Publications 2004.
[6] Bob, Suda; Assistant commission er IT solutions,G2B workshop, Buyers/ Sellers views on B2G. , 2001.
[7] Bonham G.M, Seifert; The transformative potential of E-Government in Transitional Democracies, Public Management. Electronic Journal. , 2003.
[8] Dubai e-government website: www.dubai.gov....
[9] E-government Program (2007). Jordan E-government Architecture Vision..
[10] EUROCHAMBRES; Position paper, eEUROPE 2005 Mid-Term Review. , 2003.
[11] Grunauer, Karl Maria; Fleisch, Elgar, Osterle, Hubert: E-Commerce – project in supply chain. (2000).
[12] Hijazi, A. (2008). An Arabic Book entitled: “E-government, Between Ambition and Reality”, Dar Al-Fikr Al-Jamee’, Alexandria, Egypt, 2008.
[13] Ian Mizza, MBA,http://www.miti.gov.mt/docs/Ian%20Mizz%20-%20FOI.pdf .
[14] Jordan E-government Strategy, 2006, accessed from the e-government website: www.jordan.gove.jo.
[15] Jordan e-government website: www.jordan.gove.jo
[16] Lam Wing , " Barriers to e-government integration", the journal of enterprise information management, Vol.18, No.5, pp.511-530, (2005)
[17] Launching E-government in Jordan, Ministry of information and communication technology \MOICT). http://www.moict.gov.jo/moict/downloads/Jordan_E-Gov_CH4.pdf.
[18] Matar, I. (2008). An Arabic Book entitled: “E-government, between theory and implementation”, Dar Al-Jamea’ Al-Jadeedah, Alexandria, Egypt, 2008.
[19] Mckenna, Regis: Relationship Marketing- Successful

strategies for the Age of the customer, New Yourk: Addision-Wesley 1991.
[20] Ministry of Information and Communications Technology, E-readiness Assessment of the Hashemite Kingdom of Jordan, 2006, http://www.moict.gov.jo/MoICT/MoICT_Jordan_ereadiness.aspx.
[21] Murali Raman, Maniam Kadiannan & Ming Yu. E-business and E-government: Issues and Challenges in Malaysia, Information Technology Journal 6(3) 428-434,(2007).
[22] Park, IK: National Competitiveness and EC, in : The Journal of Information (Korean), Vol.3, December 1996, pp.20-41, (1996) .
[23] Patricia J. Pascual; E-Government , e- ASEAN Task Force UNDP-APDIP. 2003
[24] Petra Schubert and Urs Hausler: E- Government meets E-Business; Institute for business Economics (IAB) University for Applied Sciences Basel, Switzerland (FHBB). (2001).
[25] Solvenian National report for the Transfer- East project, 2006.
[26] The Economist Intelligence Unit (EIU), The 2007 e-readiness rankings report. http://www.eiu.com/siteinfo.asp?infoname.
[27]Yari. e-government and Regulatory environment,2008 www.egov.iist.unu.edu/index.php?/.../E- governance%20.

Requirements Engineering in Agile Software Development

Andrea De Lucia and Abdallah Qusef
Dipartimento di Matematica e Informatica, University of Salerno
Via Ponte don Melillo, 84084 Fisciano (SA), Italy
{adelucia, aqusef}@unisa.it

***Abstract*—Finding out, analyzing, documenting, and checking requirements are important activities in all development approaches, including agile development. This paper discusses problems concerned with the conduction of requirements engineering activities in agile software development processes and suggests some improvements to solve some challenges caused by agile requirements engineering practices in large projects, like properly handling and identifying sensitive (including non-functional) requirements, documenting and managing requirements documentation, keeping agile teams in contact with outside customers. The paper also discusses the requirements traceability problem in agile software development and the relationships between the traceability and refactoring processes and their impact on each other.**

***Index Terms*—Requirements Engineering; Agile Software Development, Traceability, Refactoring.**

I. Introduction

The agile approach is creating a stir in the software development community. Agile methods are reactions to traditional ways of developing software and acknowledge the "need for an alternative to documentation driven, heavyweight software development processes" [1]. In the implementation of traditional methods, work begins with the elicitation and documentation of a "complete" set of requirements, followed by architectural and high-level design, development, and inspection. Beginning in the 1990s, some practitioners found these initial development steps frustrating and, perhaps, impossible [2]. The industry and technology move too fast, requirements "change at rates that swamp traditional methods" [3], and customers have become increasingly unable to definitively state their needs up front while, at the same time, expecting more from their software. As a result, several consultants have independently developed methods and practices to respond to the inevitable change they were experiencing. These Agile methods are actually a collection of different techniques (or practices) that share the same values and basic principles. The Agile Manifesto states valuing " individuals and interaction over processes and tools, working software over comprehensive documentation, customer collaboration over contract negotiation, and responding to changes over following a plan" [1].

Requirements Engineering (RE) is the process of establishing the services that the customer requires from a system and the constraints under which it operates and is developed. The main goal of a RE process is creating a system requirements document for knowledge sharing, while Agile Development (AD) methods focus on face-to-face communication between customers and agile teams to reach a similar goal. There are several research papers discussing the relationship between RE and AD, e.g. [4, 5, 6, 7, 8, 9]: they explain some RE practices in agile methods, compare these practices between agile and traditional development systems, and examine the problems of AD when it is dealing with the management of large projects and control critical requirements.

This paper addresses the problem of how (user) requirements can be captured and specified in the context of agile software development approaches. It therefore tries to identify how standard RE techniques and processes can be combined with agile practices and to find solutions to some of the difficulties related to their work. In addition, this article discusses the traceability problem in agile software development, since the current traceability between agile software artifacts is ill defined [10]. In particular, we discuss how to solve the traceability problem by extracting some important information from software artifacts to identify a traceability links between them, we also discuss how these links can be used to improve the decisions making process and help developers during the refactoring process. Finally, the paper comes up with a set of guidelines for agile requirements engineering.

The paper is organized as follows; the next Section sheds light on the importance of agile development in IT organizations and the benefits and limitations of agile methodologies in the software development life cycle and discusses some of agile approaches from a requirements engineering perspective. The agile RE activities are discussed in detail in Section 3, beginning with the objectives of the activity and explaining the techniques used to achieve these goals in AD, then the problems of each activity are identified and improvements to remedy these problems are discussed. In Section 4 some guidelines and enhancements are described concerned with an efficient application of RE practices in AD. Finally, Section 5 summarizes our conclusions and future work.

doi:10.4304/jetwi.2.3.212-220

II. AGILE SOFTWARE DEVELOPMENT

The goal of agile methods is to allow an organization to be agile, but what does it mean to be Agile? Jim Highsmith says that being Agile means being able to "Deliver quickly. Change quickly. Change often" [2]. While agile techniques vary in practices and emphasis, they follow the same principles behind the agile manifesto [1]:

- Working software is delivered frequently (weeks rather than months).
- Working software is the principal measure of progress.
- Customer satisfaction by rapid, continuous delivery of useful software.
- Even late changes in requirements are welcomed.
- Close daily cooperation between business people and developers.
- Face-to-face conversation is the best form of communication.
- Projects are built around motivated individuals, who should be trusted.
- Continuous attention to technical excellence and good design.
- Simplicity.
- Self-organizing teams.
- Regular adaptation to changing circumstances.

Agile development methods have been designed to solve the problem of delivering high quality software on time under constantly and rapidly changing requirements and business environment. Agile methods have a proven track record in the software and IT industries. Fig. 1 shows that about 69% of organizations are adapting one or more of agile practices for use in general project management as well as organizational development [11].

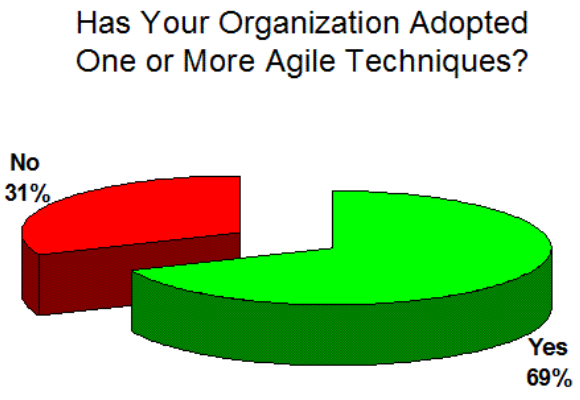


Figure 1 Agile Development Adoption

In fact, the agile development methodologies are used in organizations where there is no requirement freezing, incremental and iterative approach is used for modeling and every one in the team is an active participant and everyone's input is welcome. The main benefit of the agile development software is that it allows for an adaptive process - in which the team and development react to and handle changes in requirements and specifications, even late in the development process. Through the use of multiple working iterations, the implementation of agile methods allows the creation of quality, functional software with small teams and limited resources. The proponents of the traditional development methods criticize the agile methods for the lightweight documentation and inability to cooperate within the traditional work-flow. The main limitations of agile development are: agile works well for small to medium sized teams; also agile development methods do not scale, i.e. due to the number of iterations involved it would be difficult to understand the current project status; in addition, an agile approach requires highly motivated and skilled individuals which would not always be available, lastly, no enough written documentation in agile methods lead to information lose when the code is actually implemented. However, with proper implementation the agile methods can complement and benefit traditional development methods. Furthermore, it should be noted that traditional development methods in non-iterative fashions are susceptible to late stage design breakage, while agile methodologies effectively solve this problem by frequent incremental builds which encourage changing requirements. In the following, some common agile methods are briefly discussed from a requirements engineering perspective.

Agile Modeling (AM) is a new approach for performing modeling activities [12]. It gives the developers a guideline of how to build models - using an agile philosophy as its backbone- that resolve design problems and support documentation purposes but not 'over-build' these models. The aim is to keep the amount of models and documentation as low as possible. The RE techniques are not explicitly referred in AM but some of the AM practices support some RE techniques like brainstorming.

Feature-Driven Development (FDD) consists of a minimalist, five-step process that focuses on building and design phases [13] each defined with entry and exit criteria, building a features list, and then planning-by-feature followed by iterative design-by-feature and build-by-feature Steps. In the first phase, the overall domain model is developed by domain experts and developers. The overall model consists of class diagrams with classes, relationships, methods, and attributes. The methods express functionality and are the base for building a feature list. A feature in FDD is a client-valued function. The feature lists is prioritized by the team. The feature list is reviewed by domain members [14]. FDD proposes a weekly 30-minute meeting in which the status of the features is discussed and a report about the meeting is written.

Dynamic Systems Development Method (DSDM) was developed in the U.K. in the mid-1990s. It is an outgrowth of, and extension to, Rapid Application Development (RAD) practices [15]. The first two phases of DSDM are the feasibility study and the business study. During these two phases the base requirements are elicited. Further requirements are elicited during the

development process. DSDM does not insist on certain techniques. Thus, any RE technique can be used during the development process [9]. DSDM's nine principles include active user involvement, frequent delivery, team decision making, integrated testing throughout the project life cycle, and reversible changes in development.

Extreme Programming (XP) is based on values of simplicity, communication, feedback, and courage [16]. XP aims at enabling successful software development despite vague or constantly changing software requirements. The XP relies on the way the individual practices are collected and lined up to function with each other. Some of the main practices of XP are short iterations with small releases and rapid feedback, close customer participation, constant communication and coordination, continuous refactoring, continuous integration and testing, and pair programming [17]. Table I shows how RE activities are implemented in XP approach. In fact, XP is the most famous of any of the agile approaches.

Scrum is an empirical approach based on flexibility, adaptability and productivity [18]. The Scrum leaves open for the developers to choose the specific software development techniques, methods, and practices for the implementation process. Scrum provides a project management framework that focuses development into 30-day Sprint cycles in which a specified set of Backlog features are delivered. The core practice in Scrum is the use of daily 15-minute team meetings for coordination and integration. Scrum has been in use for nearly ten years and has been used to successfully deliver a wide range of products; Table II summarizes how RE activities are implemented actually in Scrum.

In this article some recommendations are suggested for agile development teams to help them in managing and implementing large projects and projects with critical requirements.

TABLE I.
RE IMPLEMENTATION IN XP

RE activity	XP implementation
Requirements Elicitation	• Requirements elicited as stories. • Customers write user stores.
Requirements Analysis	• Not a separate phase. • Analyze while developing. • Customer prioritizes the user stories.
Requirements Documentation	• User stories & acceptance tests as requirements documents. • Software products as persistence information. • Face-to-face communication.
Requirements Validation	• Test Driven Development (TDD). • Run acceptance tests. • Frequent feedback.
Requirements Management	• Short planning iteration. • User stories for tracking. • Refactor as needed.

TABLE II.
RE IMPLEMENTATION IN SCRUM

RE activity	Scrum implementation
Requirements Elicitation	• Product Owner formulates the Product Backlog. • Any stakeholders can participate in the Product Backlog.
Requirements Analysis	• Backlog Refinement Meeting. • Product Owner prioritizes the Product Backlog. • Product Owner analyzes the feasibility of requirements.
Requirements Documentation	• Face-to-face communication.
Requirements Validation	• Review meetings.
Requirements Management	• Sprint Planning Meeting. • Items in Product Backlog for tracking. • Change requirements are added/deleted to/from Product Backlog.

III. REQUIREMENTS ENGINEERING FROM THE AGILE DEVELOPMENT POINT OF VIEW

RE is concerned with discovering, analyzing, specifying, and documenting the requirements of the system. RE activities deserve the greatest care because the problems inserted in the system during RE phase are the most expensive to remove. As shown in Fig. 2, some studies revealed that around 37% of the problems occurred in the development of challenging systems are related to the requirements phases [19].

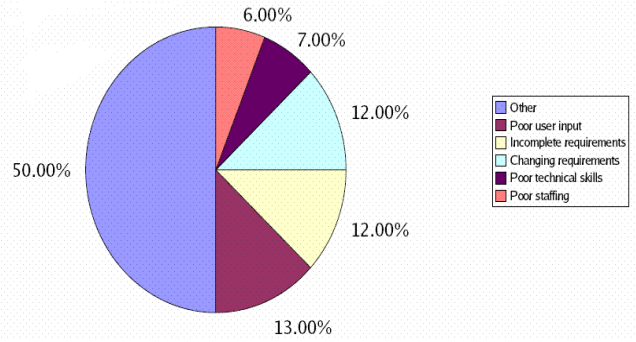


Figure 2 Problems of challenging systems

The main difference between traditional and agile development is not whether to do RE but when to do it. The RE processes in traditional systems focuses on gathering all the requirements and preparing the requirements specification document before going to the design phase, while the agile RE welcomes changing requirements even late in the development lifecycle.

Agile RE applies the focal values mentioned in the agile manifesto to the RE process. The processes used for agile RE vary widely depending on the application domain, the people involved and the organization developing the requirements. However, this paper explains the agile RE activities which are: Feasibility

Study, Elicitation and Analysis, Documentation, Validation, and Management.

A. Feasability Study

The Feasibility Study gives the overview of the target system and decides whether or not the proposed system is worthwhile. The input of the feasibility study is an outline description of the system and how it will be within an organization. The results should be a short report, which recommends whether or not it is worth carrying on with the RE and AD process. Initially, all relevant stakeholders have to be defined, in other words, all right customers who are related to the development of the system and are affected by its success or failure must be selected, and then the brainstorming session takes place to share the knowledge ideas between agile teams and "ideal" customers to answer a number of questions like:

1) **Does the system contribute to the high level objectives and the critical requirements of the organization?**
   In a first step, the high level goals and critical requirements (functional and non-functional requirements) for the system are defined upfront in order to determine the scope of the system; these requirements describe the expected business values to the customer.
2) **Is your organization ready for the AD?**
   Each agile method has its own characteristics and practices that will change the daily work of the organization. Before an organization selects one of them, it should consider whether or not it is ready for agile development. This is a very important question and many researchers tried to answer it [11, 20]. For example, Ambler [11] discusses some successful factors and questions to be answered affecting the successful adoption of agile methods.
3) **Can the system be implemented within given budget?**
   Some contracts do not allow for changing requirements. "The requirements must be complete before a contract can be made, which is often found in fixed-priced projects" [6]. In agile projects where changing requirements is welcomed, contracts often are based on time and expenses and not on fixed-priced scope. Also, "agile methods use scope-variable price contracts" [21]. This means that the features really implemented into the system and its cost evolve as well. Therefore, requirements are not specified in details at contract level but defined step by step during the project through a negotiation process between the customer and the development team [8].
4) **How to integrate the agile activities with traditional organizational activities already in place?**
   Some researches suggest tentative models for integrating agile activities with traditional organizational activities by transferring the knowledge from one process to another and how the traditional team should adopt its activities to suit the mechanisms of agile teams [22, 23].

B. Requirements Elicitation

In this activity, agile teams work with stakeholders to find out about the application domain, the services that the system should provide, the system's operational constraints, and the required performance of the system (non-functional requirement). The most important techniques used for requirements elicitation in AD are:

1) **Interviews:** "Interviewing is a method for discovering facts and opinions held by potential stakeholders of the system under development" [7]. There are two types of interviews: Closed interviews, where a predefined set of questions are answered, and the Open interviews, where there is no predefined agenda and a range of issues are explored with stakeholders. In fact, interviews are good for getting an overall understanding of what stakeholders do and how they might interact with the system, but they are not good for understanding domain requirements. All agile methods say that interviews are an efficient way to communicate with customers and to increase trust between two sides.
2) **Brainstorming:** this is a group technique for generating new, useful ideas, and promoting creative thinking. Brainstorming can be used to elicit new ideas and features for the application, define what project or problem to work on and to diagnose problems in a short time. The project manager plays an important role in brainstorming. He/she determines the time of creative session, makes sure that there is no escalating discussions about certain topics, and comes to make sure that every body expresses his/her opinion freely. After the creative session is ended, the topics are evaluated by the team. Also, the connections and dependences between the discussed ideas are represented by (for example) graph visualization, so the conflicts with other requirements are found and evaluated.
3) **Ethnography:** it is an observational technique that can be used to understand social and organizational requirements [24]. In agile development ethnography is particular effective at discovering two types of requirements: the first one refers to requirements that are derived from the way in which people actually work rather than the way in which process definitions say they ought to work, and the second one refers to requirements that derived from cooperative and awareness of other people's activities. Ethnography is not a complete approach to elicitation and it should be used with other approaches such as use case analysis [19, 24].
4) **Use Case analysis:** this is a scenario based technique used in UML-based development

which identifies the actors involved in an interaction and describes the interaction itself. A set of use cases should describe possible interactions that will be presented in the system requirements; each use case represents a user-oriented view of one or more functional requirements of the system [24].

C. Requirements Analysis

The main task here is to determine whether the elicited requirements are unclear, incomplete, ambiguous or contradictory, and then resolve these issues. Conflicts in requirements are resolved through prioritization negotiation with stakeholders. The main techniques used for requirements analysis in agile approaches are:

1) **Joint Application Development (JAD):** this is a workshop used to collect business requirements while developing a system. The JAD sessions also include approaches for enhancing user participation, expediting development, and improving the quality of specifications [24]. In agile environment, in case of conflicts between stakeholders' requirements the use of JAD can help promoting the use of a professional facilitator who can help to resolve conflicts. In addition, the JAD sessions encourage customer involvement and trust in the developed system.
2) **Modeling:** system models are important bridge between the analysis and the design process [7]. In agile environment the pen board (or pin board also) is divided into three sections: models to be implemented, models under implementation, and models completed. "This layout provides a visual representation of the project status" [8]. These models must be documented and not throw-away.
3) **Prioritization:** agile methods specify that the requirements should be considered similar to a prioritized stack. The features are prioritized by the customers based on their business value, so that the agile teams estimate the time required to implement each requirement. The agile team must distinguish between "must have" requirements from "nice to have" requirements, this can be done by frequent communications with the customers.
   Fig. 3 shows the Requirements prioritization process: "at the beginning of each iteration, there is a requirements collection and prioritization activity. During that, new requirements are identified and prioritized. This approach helps to identify the most important features inside the ongoing project. Typically, if a requirement is very important it is scheduled for the implementation in the upcoming iteration; otherwise it is kept on hold. At the following iteration, the requirements on hold are evaluated and, if they are still valid, they are included in the list of the candidate requirements together with the new ones. Then, the new list is prioritized to identify the features that will be implemented, if a requirement is not important enough, it is kept on hold indefinitely" [8].

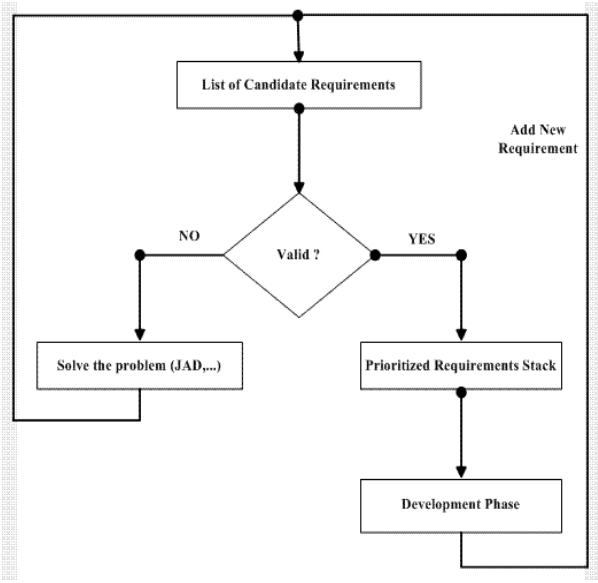


Figure 3 Requirements prioritization process

D. Requirements Documentation

The purpose of requirements documentation is to communicate requirements (or knowledge sharing) between stakeholders and agile teams. In fact, no formal requirements specification is produced in agile development methods since agile focuses on minimal documentation. The features and the requirements are recorded on story boards, index cards, and paper prototypes like use cases and data flow diagrams.

The lack of documentation might cause long-term problems for agile teams [7], so, we suggest some techniques to solve this problem:

1) The agile team leader assigns two or three members to produce documentation in parallel and concurrence with development. The two (or three) members will be responsible for handling requirements (functional and non-functional requirements), writing, reviewing, and maintaining documentation consistent with development. Furthermore, efficient practices like peer interviews will help to ensure the accuracy and quality of the documentation. The reason for choosing two or three members is because the resources are limited and the other members must adhere to the agile manifesto of producing working software rather than documentation. In addition, we can not have just one person doing it, because that violates one of the agile manifesto principles [1] "Business people and developers must work together daily throughout the project".
2) Using computer-based tools like UML modeling and project management tools to specify a high level description of the project, and to document certain practices and requirements used in agile projects in an electronic format.
3) Developing a reverse engineering process [25] to be applicable on agile projects, so that we can use it to reverse engineer the code to produce

documentation using for example UML modeling tools.

E. Requirements Validation

The goal of requirements validation is to ensure that requirements actually define the system which the customer wants. The requirements validation checks the consistency, completeness and realism of requirements. The main practices used for requirements validation in agile approaches are:

1) **Requirements reviews:** it is a manual process that involves multiple readers from both agile team and stakeholders checking the requirements against current organizational standards and organizational knowledge for anomalies and omissions. In agile projects the requirements reviews must be formal reviews: we mean that the agile team should walk with the customers through each requirement; conflicts, errors, extra, and omissions in the requirements should be formally recorded.
2) **Unit testing:** In agile, unit testing is a method for requirements validation and therefore also part of requirements engineering. In some agile methods like XP, the requirements are implemented and tested using the TDD technique. By applying this technique developers create tests before writing code. The developed code is then refactored to improve its structure [32]; the rule here is to write a code if and only if a test fails. This technique has some advantages; it is the greatest advantage to set test cases that test your requirement very accurately. The requirement from which the test case was created is now presented in a form in which it is completely validated, in the sense that it can be automatically (after each iteration) determined whether a requirement is implemented by the software or not. This makes the developers aware for the progress of the project and the state of the current iteration of the project. Also, supports the refactoring process to get an improved design by reduced coupling and strong cohesion [26]. A common misconception is that all of the tests are written prior to implementing the code [9]. Rather, TDD contains short iterations which provide rapid feedback. Code refactoring and unit tests ensure that emerging code is more simple and readable. In fact, unit tests can be considered as a live and up-to-date documentation: they represent an excellent repository for developers trying to understand the system, since they show how parts of a system are executed.
3) **Evolutionary prototyping:** a prototype is an initial version of the system. Evolutionary prototyping starts with a relatively simple system which implements the most important customer requirements which are best understood and which have the highest priority. The system prototypes allow customers to experiment to see how the system supports their work (requirements elicitation), and may reveal errors and omission in the requirements which have been proposed (requirements validation). As shown in Fig. 4, the main objective of evolutionary prototyping in AD is to deliver a working system to customers by focusing on customer interaction, [24]. The verification (Are we building the system right?) and validation (Are we building the right system?) [27] of agile projects which have been developed using evolutionary prototyping can only therefore check if the system is adequate, that is, if it is good enough for its intended purpose; in other words, verification and validation of requirements in agile systems usually rely on the validation process.
4) **Acceptance testing:** acceptance testing is a formal testing conducted by the customer to ensure the system satisfies the contractual acceptance criteria. The acceptance tests are not different than the automated system tests, but they are performed by the customer. Delivering working software to the customer is a fundamental agile principle and hence. The customers create acceptance criteria for the requirements and test the requirements against these criteria. Being AD an incremental process, the customers can give feedbacks to the developers to enhance the development of future increments of the system. However, as a general problem there are often no formal acceptance tests for non-functional requirements.

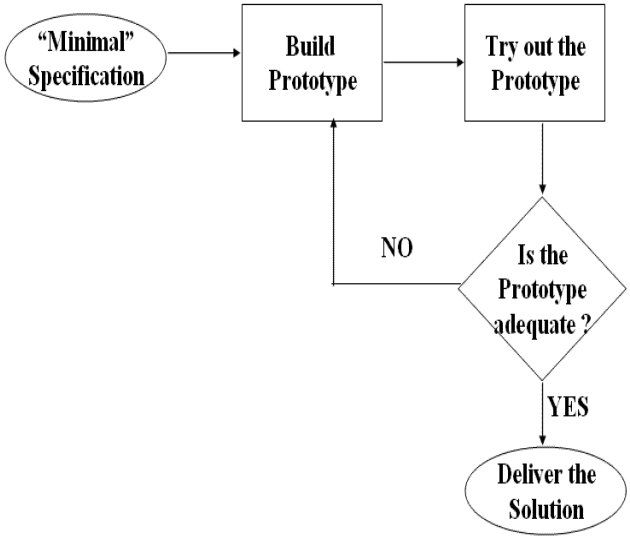


Figure 4 Evolutionary prototype processes

F. Requirements Management

Understanding and controlling changes to system requirements take place in this activity. In order for requirements management tools to work efficiently, "they must be able to store requirements, prioritize requirements, track requirement changes and development progresses, and provide a level of requirements traceability" [28, 29].

In agile projects, managers have to create and maintain a framework for the interaction between the agile teams and the stakeholders, by identifying the ideal people who

can be members of agile teams and ideal customers who can answer all the developers questions correctly [7], strengthening the collaboration, and negotiating contracts with the customers [8].

We believe that agile methods can play an important role in the management of large projects. The decomposition of the larger parts of the project into smaller components, called sub-components, lends itself to the employment of more agile teams. These agile teams can work in other time zones and other countries provided that frequent communications and self organization are established. Agile teams working in parallel on sub-components allows for quick development and an early design. An early design leads to an early review. Consequently, the iterative schedule and emphasis on delivering the product allows the agile teams to assess the successes and shortcomings, and plan for the next iteration. Once a specific agile team has successfully completed a sub-component, the team is available to work on another component or sub-component. Each of these smaller agile teams will still be responsible for assigning two members to complete the previously described documentation which is necessary to satisfy the other stakeholders.

Agile teams should use modern communications like web-based shared team projects and instant messaging tools; these tools are useful to keep in touch with the customer and other agile teams in order to discuss requirements when they are not on-site.

The ability to trace the software artifacts through the system lifecycle (source code, acceptance tests, requirements, and design logic) is critical to the success of large complex projects. Requirements traceability refers to the ability to describe and follow the life of a requirement, in both a forwards and backwards direction [30]. One of the problems is that traceability is an important part in traditional software development but it is not a standard practice for the agile methods. There are many techniques that have been presented to solve traceability issues. These techniques have been intended to work with traditional software development methodologies and therefore designed under the assumption that a formal requirements process is in place, but in agile software development the situation is different because the main development artifact in agile methods is a source code. In agile process, requirements, acceptance tests, unit tests and code change at the same time, so the unit tests should be traced to code, and the acceptance tests must include references to the requirements they test, see Fig. 5 [31].

As we say before, the main software development practice used in agile is TDD. The key aspect of TDD is that it can be viewed as a source of free traceability information. In turn, if such information is available to the developer, it may improve the efficiency with which tests are produced and code is written for each iteration. In TDD a traceability matrix is obtainable by matching new tests with changes in the code [31].

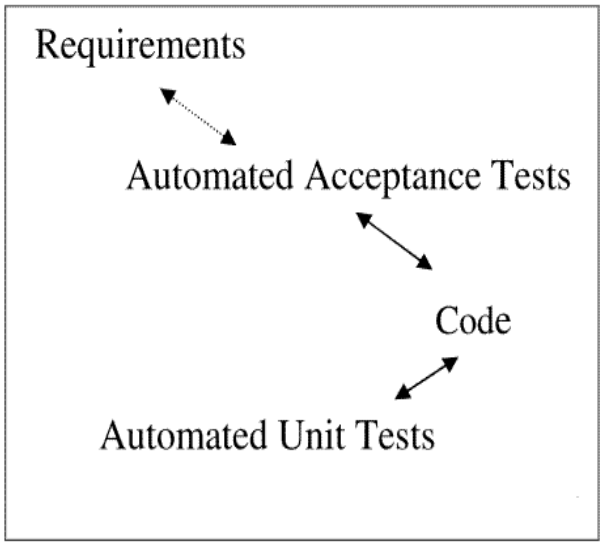


Figure 5 Traceability from requirements to code

Refactoring is an important aspect of TDD, but can represent a serious challenge to traceability. Refactoring of code may lead to the appearance of new traceability links and the disappearance of old traceability links between tests/requirements and code. Additionally, refactoring may lead to temporary code degradation, when some of the existing tests fail to pass. When refactoring, the TDD developer must ensure that all unit tests continue to pass, so unit tests might need to be refactored together with the source code.

IV. GUIDELINES FOR AGILE RE

This section introduces some guidelines to improve the performances of requirements engineering processes in agile environment and to enhance the quality of requirements.

- **Customer Involvement:** agile development focuses very strongly on customer interaction. At the beginning, all relevant ideal stakeholders have to be identified. Selecting the right customers and prioritizing their respective requirements is a key issue. The different elicitation practices aim to get as much knowledge as possible from all stakeholders and resolve inconsistencies.
- **Agile Projects Contracts:** at the beginning, the most critical requirements are expressed by the stakeholders as well as they can, so that the experienced project leaders can determine an initial cost for agile projects and guess the cost of later changes.
- **Frequent Releases:** frequently delivering parts of the system provides the ability to release faster expected results to the customers in order to get feedbacks from them. Hence, the requirements are implemented in an iterative and incremental fashion.
- **Requirements Elicitation Language:** use linguistic methods for requirements elicitation, derived from Natural Language Processing (NLP)

[7]. In other words, requirements are collected using the language of the customer, not a formal language for requirements specification.

- **Non-Functional Requirements (NFR):** in agile approaches handling of NFR is ill defined [9]. We propose the customers and agile team leaders to arrange for meetings to discuss NFR (and all critical requirements) in the earliest stages. Once the initial NFR of a project have been identified and documented, the agile teams can begin with development.
- **Smaller agile teams are flexible:** smaller agile teams allow continuous communications between them and stakeholders in efficient way, and the requirements changes are controlled. Fig. 6 shows that whenever the agile teams are smaller, the chances of the project success increased [16].

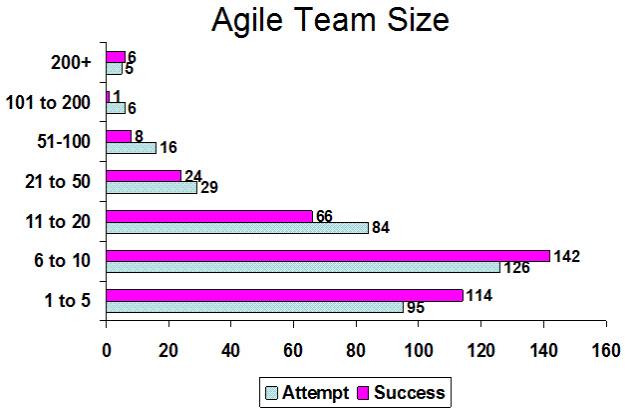


Figure 6 Agile team sizes

- **Evolutionary requirements:** RE in agile methods accommodate changing requirements even late in the development cycle, but that changes to the requirements must wait until the culmination of each iteration. Therefore, agile development does not spend much time in initial requirements elicitation. Consequently, this methodology will ensure that iterations are consistent with expectations, and that the development process will remain organized.
- **No early documentation:** any documents produced in the early stages can quickly become irrelevant because the agile principles encourage requirements change. By allocating only 5%-15% of the resources to requirements we think development team can still address shortcomings in agile development while complying with the agile principles in general.
- **Requirements splitting:** if the agile team considers a requirement too complex, this technique helps the customer to divide it into simpler ones. This helps agile teams to better understand the functionalities requested by the customer, and helps agile teams working in parallel with frequent communications between them. In XP [16], the requirements are written on story cards, the complex user stories are broken down smaller stories. Of course not all user stories can be divided since somecontain several sub-requirements, or record non-functional requirements. If a story card could be successfully divided, the original story card is discarded, since it no longer needed. All requirements are now included in the union of the new story cards.
- **Requirements Traceability:** a major upset in the development of large systems, especially those with evolving requirements is ensuring that the design of the system meets the current set of requirements. We are persuaded that agile projects would work better if they include requirements traceability tools together with validation tools. A good practice would be to identify the traceability links in TDD environment. In other words, the traceability links between test cases and related code should be identified and evolved to control co-changes. In this way, once the code is refactored, the agile team is able to re-build the traceability matrix again and determine what are the test cases needed to be re-run. In particular, the focus should be on the identification of the traceability links added or deleted after the refactoring process. In case the traceability links between source code and the related unit tests are broken during refactoring, this may be treated as a warning for possible code and/or unit test review [31]. Traceability information between requirements, source code and unit tests can also be used to drive software development, by identifying requirements for which unit tests and/or source code has not been implemented yet. In addition, traceability information can be used to support refactoring. Similar test cases can be grouped in test suite and traced onto source code classes. Source code classes related to more than one test suite are good candidates for refactoring.

V. CONCLUSION AND FUTURE WORK

The agile methodology manifesto supports a very efficient RE; this paper surveys the real process and activities of agile RE including feasibility study, elicitation, analysis, documentation, validation, and management. The secret of the success of agile RE is customer collaboration, good agile developers, and experienced project managers. This article provides some recommendations to solve the requirements documentati--on problem in agile projects, to make agile methodology suitable for handling projects with critical (functional and non-functional) requirements, to allow agile teams involved in large software projects to work in parallel with frequent communications between them. As future work, we will present industrial case studies that support our ideas, and try to develop a tool that support the distinction between functional and non-functional requirements; also we ignite debates for solving the traceability problem in TDD environment to re-establish

traceability after refactoring and to use traceability to improve refactoring.

AKNOWLEDGMENTS

We would like to thank Mr. Avishek Shrestha for his help, valuable ideas, and various references.

REFERENCES

[1] K. Beck, A. Cockburn, R. Jeffries, and J. Highsmith, (2001). "Agile manifesto". http://www.agilemanifesto.org, 23-02-2010.

[2] J. Highsmith, "Agile Software Development Ecosystems," *Addison–Wesley, Boston*, MA, 2002.

[3] J. Highsmith, K. Orr, A. Cockburn, "Extreme programming," *E-Business Application Delivery*, pp. 4–17, February, 2000.

[4] S. Bose, M. Kurhekar, J. Ghoshal, "Agile Methodology in Requirements Engineering," *SETLabs Briefings Online*. http://www.infosys.com/research/publica-tions/agile-requirements-engineering.pdf, February, 2010.

[5] A. Eberlein and J. Leite, "Agile Requirements Definition: A View from Requirements Engineering," Proceedings of the International Workshop on Time-Constrained Requirements Engineering (TCRE'02), Germany, 2002.

[6] R. Goetz, "How Agile Processes Can Help in Time-Constrained Requirements Engineering," *International Workshop on Time-Constrained Requirements Engineering*, 2002

[7] F. Paetsch, A. Eberlein, F. Maurer, "Requirements engineering and agile software development," *Eighth International Workshop on Enterprise Security, Linz, Austria*, 9 - 11 June, 2003.

[8] A. Sillitti, G. Succi, "Requirements Engineering for Agile methods," *Engineering and Managing Software Requirements, Springer*, 2005.

[9] B. Ramesh, L. Cao and R. Baskerville, "Agile requirements engineering practices and challenges: an empirical study," *Info Systems J*, "doi:10.1111/j.1365-2575.2007.00259.x", 2007.

[10] B. Rompaey and S. Demeyer, "Establishing Traceability Links between Unit Test Cases and Units under Test," *13th European Conference on Software Maintenance and Reengineering, Germany*, 2009.

[11] S. Ambler, "When does(n't) agile modeling make sense?,".http://www.agilemodeling.com/essays/whenDoesAMWork.htm, February, 2010.

[12] S. Ambler, "Agile Modeling: Effective Practices for Extreme Programming and the Unified Process," *John Wiley & Sons, Inc. New York*, 2002.

[13] S. R. Palmer and J. M. Felsing, "A Practical Guide to Feature-Driven Development," *The Coad Series*, 2002.

[14] Peter Coad, Eric Lefebvre and Jeff De Luca, "Java Modeling in Color with UML," *Prentice Hall PTR*, Chapter 6, 1999.

[15] Jennifer Stapleton, "DSDM - Dynamic System Development Method," *Addison-Wesley*, 1995.

[16] K. Beck, "Extreme programming explained. Reading, Mass," *Addison-Wesley*, 1999.

[17] Pekka Abrahamsson, Outi Salo, Jussi Rankainen and Juhani Warsta, "Agile software development methods - Review and analysis," *VTT Electronics*, 2002.

[18] K. Schwaber and M. Beedle, "Agile Software Development With Scrum," *Upper Saddle River, NJ: Prentice-Hall*, 2002.

[19] A. Polini, "Software Requirements," http://www1.isti.cnr.it/~polini/lucidiSE/Requirements1.pdf , February, 2010.

[20] A. Sidky and J. Arthur, "Determining the Applicability of Agile Practices to Mission and Life-Critical Systems," *Proceedings of the 31st IEEE Software Engineering Workshop, IEEE Computer Society*, pp. 3-12, 2007.

[21] T. Poppendieck and M. Poppendieck, "Lean Software Development: An Agile Toolkit for Software Development Managers," *Addison-Wesley*, 2003.

[22] O. Salo, "Systematical Validation of Learning in Agile Software Development Environment," *7th International Workshop on Learning Software Organizations, Kaiserslautern, Germany*, April, 2005.

[23] O. Salo and P. Abrahamsson, "Integrating Agile Software Development and Software Process Improvement: a Longitudinal Case Study," *4th International Symposium on Empirical Software Engineering, Noosa Heads, Australia*, November, 2005.

[24] I. Sommerville and P. Sawyer, "Requirements Engineering – A Good Practice Guide," *John Wiley & Sons*, 2000.

[25] E. Chikofsky and J. I. Cross, "Reverse engineering and design recovery: A taxonomy," *IEEE Software*, vol. 7, no. 1, pp. 13–17, 1990.

[26] K. Beck and M. Flower, "Planning Extreme Programming," *Addison-Wesley*, 2001.

[27] B. Boehm, "Verifying and validating software requirements and design specifications," *IEEE Software*, Vol. 1(1), 1984.

[28] L. Delgadillo, "Story-Wall: Lightweight Requirements Management for Agile Software Development," *15th IEEE International Requirements Engineering Conference*, pp.377-378, 2007.

[29] A. De Lucia, F. Fasano, R. Oliveto, and G. Tortora, "Recovering traceability links in software artifact management systems using information retrieval methods," *ACM Transactions on Software Engineering and Methodology*, vol. 16, no. 4, 2007.

[30] O. Gotel, and A. Finkelstein, "An Analysis of the Requirements Traceability Problem," *Proceedings of the IEEE International Conference on Requirements Engineering*, pp 94-101, April, 1994.

[31] J. Hayes, A. Dekhtyar and D. Janzen, "Towards Traceable Test Driven Development," *Proceedings of the 2009 ICSE Workshop on Traceability in Emerging Forms of Software Engineering*, pp.26-30, 2009.

[32] M. Fowler, "Refactoring: Improving the design of existing code," *Addison Wesley*, 1999.

A New Approach in Requirements Elicitation Analysis

William J. Tastle
Dept of Management, School of Business, Ithaca College
Ithaca, New York, USA
tastle@ithaca.edu

Amjad Abdullat
Dept of Computer Information Systems, West Texas A&M University
Amarillo, Texas, USA
aabdullat@wtamu.edu

Mark J. Wierman
Dept of Computer Science, Creighton University
Omaha, Nebraska, USA
wierman@creighton.edu

***Abstract* — In requirements analysis the task of elicitation of stakeholder need has been a continuing source of error and frustration in systems development. To aid in the acquisition of a set of proper needs that are critical to the design of an effective system, the systems analyst is provided with a new tool to assist in determining when group consensus has been met with respect to the identification of one or more needs. A recently developed measurement tool for measuring subjective concepts like consensus, agreement, and dissent is described. Categorical data are frequently collected using an ordinal scale such as the Likert scale and a new method is available that gives the analyst a different perspective of group-think. The agreement measure is also extended to an agreement distribution and used to calculate a mathematical distance between two separate agreement distributions. With these measures it is easy to calculate the proximity of agreement between two or more groups of stakeholders. This measure is then applied to requirements analysis.**

Keywords — requirements analysis, requirements elicitation, problem identification, agreement, consensus

I. INTRODUCTION

This paper builds on and extends a recently given paper at the International Conference on Information and Communication Systems in Jordan on December 2009. In the domains of software engineering, systems engineering systems analysis, and systems design, it is commonly acknowledged that difficulties in understanding challenging and complex problems, sometimes even perceived as being rather intractable, usually presage the more interesting efforts of building information systems in organizations. Determining the needs for new systems, improvements to existing systems or for new or altered products, can require the systems analyst to address conflicting requirements from the various stakeholders, each viewing the new (or revised) system from a particular perspective.

Being able to sift through the myriad of differing views to identify the set of necessary conditions, rather than a set of symptoms that actual problems may manifest, is critical to the success of the project. To accomplish this end, analysts are charged with the task of producing a set of requirements that are measurable, testable, actionable, and in sufficient detail such that system design can occur with the delivery of a system that truly does solve business needs. Mistakes made at this investigatory level and propagated through to the project conclusion result in very expensive corrections. Textbooks on systems analysis and design (and software engineering) usually identify requirements analysis as consisting of three types of activities: the eliciting of the requirements, the analysis of those requirements, and methods by which these requirements are recorded and documented.

Since the analysis process can be long, arduous and tempting of the patience of management, it is important to identify all the stakeholders (as well as those who move into, and out of, the organization), and take into account their needs while not detracting from the needs of others. There are many techniques used to elicit information from stakeholders and, in fact, this area of requirements analysis, that of *requirements elicitation*, is an area that transcends the domains of business, psychology, sociology, human relations, organizational behavior, statistics, and other disciplines as well as emerging disciplines. The list of domains from which the tools used in requirements elicitation are derived can be quite long, and can vary from project to project. This area of elicitation is fundamental to the entire systems development area, for an error made at this level of

doi:10.4304/jetwi.2.3.221-231

analysis is the most costly (in terms of money and time) to resolve later in the development process.

Requirements elicitation, sometimes referred to as requirements gathering, is the practice of obtaining requirements from the stakeholders (users, customers, clients, suppliers, other companies, etc). The process of elicitation is non-trivial because it is far more than merely the asking of a few questions. One can never really be sure one has all the requirements in spite of using interviews, questionnaires, personal observation, brainstorming techniques, use cases, prototyping and so forth. One frequently used method is the brainstorming technique in which all ideas, regardless of their accuracy or relevance, are encouraged. Unfortunately, those individuals higher in the pecking order sometimes monopolize the discussions until their view becomes the accepted view of all the stakeholders though there are methods by which this kind of behavior can be controlled (see [29] for examples). Discussions at the brainstorming table frequently lack any kind of measurable rigor for the intent is to encourage all ideas without placing anyone in an adverse position. Thus, ideas that might be shared by a large number of people may not be immediately recognized as belonging to a large constituency of stakeholders but rather in varying degrees of acceptance. We offer a method by which such discussions, as well as the results of questionnaires and surveys, can be quickly analyzed.

II. THEORY OF ELICITATION

The motivation for this study is to identify a function that can be easily computed and containing enough meaning such that all individuals engaged in a brainstorming session can easily express their feelings about any situation, idea, suggestion, plan, diagram, discussion, etc. without embarrassing anyone or causing a rift or conflict in the meeting or with one's colleagues. If an idea is being put forth by a stakeholder, and another stakeholder does not believe that the idea being put forth is relevant, or perhaps only practically relevant, an indication as to the degree of agreement can be determined by means of the recently developed agreement measure [14, 26].

Individuals naturally compartmentalize their thinking into categories or fuzzy numbers [30]. A fuzzy number is a convex, normalized fuzzy set whose membership function is at least segmentally continuous. An entire discipline has evolved around fuzzy sets, numbers and systems as reflected in numerous journals (e.g., Fuzzy Sets and Systems, J of Approximate Reasoning). A physician might ask a patient about the degree of pain associated with an injury by giving a number between 1 and 10. The physician has no expectation that the response from the patient is anything more than a categorical estimate, essentially a fuzzy number.

Humans simply do not have the ability of ranking or evaluating anything along such the continuous number line unless they use some piece of equipment by which their perceptions can be augmented. It is the use of categories that people are comfortable utilizing as in their propensity to agree or disagree, with an issue. Human vocabulary is replete with categorical rankings: cold, warm, tepid, hot; low vs. high; innocent vs. guilty. Thus, stakeholders around a table engaged in dialogue or brainstorming react to ideas in the form of categories: they will strongly agree, moderately agree, moderately disagree, or strongly disagree with what is on the table at the moment. If all the stakeholders are in complete agreement with the idea being presented, then the overall consensus around the table should be 100%. If, on the other hand, the stakeholders are completely split with half being in strong agreement and the other half in strong disagreement, then the overall consensus around the table should be 0%. Every other possible category assignment by the stakeholders would be somewhere between these extremes. Thus an interval exists between 0 and 1, or 0% and 100%. This is simply to understand and offers a quick guide as to the degree to which the group is in consensus or, stated another way, the degree to which the stakeholders are in agreement as to a need the new system must address.

It is not required that the stakeholders agree completely on an idea in order to accept it for preliminary analysis shows that a consensus of 80% approximates 95% significance [28]. What is important is that stakeholders can discuss their perceptions of system needs in an open and non-threatening environment but can also "quietly" indicate their level of support to the current idea.

It is difficult enough to get stakeholders to a table to discuss their perceptions of system inadequacies or new system requirements, let along control the discussions along some particular direction. The systems analyst must engage in this kind of activity and do so in an efficient manner. By using this measurement tool it may be possible to direct, in real time, discussions on ideas that appear to have stronger merit based on a consensus calculation. It is to this concept that the remaining portion of the paper is directed.

III. THE CONNECTION BETWEEN CONSENSUS AND ORDINAL SCALES

The basis for the study of uncertainty, in the sense of imprecision, was first established by Zadeh [24] when he characterized grades of set membership by a function that assigned a value between zero and one to each member. The succeeding papers (many hundreds of them that have since spawned the discipline of fuzzy sets and fuzzy measures (now sometimes referred to as general measures)) have further developed the new discipline, but the overwhelming majority or papers have dealt with the collection of data based on interval and/or ratio scales although laudable efforts were made to connect, for example, category theory and systems theory [2] and fuzzy clustering to ordinal and nominal scales [3].

The original purpose in studying fuzzy sets as they relate to ordinal scales was to establish a connection between group consensus making and a calculable measure. It was intuitively felt that the various forms of fuzzy set and fuzzy

measure theory could be the basis for such a measure. The result of the initial meeting was the consensus measure (see below). In the original formulation an "average" was calculated (the mean), but after further study it was decided that the median was more conceptually correct when dealing with ordinal measures than the mean, for an mean calculation (addition and division) is predicated on the values being summed, and then divided by the number of values, properties that properly belonged to the interval (or ratio) scale and was thus inappropriate for the consensus measure.

The difficulty in using interval and ratio scale measures on ordinal scales lies in the use of permissible statistics (see figure 9 in the Appendix). While the median and percentile are permitted (and the mode and chi square, permissible on nominal scales and thus permitted on higher-level scales), the mean, standard deviation, and other statistical measures are limited to at least interval scales and hence not permitted on nominal or ordinal scales. It is unfortunate that means are regularly applied to ordinal scale data, typically in the form of evaluations of Likert measures, no evidence exists to support an existence of a regular interval between ordinal categories, though totally ordered and monotonically increasing; the literature, however, is replete with examples (none are specifically cited for the purpose of propriety) of studies that assume an interval is the absence of evidence [8].

There are several examples below to illustrate the value of using consensus theory measures to evaluate ordinal data.

IV. Typical Methods of Ordinal Scale Analysis

It is typical for researchers to apply the mean to ordinal data, for the available statistical tools are quite limited. There are nonparametric tests (see any introductory statistics text) that can be used to make sense of ordinal data (e. g., Wilcoxon signed ranks test, Spearman test, Gamma coefficient, and the like), but all these methods were devised before the onset of fuzzy set theory, fuzzy measure theory and information-theoretic measures. These are the basis for the mathematics of vagueness and impreciseness, two qualities that are present in most systems investigations undertaken by the systems analyst as one pursues the requirements identification phase. Papers abound (see example below) in information systems (as well as other disciplines) in which Likert data (based on the selection of a category from an ordered sequence such as "strongly agree," "agree," "neutral,", "disagree," and "strongly disagree" which is represented in this paper as SA, A, N, D, and SD, respectively) is presented in terms of a category mean, standard deviation, confidence intervals, t-test, and other such statistics. The authors argue this is equivalent to saying that the average of warm and hot is warm-and-a-half. Ordinal scales possess no inherit interval between categories. Sometimes researchers place category labels on a number scale, i.e., SA = 1, A = 2, etc. to give an impression that they are interval. These are called *Likert-like* measures but are actually ordinal in disguise. To paraphrase an old phrase, painting stripes on a horse does not make it a zebra! We offer a different method that is conceptually sound and mathematically proven [12-19].

V. Measures of Consensus, Dissent and Agreement

A. Concept

The underlying concept behind the measures of consensus and agreement, and their complementary measures of dissent and disagreement, respectively, (disagreement is not discussed in this paper) is centered on the existence of a perceived relative distance between ordered categories (called the intra-categorical distance) that may or may not be equal, and may or may not be similar to the distances in the minds of others, but the distance from one extreme category to the other extreme category is always 100% of whatever the mindset. Hence, from SA to SD the overall intra-categorical distance is 100%; from cold to hot is 100%. Given a stakeholder team or an even number of people, if half select SA and the other half select SD, then the group consensus should be zero, for the group is equally partitioned at their extremes (SA and SD). Similarly, because the group is at maximum opposition the dissent should also be maximized at 100%. Consensus and dissent are measures that characterize the entire set of stakeholders and are thus measures of the collective and are directly related.

This is similar to the Congress of the US in which the principal two parties (Democrats and Republicans) each hold half of the membership. A consensus may never be attained. If one person moves from SD to D, or from SA to A, then the consensus should increase to some value above zero for the group is no longer balanced on the extremes. A consensus does not require 100% agreement, and it is usually the committee chair who must recognize when a consensus has been met in order to move the group on, but how does the committee chair (or the systems analyst) know that point has been reached? This is a matter for the psychologists and sociologists to research, but we can establish a criterion *a priori*. A percentage value that determines the threshold for consensus, from 0 to 100%, should be agreed upon before this analysis is applied.

If is reasonable to assume that a consensus is represented by a super majority, hence 51% probably does not represent a consensus. Clearly a consensus is met when 100% of the participants agree on a single Likert category, be it to agree or disagree with the statement under review. A group could even form a consensus around neutral in the sense that they have come to an agreement that they are all unsure. It is the selection of a number in the gray area between 50% and 100% that is the challenge. The US Senate requires a 60% super majority to pass legislation, and that value could be used to indicate a consensus. A recent study [11] on the establishment of a curriculum for dermatology students has used 80% as an indicator of consensus among the dermatol-

ogy medical community as to the importance of certain items being in a basic curriculum for dermatology students. Whatever the threshold, the following procedure can be used to determine the degree of consensus (and/or dissent) of the group towards a Likert statement.

B. Measure of Consensus

A consensus measure has been introduced, desirable properties have been proven, and applications have been demonstrated [12-21]. It is starting to receive outside attention.

The measure of consensus, whose form was inspired by the Shannon entropy, requires an ordinal scale, but it can be used with interval and ratio scales, though the standard statistical measures are probably a better choice in those particular cases. The equation for consensus is:

$$Cns(\mathbf{X}) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu_x|}{d_x} \right) \quad (1)$$

where X represents the list of categories (Strongly Agree (SA), Agree (A), Neutral (N), Disagree (D), and Strongly Disagree (SD)), X_i is an element of X, μ_X is the mean of X and d_X is the width of X, $d_X = X_{\max} - X_{\min}$. Let us assume that we have a five-attribute Likert scale: SA, A, N, D, and SD. Let us further assign an arbitrary numerical scale of SA = 1, A = 2, N = 3, D = 4, and SD = 5. Then X = {1, 2, 3, 4, 5} and X_1 = 1, X_2 = 2, etc. The width of X, d_X, is $X_{\max} - X_{\min}$. In this case d_X = 5 - 1 = 4. The mean, or expected value, of X is given by the usual formula $E(X) = \sum_{i=1}^{n} p_i X_i = \mu_X$.

C. Measure of Dissent

Let X be a discrete random variable of size $n > 2$ with probability distribution $p(x)$. As usual $E(X) = \sum_{i=1}^{n} p_i X_i$ is the mean μ_X of X. Let $d_X = X_{\max} - X_{\min}$ be the width of X and set $d_i = |X_i - \mu_x|$ as the absolute deviation of X from the mean. The Dissention, $\mathbf{Dnt}(\mathbf{X})$ is then defined to be

$$\mathbf{Dnt}(\mathbf{X}) = -\sum_{i=1}^{n} p_i \log_2 \left(\frac{d_X - d_i}{d_X} \right) = -\sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu_X|}{d_X} \right)$$

If there is no chance of confusion then we will drop the subscripts and write:

$$\mathbf{Dnt}(\mathbf{X}) = -\sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu|}{d} \right) \quad (2)$$

If $n = 1$ then there is no dissention and we will set $\mathbf{Dnt}(\mathbf{X}) = 0$.

$Cns = 1 - Dnt$ or

$$\mathbf{Cns}(\mathbf{Y}) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|Y_i - \mu_Y|}{d_Y} \right) \quad (3)$$

Table 1 A simulation of categorical assignments made by a group of 10 stakeholders. Consensus is compared to the mean and standard deviation.

	SA	A	N	D	SD	Cns	Mean	Std Dev
1	7	3	0	0	0	0.838	1.3	0.458
2	0	0	0	3	7	0.838	4.7	1.458
3	6	4	0	0	0	0.815	1.4	1.490
4	5	5	0	0	0	0.807	1.5	0.500
5	8	1	1	0	0	0.802	1.3	0.640
6	0	0	8	1	1	0.802	3.3	0.640
7	7	2	1	0	0	0.773	1.4	0.663
8	5	4	1	0	0	0.760	1.6	0.663
9	6	3	1	0	0	0.759	1.5	0.671
10	8	1	0	1	0	0.703	1.4	0.917
11	5	4	0	1	0	0.693	1.7	0.900
12	7	2	0	1	0	0.685	1.5	0.922
13	6	3	0	1	0	0.682	1.6	0.917
14	7	0	3	0	0	0.649	1.6	0.917
15	8	0	1	1	0	0.637	1.5	1.025
16	5	4	0	0	1	0.577	1.8	1.166
17	7	0	2	1	0	0.569	1.7	1.100
18	6	3	0	0	1	0.548	1.7	1.187
19	6	0	3	1	0	0.533	1.9	1.136
20	7	2	0	0	1	0.532	1.6	1.200
21	5	0	4	1	0	0.528	2.1	1.136
22	8	1	0	0	1	0.527	1.5	1.204
23	7	0	0	3	0	0.420	1.9	1.375
24	8	0	0	1	1	0.403	1.7	1.418
25	8	0	0	0	2	0.278	1.8	1.600
26	6	0	0	3	1	0.258	2.3	1.616
27	5	0	0	4	1	0.251	2.6	1.625
28	7	0	0	1	2	0.210	2.1	1.700
29	7	0	0	0	3	0.119	2.2	1.833
30	6	0	0	1	3	0.101	2.5	1.857

D. Measure of Agreement

The measure of agreement has been previously introduced [14-15] and shown to be a method by which differing opinions of stakeholders can be justifiably assembled to

yield a single value upon which there exists maximal agreement [19]. This does not mean that all stakeholders need to have selected a particular category, for it is possible to agree on a category to which no one has made a selection. For example, row 21 in Table 1 shows a mean of 2.1, or 1/10 of the way between Agree and Neutral. If the Agree category is selected, it is apparent that the category in greatest agreement is that one which no one selected. Having an "average value" (see the *mean* column in table 1) refer to a category that reflects a zero frequency is not uncommon (note rows 23-30 with a mean representing a category with an assignment of zero).

We define agreement as a harmony of opinion or action. To attain a harmony of opinion does not require that all involved individuals express the same view.

The consensus measure was found to be modifiable to determine the degree of agreement associated with every frequency in a given distribution by assigning each category to the position of the *mean* in equation (1). In other words, consensus depends on the calculation of a mean value against which the distance of each category from that mean value is calculated. Hence the consensus is a measure of the degree of attraction to a mean value. This equation is modified slightly to calculate the degree of attraction to each individual category value. It should be noted that in lieu of the mean it is appropriate to use the median. In fact, the median is more conceptually accurate than the mean when working with ordinal scales.

The desire to *target* the consensus led the authors to examine different expressions for the log term. This has led to the development of an Agreement measure:

$$\mathbf{Agr}(\mathbf{X},\tau) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \tau|}{2d_X} \right) \qquad (4)$$

τ represents the target category such as SA, A, etc. and the denominator is changed to $2d_X$ to reign in the range of the measure.

VI. Example

A. Illustration of Consensus

Assume a group of 10 stakeholders in a meeting conducted by a systems analyst for the purpose of defining the problem that needs to be addressed. Each stakeholder has an interest in having the eventual solution satisfy their particular set of needs, and no individual stakeholder has full knowledge of the extent of the problem under investigation. It is up to the systems analyst to determine the actual nature of the problem. Suppose the analyst begins by raising a series of issues that seem to be plausible indicators of the problem, and each stakeholder is asked to respond to the statement indicating their level of agreement. A Likert scale is chosen and the questions are similar in format to the following example:

The problem centers on the inability of our customers to adequately access our web site to initiate orders.

For purposes of illustration let us select a subset of possible responses that compare the consensus with the traditionally calculated mean and standard deviation.

The mean is expectedly different for each of these frequencies indicating the average value without any dispersion as evidenced by the standard deviation (see Table 2). The consensus yields a value of 1.0 to show that these frequencies are in complete agreement. However, means (μ) and standard deviations (SD) do not convey a sense of agreement in the same way as the consensus measure (Cns).

Table 2 Three frequency distributions centered on strongly agree, neutral, and strongly disagree respectively, with the consensus, mean and standard deviation for each distribution.

SA	A	N	D	SD	Cns	μ	SD
10	0	0	0	0	1.000	1	0
0	0	**10**	0	0	1.000	3	0
0	0	0	0	**10**	1.000	5	0

Table 3 shows another possible distribution of responses from our group of 10 stakeholders. Note that the mean and standard deviation properly and accurately depict the information gathered by these distributions, but the interpretation of these results is in question. Each distribution has the same mean of 3.0 = Neutral, but the standard deviations show different dispersions. Reflecting only the meaning of the standard deviation, the reader visualizes a normal distribution with a slightly wider stance in the left and right legs of the curve. There is nothing special in this statistical interpretation. However, the consensus measure of 0.585. or 58.5%, induces a mental image of a consensus that reflects too much dissention to be acceptable.

Row 2 of table 3 shows an even further dispersion as the consensus is 0.0 or 0%. This is the situation when neither side accepts the argument of the other. This gridlock is serious and reflects a situation that the analyst needs to further investigate. When the stakeholders differ in their perception of the problem at this great a degree, the actual problem is likely to be something other than that being presented.

Table 3. Two possible distributions having the same mean value, with increasing standard deviations, and consensus values that tell strikingly different stories.

	SA	A	N	D	SD	Cns	Mean	Std Dev
1	0	5	0	5	0	0.585	3.0	1.054
2	5	0	0	0	5	0.000	3.0	2.108

Table 4 shows four different distributions that contain identical pairs of means, standard deviation and consensus measures. Note that the presence of a mean does not require the presence of any values in that category (e.g., row 1 has a mean of 2 = Agree, but the frequency for that element is zero. Means can change and still retain the same standard deviation. The consensus measures show about a 43% consensus around the frequencies in Rows 1 and 2, and a 31% degree of consensus in Rows 3 and 4. Since these frequencies are symmetrical to each other (row 1 and row 2 are reversals of each other), it should be expected that the degree of consensus is the same. What matters is whether or not a degree of consensus has been met to move forward in identifying the problems to be addressed in the new system.

Table 4 Pairs of consensus', means, and standard deviations associated with different distributions.

	SA	A	N	D	SD	Cns	Mean	Std Dev
1	6	0	3	0	1	0.426	2.0	1.414
2	1	0	3	0	6	0.426	4.0	1.414
3	7	0	0	2	1	0.309	2.0	1.633
4	1	2	0	0	7	0.309	4.0	1.633

	SA	A	N	D	SD	Agr(SA)	Dnt	Cns	Mean	StDev
1	0	0	0	0	15	0.000	0.000	1.000	5.000	0.000
2	0	0	0	1	14	0.021	0.048	0.952	4.933	0.249
3	0	0	0	2	13	0.043	0.089	0.911	4.867	0.340
4	0	0	0	3	12	0.064	0.124	0.876	4.800	0.400
5	0	0	0	4	11	0.086	0.151	0.849	4.733	0.442
6	0	0	1	4	10	0.125	0.213	0.787	4.600	0.596
7	0	0	2	4	9	0.164	0.259	0.741	4.467	0.718
8	0	0	3	4	8	0.203	0.291	0.709	4.333	0.798
9	0	1	3	4	7	0.257	0.347	0.653	4.133	0.957
10	0	2	3	4	6	0.310	0.389	0.611	3.933	1.062
11	1	2	3	4	5	0.377	0.490	0.510	3.667	1.247
12	3	3	3	3	3	0.543	0.566	0.434	3.000	1.414
13	5	4	3	2	1	0.709	0.490	0.510	2.333	1.247
14	6	4	3	2	0	0.775	0.389	0.611	2.067	1.062
15	7	4	3	1	0	0.820	0.347	0.653	1.867	0.957
16	8	4	3	0	0	0.866	0.291	0.709	1.667	0.798
17	9	4	2	0	0	0.893	0.259	0.741	1.533	0.718
18	10	4	1	0	0	0.921	0.213	0.787	1.400	0.596
19	11	4	0	0	0	0.949	0.151	0.849	1.267	0.442
20	12	3	0	0	0	0.961	0.124	0.876	1.200	0.400
21	13	2	0	0	0	0.974	0.089	0.911	1.133	0.340
22	14	1	0	0	0	0.987	0.048	0.952	1.067	0.249
23	15	0	0	0	0	1.000	0.000	1.000	1.000	0.000

Table 6 Twenty-three distributions from a stakeholder group of 15 individuals with agreement (Agr) calculated on the Strongly Agree (SA) category.

As the distributions become more random in the permutations of stakeholder values, making sense of the means and standard deviations become increasingly difficult. Table 1 above shows a subset of randomly selected rows, sorted by consensus, then by mean.

If we assume that Salmoni [11] is correct and we accept 80% as the threshold for consensus, it is easy to justify rows 1 through 6 in Table 1 as being acceptable to the systems analyst. While we can accept that the group of stakeholders has arrived at a consensus, it is still a matter of interpretation as to the winning category. There is also the matter of those means that are midway between categories, like rows 4, 9, 12, 22, and 26. What is the arbiter of the consensus category in these situations? While the consensus measure gives a theoretical justification for having attained consensus, it does not permit the unambiguous selection of a particular category of ownership (see table 5 for the identification of first 10 rows

Row	Category
1	SA
2	SD
3	SA
4	SA or A
5	SA
6	N
7	SA
8	A
9	SA or A
10	SA

Table 5. A subjective determination of category based on the mean.

of table 1 into categories based on the mean value). To solve this problem we must modify the consensus equation to permit us to calculate the amount of agreement the group of stakeholders has for each individual category.

B. Illustration of Dissent

An examination of Table 6 shows another set of 23 distributions that could be elicited from some group of 15 stakeholders. For each distribution the agreement (Agr), dissent (Dnt), consensus (Cns), mean and standard deviation (StDev) are calculated.

For each value associated with Cns the value of Dnt can be seen to equal 1 – Cns (see equation 3 above). Note that the consensus is maximized in rows 1 and 23 and minimized, in this particular set of distributions, at row 12. Since 15 cannot be evenly divided by 2, it is not possible to have a zero consensus. The dissention can be interpreted as a measure of dispersion around the mean. Dissention as a measure of dispersion gives no new information to the statistical measure of the standard deviation except to say that it is far easier to understand. Figure 1 (see Appendix) shows the linear relationship between dissent and the standard deviation. The covariance of Dnt and StDev in Table 6 is 0.0635, and the R^2 is 0.9913 (see Figure 1 after references). Dissention as a measure of dispersion gives no more information than the standard deviation except that it is a far easier statistic to understand.

Since both measures give the same mathematical sense of variance (dispersion), one could use either one. The authors argue that it is easier to visualize and understand dispersion when it is represented in terms of a percentage than by a number that grows with the magnitude of the numbers that comprise the calculation, and since the calculation does not assume an interval value, it is the more conceptually correct measure. Thus, given our group of stakeholders, it is likely that everyone will better understand the dispersion in their ratings when it is represented in row 11 of table 6 as 49% rather than the 1.291 given by the calculated standard deviation; if the numbers in the distribution changed from {1,2,3,4,5} to {100,200,300,400,500} the standard deviation would change from 1.464 to 146.385. Again, the value of the standard deviation is driven by the size of the numbers used in the calculation while the dissent (as dispersion) is a percentage that retains a constant mental reference of a value between 0-100%.

C. Agreement

Table 6 shows the Agreement based only on the strongly agree category, and table 7 shows the Agreement for each category in the grayed row below its frequency distribution. Each gray row is called a *set of agreement measures* (that can also be considered an *agreement 5-tuple*). Since we assume that the survey data represents an entire population (like a group of stakeholders attempting to identify a systems problem) we make no claims with respect to samples and populations and leave that area for future research activities.

Table 7 Agreement calculated for each response.

	SA	A	N	D	SD
1	0	0	0	0	15
$Agr_1(\tau)$	0.000	0.322	0.585	0.807	1.000
2	0	0	0	1	14
$Agr_2(\tau)$	0.021	0.339	0.600	0.820	0.987
3	0	0	0	2	13
$Agr_3(\tau)$	0.043	0.357	0.615	0.833	0.974
4	0	0	0	3	12
$Agr_4(\tau)$	0.064	0.375	0.629	0.846	0.961
5	0	0	0	4	11
$Agr_5(\tau)$	0.086	0.392	0.644	0.859	0.949
6	0	0	1	4	10
$Agr_6(\tau)$	0.125	0.424	0.672	0.859	0.921
7	0	0	2	4	9
$Agr_7(\tau)$	0.164	0.457	0.700	0.859	0.893
8	0	0	3	4	8
$Agr_8(\tau)$	0.203	0.489	0.727	0.859	0.866
9	0	1	3	4	7
$Agr_9(\tau)$	0.257	0.534	0.742	0.844	0.820
10	0	2	3	4	6
$Agr_{10}(\tau)$	0.310	0.580	0.757	0.829	0.775
11	1	2	3	4	5
$Agr_{11}(\tau)$	0.377	0.612	0.757	0.797	0.709
12	3	3	3	3	3
$Agr_{12}(\tau)$	0.543	0.704	0.757	0.704	0.543

Notice row 12 in table 7. The agreement is not equidistributed over the categories because each category in this ordered set gives strength to those other categories contiguous to itself. Hence, the three individuals who selected SA would also permit the selection of A. Those who chose N are strengthened by those who selected A and D. Row 1 show the strongest possible agreement with the SD category,

and the agreement drops off rather quickly. If we adopt Salmoni's [11] threshold of 80% for acceptance of agreement, then row 10 shows an agreement in support of the Disagree category even though more people selected SD.

This measure is particularly useful when soliciting the response from a set of knowledgeable individuals who are acting separately in making a categorical assignment. It is apparently not atypical to make a solicitation of experts[1] and then to make a claim that the experts assert a particular position or category. With this new agreement measure it is possible to assign a real number to that level of agreement. This was particularly significant in an application of this measure to the assignment of colors to the terrorist threat levels (green, yellow, orange, and red) [19]. One problem with that application was the inability to map changes in the agreement distribution over time. That is to say, as the agreement frequencies changed over time, so would the agreement measures associated with each category, but how could one follow the degree by which the change occurred? It is now possible to measure the distance between these agreement distributions.

D. Measuring Distance between Agreement 5-Tuples

Given two frequency distributions (see table 8), F_1 and F_2, for which the agreement distributions, Agr_1 and Agr_2, are calculated using equation (5), a distance between the distributions can be determined. For each category in each frequency distribution there is a corresponding agreement value (see figure 2 for an illustration). The distance is calculated using

$$c_n \sqrt{\sum_{i}^{n} (Agt_1^i - Agt_2^i)^2} \qquad (5)$$

where n = the number of categories, c_n is a constant for each n (for a five category Likert scale $c_n = 0.63612$), Agr_1 uses probabilities derived from the frequencies F_1, and Agr_2 uses probabilities derived from the frequencies F_2. Agr (X, X_i) is the Agreement of the categories X with the i^{th} category.

The illustration in row 1 of Table 8 shows the maximum possible distance between two distributions in which the survey participants have chosen extreme positions.

		SA	A	N	D	SD	Dist
1	F_1	0	0	0	0	5	1.000
	F_2	5	0	0	0	0	
2	F_1	0	0	0	1	4	0.957
	F_2	5	0	0	0	0	
3	F_1	0	0	1	1	3	0.864
	F_2	5	0	0	0	0	
4	F_1	0	0	1	2	2	0.830
	F_2	5	0	0	0	0	
5	F_1	0	0	1	2	2	0.777
	F_2	4	1	0	0	0	
6	F_1	0	0	1	2	2	0.670
	F_2	3	1	1	0	0	
7	F_1	0	0	1	2	2	0.512
	F_2	2	1	1	1	0	
8	F_1	0	0	1	2	2	0.283
	F_2	0	1	2	2	0	
9	F_1	0	0	1	2	2	0.063
	F_2	0	0	2	1	1	
10	F_1	0	0	1	2	2	0.000
	F_2	0	0	1	2	2	

Table 8. Computed distance between agreement distributions. First column indicates a row number (from 1 to 10), column 2 distinguishes the two rows of distributions, columns 3-7 represent the values in the Likert distribution, and the last column is the calculated distance.

Row 1 of table 8 shows two frequency distributions, each of five stakeholders. The agreement distribution for each row is calculated (not shown in this table) and compared to the agreement distribution for the other row in this couple. Each row in table 8 contains two distributions, a "top" distribution and a "bottom" distribution. It is the agreement distribution for each of these that are compared and a distance calculated. Thus, the agreement distances become smaller as the agreement values become equal (see rows 8-10). It is important to understand that actual frequency values are not compared, rather, the agreement measure calculated on those frequencies. This permits us to calculate a distance without regard to the number of items constituting the frequency distribution.

VII. CONCLUSION

One of the most challenging areas in requirements determination is the identification and selection of the actual problem to address. The challenge is compounded in that finding solution methods frequently means crossing disciplines from organizational behavior to psychology, from statistics to measure theory, from the structured to the subjective. Currently available statistical tests such as the Wilcox-

[1] Private discussions with corporate representatives at the Risk Symposium 2008, Santa Fe, NM, 11-13 March 2008.

on signed rank test, are commonly used, but the underlying concept for their use is circumspect. The use of the mean and the standard error assumes an interval measure, something for which the evidence is, at best, sketchy using a Likert or similar scale. Scientists have, nonetheless, been using these kinds of tests since they were created in mid-1940 with good results. Now there are other concepts that have been developed such as the Dempster-Shafer evidence theory in 1976 and fuzzy set theory by Lotfi Zadeh in 1965. Building upon the works of others produced the concept of applying a variation from information theory to what is presented here as consensus theory. This method offers a different way of examining ordered category data, not to exclude or otherwise dispose of the currently used measures, but rather as an additional tool that brings a different view of categorical data analysis.

Beginning with the basic concept of deriving a measure of consensus given an ordered set of choices, such as the Likert scale, the concept was extended to provide a degree of focus on each category in a distribution (the agreement and dissent measures), and finally to being able to make a reasonable calculation of distance between two such distributions (the agreement distribution and the agreement distance). Using the agreement distance it is possible to easily track the success (or lack of success) in securing an agreement among stakeholders as to a particular position and when to claim that there is sufficient consensus from the stakeholders to accept the position and move onto another item.

REFERENCES

[1] Antunes, P., Costa, C.J. and Pino, J.A. (2006) "The use of genre analysis in the design of electronic meeting systems" *Information Research*, **11**(3) paper 251 [Available at http://InformationR.net/ir/11-3/paper251.html].

[2] Arbib, M. A. and Manes, R. G. (1975) "A Category-Theoretic Approach to systems in a Fuzzy World," Synthese, 30(3-4), pp. 381-406.

[3] Benedictis, L D and Maio, M D (2009) "Economists' View about the Economy. Evidence from a Survey of Italian Economists," Centre for Economic and International Studies, University of Rome.

[4] Brouwer, R. (2007) "Fuzzy Clustering of Feature Vestors with some Ordinal Valued Attributed using Gradient Descent for Leanring," Unternational Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 16(2), pp. 195-218.

[5] ComputerWorld.com 2008 "Top 10 Corporate Information Technology Failures," accessed 2008 at http://www.computerworld.com/computerworld/records/images/pdf/44NfailChart.pdf.

[6] Dennis, Alan and Wixom, Barbara H. (2000) Systems Analysis and Design: An Applied Approach. John Wiley & Sons, New York.

[7] Farley, J and Ceroni, M (2007) "Redefining "Productive": Implications for Sustainable Forest Management," Northeastern States Research Cooperative, www.uvm.edu/envnr/nsrc.

[8] Jamieson, S. (2004) "Likert scales: How to (ab)use them?." Medical Education, 38, pp. 1217-1218.

[9] Loughman, T., Fleck, Jr,, R. and Snipes, R. 2000 "A cross-disciplinary model for improved information systems analysis," Journal of Industrial Management & Data systems, 100(8), pp. 359-369.

[10] Saldarini, R., (1989) Analysis and Design of Business Information Systems, Macmillan, New York.

[11] Salmoni, A. J., Coxall, S., Gonzalez, M., Tastle, W., and Finley, A. (under review) " Defining a postgraduate curriculum in dermatology for general practitioners: a needs analysis using a modified Delphi method."

[12] Sidky, A. S., Sud, R. R., Bhatia, S., and Arthur, J. D. 2002 "Problem Identification and Decomposition within the Requirements Generation Process," Technical Report TR-02-31, Computer Science, Virginia Tech.

[13] Stevens, S. S., 1946, "On the theory of scales of measurement. *Science*, 103, pp. 677-680.

[14] Tastle, W. and Wierman, M. (2005b) "Consensus and Dissention: Theory and Properties," North American Fuzzy Information Processing Society (NAFIPS) Conference, Ann Arbor, MI.

[15] Tastle, W. and Wierman, M. (2005c) "A Tool for the Analysis of Ordinal Scale Sate: Measuring Consensus, Agreement, and Dissent," 5th International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands.

[16] Tastle, W. and Wierman, M. (2005d) "Measuring Consensus in Team Settings Doing Non-routine Knowledge Work," 4th AIS Symposium on Research in Systems Analysis and Design, Univ of Cincinnati [2005], April 23-24, 2005, Univ of Cincinnati.

[17] Tastle, W. and Wierman, M. (2006b) "An Information Theoretic Measure for the Evaluation of Ordinal Scale Data," Journal of Behavior Research Methods, 38(3), pp. 487-494, August 2006.

[18] Tastle, W. and Wierman, M. (2006c) "Consensus and dissention: A measure of ordinal dispersion," International Journal of Approximate Reasoning, 45(3), pp. 531-545.

[19] Tastle, W. and Wierman, M. (2007) "Determining Risk Assessment Using the Weighted Ordinal Agreement Measure," Journal of Homeland Security, http://www.homelandsecurity.org/newjournalArticles/displayArticle2.asp?article=157, June 2007.

[20] Tastle, W. and Wierman, W. (2005a) "Consensus and Dissention: A New Measure of Agreement," North American Fuzzy Information Processing Society (NAFIPS) Conference, Ann Arbor, MI.

[21] Tastle, W. and Wierman, W. (2006a) "Adjusting the Consensus Measure to Satisfy Ordinal Scale Arguments," North American Fuzzy Information Processing Society (NAFIPS 2006) Conference, Montreal, Canada.

[22] Tastle, W., Russell, J. and Wierman, M. (2005) "Ordinal Scales: Using the Consensus Measure to Compare Likert Scale Data," Information Systems Education Conference (ISECON 2005), Columbus, Ohio.

[23] Tastle, W., Wierman M. and Dumdum, U. R. (2005) "Ranking Ordinal Scales Using the Consensus Measure," Int'l Assn of Computer Information Systems Conference (IACIS 2005), Atlanta, GA, pp. 96-102.

[24] Van Horn, R. L., Schwarzkopf, A. and Price, R. (2006) *Information Systems Solutions: A Project Approach*, McGraw-Hill-Irwin, New York City.

[25] Velleman, P. F. and Wilkinson, L. (1993) "Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. The American Statistician, 47(1), pp. 65-72.

[26] Wierman, M. and Tastle, W. (2006) "Placing the dissonance measure in the context of generalized information theory," In North American Fuzzy Information Processing Society (NAFIPS 2006) Conference, Montreal, Canada.

[27] Zadeh, L. (1965) "Fuzzy Sets," Information and Control, 8, pp. 338-353.

[28] Tastle, W. (2009) "Measuring Faculty Instructional Performance." In The Proceedings of the Information Systems Educaiton Conference 2009, v 26 (Washington DC): §1532. ISSN: 1542-7382.

[29] Toubia, O. (2006) "Idea Generation, Creativity, and Incentives," Marketing Science, 25, pp. 411-415.

[30] Nasseri, H (2008) "Fuzzy Numbers: Positive and Nonnegative," International Mathematical Forun, 3(36), pp. 1777-1780.

William J. Tastle is an associate professor in the department of management at Ithaca College, New York, a Research Fellow at the Semeion Research Centre of Rome, Italy, a Research Professor at the University of Iceland, and a Fellow of the Association of Information Technology Professionals (educaton). He is also a senior member of both IEEE and ACM. His PhD is from the State University of New York at Binghamton in Advanced Technology with specialization in Systems Science, holds MS and MBA degrees, and his research interests include measure theory, offshoring of IT functions, and information systems curriculum development. He is currently the conference chair for the North American Fuzzy Information Processing Society Conference in 2010.

Amjad Abdullat serves as the head of the Department of Computer Information Systems and associate professor in the College of Business. He has a B.S. and B.A. from California State University at Sacramento, an M.B.A. from National University of San Diego and a Doctorate. from Pepperdine University. Dr. Abdullat's dissertation was "The Need for Paradigm Shift in Information Technology Planning." Dr. Abdullat is one of the founding principles and member of the board of directors of Edmin.com. Edmin is a comprehensive technology enterprise providing learning organizations with the next generation of web-based products, applications and services.

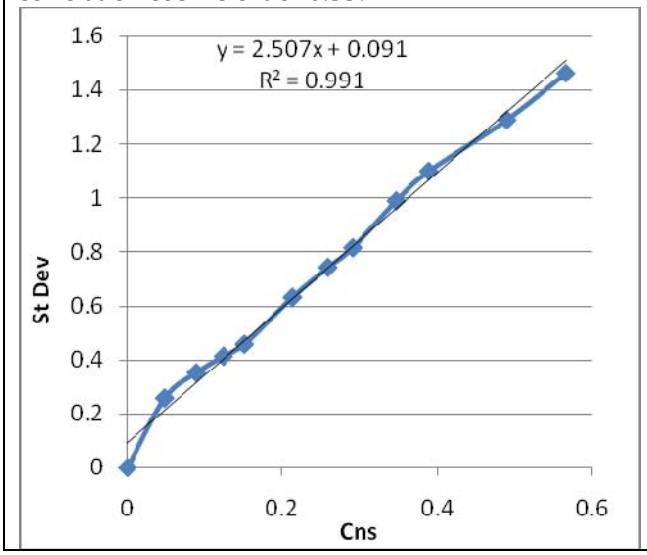
Mark J. Wierman BA mathematics, Purchase College, Purchase NY, 1978. MA mathematics, Binghamton University, Binghamton NY, 1980. PhD systems science, Binghamton University, Binghamton NY, 1994.

He has been an Instructor at SUC Oneonta, done research for Rome Labs, and programmed for IBM. Currently is an Assistant Professor of Computer Science at Creighton University in Omaha, NE. He co-authored two books, *Uncertainty-Based Information* with George Klir and *Applying Fuzzy Mathematics to Formal Models in Comparative Politics* with Terry D. Clark, Jennifer M. Larson, John N. Mordeson, and Joshua D. Potter.

Dr. Wierman major areas of research are Generalized Information Theory and the mathematics of Fuzzy Set Theory. Recently he has been studying measures of consensus as well as the application of fuzzy set theory to Political Science. He is a member of the Society for New Mathematics as well as a NAFIPS board member.

APPENDIX A

Figure 1 Comparison of Dissent (Dns) to Standard Deviation (StDev) as measures of dispersion. Note the correlation coefficient of 0.99.



Scale Type	Permissible Statistics	Admissible Scale Trans	Mathematical structure
Nominal (also denoted as categorical or discrete)	Mode, chi square	One to one (equality (=))	Standard set structure (unordered)
Ordinal	Median, percentile	Monotonic increasing (order (<))	Totally ordered set
Interval	Mean, standard deviation, correlation, regression, analysis of variance	Positive linear (affine)	Affine line
Ratio	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Positive similarities (multiplication)	Field

Table 9 Classification of the four different types of scales, Stevens (1946, 1951), taken from Wikipedia. Note that each scale type includes the permissible statistics of the previous types; hence, ordinal statistics include those in the nominal category (mode, chi square).

Appliance Controlling Using IPAC

Sandeep Kumar
Department of Computer Engineering, Aligarh Muslim University, Aligarh India 202001
sandeepkumar@zhcet.ac.in

Saiful Islam, Archana Gupta and Himanshu Bhardwaj
Department of Computer Engineering, Aligarh Muslim University, Aligarh India 202001
{ saifulislam, archanagupta,himanshubhardwaj}@zhcet.ac.in

***Abstract*— Today, because of the advancements in the computer and electronic sciences everything is going to be automated. In fact, some devices or infrastructures are capable to change the behavior according to situations; these devices are called Smart Devices or Smart Infrastructures. This system is designed to meet the requirement of appliance control in automated or smart infrastructures which includes home, offices, industries or may be sophisticated vehicles like aero planes. Appliance control basically refers the process or technique of controlling a device (including complete machines, mechanical devices, electronic devices, electrical devices etc.) using some comfortable, luxurious and reliable means based on some automation methods.**

Even a number of standards have been defined for wired and wireless controlling and automation of home appliances including Bluetooth, UPnP, X10 etc, this field is still in developing state. In this document we have proposed an appliance controlling system, named as Internet and PC Based Appliance Control (IPAC), using concepts of parallel port programming.

IPAC is designed to control a device from PC and from Internet, and can be applied in any smart infrastructure to automate the device and can work with almost every type of automation method either it is wired (e.g. LAN) or wireless (e.g. Bluetooth). This system can be applied in designing smart homes, secure homes, centralized device controlling system, Bluetooth control system, WAP control system.

***Index Terms*— Home Networking, Smart Homes, Secure Homes, UpnP Devices, X10 protocol, WAP Devices, IR Devices, Bluetooth Device**

I. INTRODUCTION

Why Home Networking is not so common?

There are several key problems associated with creation of home networked. Some of them are discussed below:

1. **Consumers are unaware of the benefits of the networked or smart home**

At this point in time, most home networks are used to connect PCs for tasks such as printing and shared Internet connectivity. Consumers still do not see the other potential benefits, such as on demand video, enhanced voice communications, and remote security control. Because of this lack of awareness, the demand for home networking products is still minimal

2. **Running additional wires through homes is costly and a hassle for consumers.**

In order to counteract this problem, the industry is developing wireless and other standards which will allow users to interconnect information devices without installation of new wires.

3. **Technology is too complex for most household users.**

Unlike other home electronics, the technology behind home networking is not intuitive and requires more technological expertise than the average household possesses.

4. **Lack of incentive for Internet providers to push networking technology.**

The home providers of broadband Internet (i.e., cable Internet providers and DSL providers) are currently surviving well enough on the strength of their connectivity service sales and do not need to push additional products. In addition, these communications carriers are too busy building network infrastructures and too swamped with customers demanding their high-speed access to spend time worrying about home networking.

5. **Potential privacy issues.**

Because the networked home would enable information to flow out of the home in ways that households are not accustomed to, privacy could be compromised. Additionally, the new technology behind information appliances and smart homes could introduce new security holes not before encountered.

6. **Interface issues.**

In smart home test beds, control interfaces have ranged from touch-screen devices to PDAs. Data on the effectiveness of the various interfaces seems scarce.

So, these were the some issues regarding the popularity of home networking. Now we shift our attention towards the home networking.

Imagine a completely networked home, in which every appliance can be remotely managed[14] from anywhere on the Internet with a simple Web browser[1][12][17][19]. The general goal of the automatic-home movement is to use networking technology to integrate the devices, appliances and

doi:10.4304/jetwi.2.3.232-238

services found in homes so that the entire domestic living space can be controlled centrally or remotely[16].
Following is a snap displaying a typical automated home. [2]

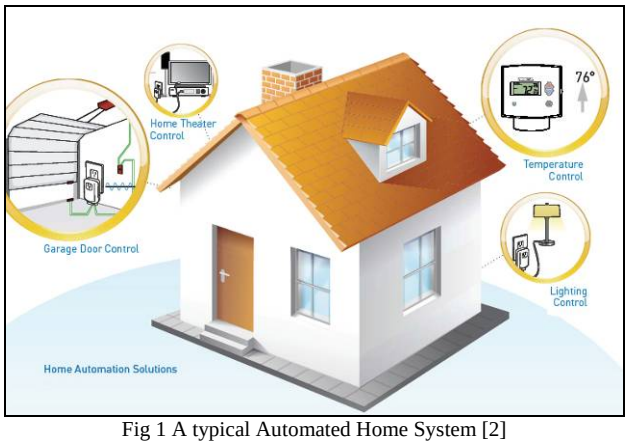


Fig 1 A typical Automated Home System [2]

Home wiring, the advance home developers are installing, typically adds several thousand dollars to the cost of a new home, and it is usually Ethernet or coaxial cable -- or some combination of both -- with other technologies in the mix. The network is being designed to make possible remote operation of appliances connected to the network.

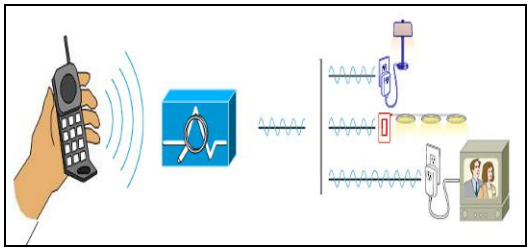


Fig 2 Already available extra wiring for device controlling in Smart Infrastructure

Other technology developers are generating buzz in this area as well. In June 2008, at the Bluetooth World Congress, vendors were touting the expansion of wireless networking technology into everything from air conditioners to cable television boxes."Bluetooth was originally developed as a wireless technology -- primarily for short-range exchange of data between laptops, PDAs and mobile phones," said Nick Hun, managing director at TDK Systems, whose Blu2i adapters are being used in such home applications. But, he noted, when early adapters were released to industrial engineers at the end of 2002 demand soon proved overwhelming.

Secure home
It is a highly cute smart home environment in which every device is automated with maintaining sufficient security. E.g. In the following figure home door is locked by software lock (by using a password) and can be opened only by software methods.

FIG 3 A SOFTWARE LOCK IN SECURE HOME ENVIRONMENT[2]

II. PRELIMINARIES

Summarized from [3][4][5][20][22] there are following technologies used to create a networking environment where home appliances work as a network nodes.

Direct Cable
In this devices are connected through serial, parallel or USB port. Generally desktop software is also supplied for making the device management a comfortable and easy task.

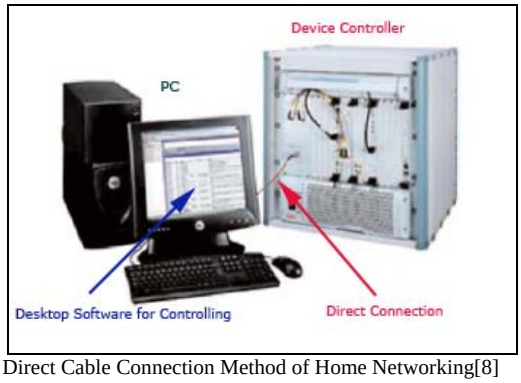


Fig 4 Direct Cable Connection Method of Home Networking[8]

Bluetooth
This is cross device wireless standard created for cell phones and PDAs, and can link up to eight devices.

Phone Line
Data shares the phone line frequency and requires phone jack every where a networked device is located. Also requires special cards and drivers.

Ethernet
Connections are made using hub system and network cards in each device. It requires driver installation and wiring. There are more expensive and chances of hardware conflicts are there.

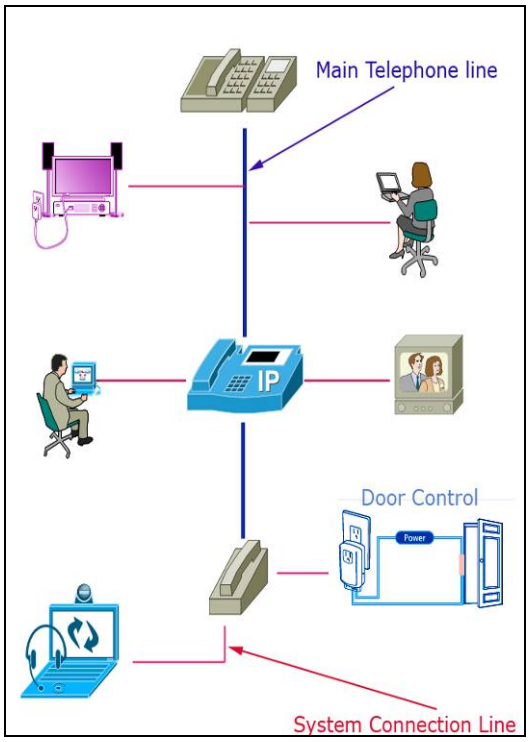


Fig 5 Use of telephone media for controlling home devices

Radio Free Network
Uses radio frequency waves to transmit data through walls and doors up to 800 feet, requires network card, can have some interference.

AC Network
Uses power lines and wiring already within home to connect parallel port to adapter in outlet. It is difficult to set up, slow networking, problems with interference from other devices and is also expensive [4].

Related Work
At 2Wire's R&D laboratory, researchers are currently developing [2009] wireless applications to control lighting and home security devices, Software and other IT companies are also not lagging behind in this advancement. Following figure shows a desktop PC Controlled automated home[2] from INSTEON.

Fig 6 Home controlling desktop software from INSTEON [2]

Related Technologies
UPnP(Universal Plug and Play Devices) UPnP technology is a distributed, open networking architecture that employs TCP/IP and other Internet technologies to enable seamless proximity networking, in addition to control and data transfer among networked devices in the home, office, and public spaces. Intel software for UPnP technology helps hardware designers and software developers build easy connectivity into common electronic devices [4] [6] [9].

X-10 Devices
X10 [7][21] is a communication language protocol that allows compatible products to talk to each other via existing 110 v electrical wiring in the home. Upto 256 different addresses are available and each device you can use usually requires a unique address.

Infra Red Device
IR data transmission is also employed in short-range communication among computer peripherals and personal assistance. Remote controls and IrDA devices use infrared light-emitting diodes (LEDS) to emit infrared radiation which is focused by a plastic lens into a narrow beam.

Bluetooth Devices [10][13]
It is a small form factor, low cost technology that provides low-power, short range (up to 10 m) links between mobile PCs, cell phones, printers or other devices arranged in ad hoc 'piconets' of up to 8 devices. For promoting interoperability between different Bluetooth devices Bluetooth Special Interest Group (SIG) has produced over 400 pages of profiles (published as volume 2 of the version 1.1 specifications). Bluetooth is a simpler technology than any other popular IEEE 802.1 1 standard for wireless local area networks. By 'simpler' is meant here fewer and/or less demanding RF semiconductor chips, fewer passive components and less complex digital base band chips.

III. Design of IPAC

Organization of IPAC

Figure[7] shows the complete organization structure of IPAC. In the figure only four devices are shown but using IPAC system we can control up to 128 devices (Why and How, This will be clear in the next section).

Functions of Different Units of IPAC

Web Interface (WI)
This is the interface available over the internet and appears on the browsers screen. Its function is to just provide an interface to user over the net through web browser so that user can access the appliance at the distant place (home or office).
User Interface (UI)

This is the WI counterpart for local users. This is the interface that is used by the users who own the server.

Fig. 7 Organization of IPAC

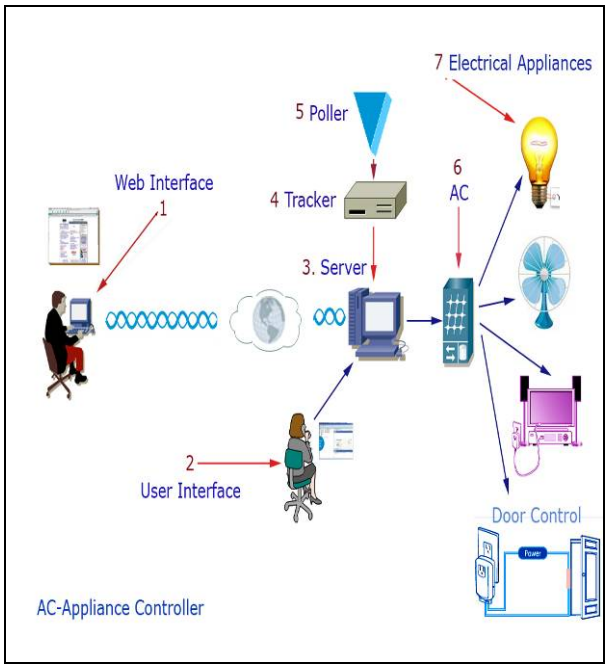


Fig. 7 Organization of IPAC

This is basically similar to INSTEON desktop[2] software and has aim of the device management a comfortable and easy task for user's point of view.

Server

This runs web-server so that the system can be accessed over the internet. Second important part of server is the database which stores the information about status of different devices.

Tracker

This system reads the status entries from database and generates proper control word to be PC port for generating proper signals. This is one of the most important parts of the system.

Control Word Format

For the proposed IPAC system an 8-bit control word will be used to control various home-appliances. Following figure shows the interpretation of different bits:

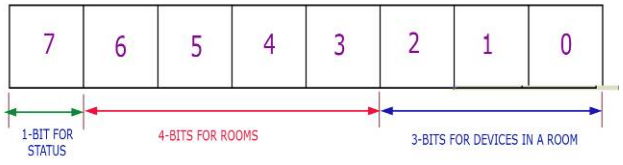


Fig.8 Control Word Format for proposed IPAC System

• 4-bits are used to address maximum of 16 rooms. (R-Field)

• 3-bits are used to address up to 8 devices within a single room. (D-Field)

• Single bit is used to set or reset (ON or OFF) the device selected by R-field and D-field bits. (SField)

In this structure we can address up to 128 devices because we are using 7-bits for addressing. All available 128 addresses (devices) are grouped into the 16 groups (rooms). Table1 shows these 16 groups and corresponding device address range.

Room Number	Device Address Range
1	0-7
2	8-15
3	16-23
4	24-31
5	32-39
6	40-47
7	48-55
8	56-63
9	64-71
10	72-79
11	80-87
12	88-95
13	96-103
14	104-111
15	112-119
16	120-127

Table:1 Grouping of 128 addresses into 16 different groups

Encoding

When converting the status record from database into control word we will first find the binary equivalent of room and device separately. From the control word structure it is clear that we have to left shift the room bits by three position to place them at the correct position. Then we can OR the 'room bits' and 'device bits' to determine the absolute address. .Similarly we have to 'OR' the absolute address with $(10000000)_2$ for making the status field '1' if the device is ON. Otherwise if the device is OFF there is no need to change the MSB because by default it is '0'. Moreover, one thing to be noted here is that the address of first room is '0000' but we say it 'Room 1" just to keep the room number in natural domain. Similar is the case with device address. So while decoding we will decrease the room number and device number by 1 prior to converting to binary equivalent. This is shown in the flow chart given below. While implementing the system in HLL it is important that in place of decreasing room-no by 1 then covering to binary and then shifting the room bits by 3-bits is just equivalent to multiplying room no by 8 after decreasing by 1. Similarly ORing with $(10000000)_2$ is equivalent to

adding with 128 in decimal number system (only in this particular case).

Example

If room number is 5, device number is 8 and we want to set this device in ON state.

Binary Method

$(5\text{-}1)_{10}=(4)_{10}=(00000100)_2$ (room bits)
$(8\text{-}1)_{10}=(7)_{10}=(00000111)_2$ (device bits)
Status =1
Shifting room bit left by 3-position we get $(00100000)_2$
After ORing with device bits we get $(00100111)_2$
Since the status is ON so we have to OR with $(10000000)_2$
After ORing we get $(10100111)_2$ which is the required control word.

Denary Method

$(5\text{-}1)_{10}=(4)_{10}$ (corrected room no)
$(8\text{-}1)_{10}=(7)_{10}$ (corrected device no)
Status =1
Multiplying corrected room no by 8 we get 4X8=32
Adding with corrected device number we get 32+7=39
Since the status is ON so we have to add 128
After adding 128 we get $(167)_{10}$ which is the required control word.

Since binary equivalent of $(167)_{10}$ is $(10100111)_2$ we are at the right track. Following figure the complete process flowchart

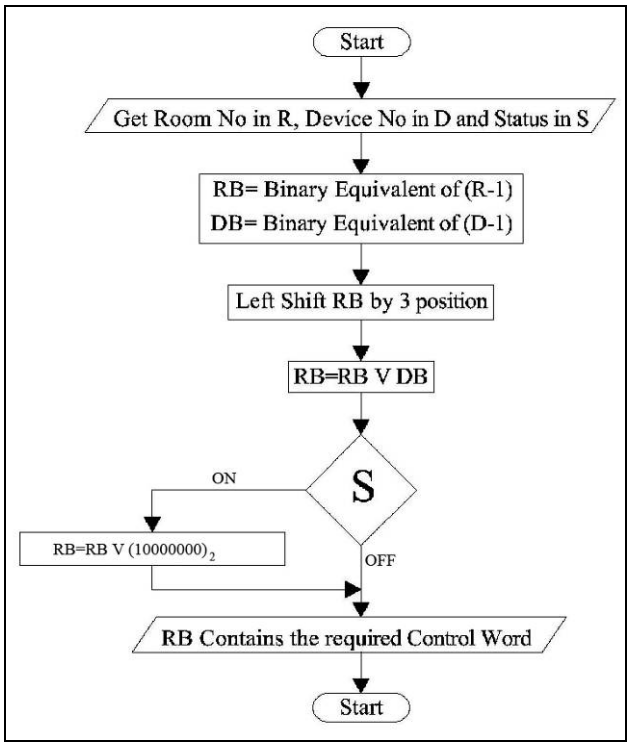


Fig 9: Flowchart for Encoding the room number, device number and their associated status into corresponding control word.

Poller

This part of IPAC remains in running state as long as your system (server) is ON (if you do not want to exit the IPAC service). This system, after a periodical time, executes the tracker so that if any changed had been made in the device status it should be propagate to corresponding device. Its function is to watch (poll) the database continuously so it is designated as Poller.

Appliance Controller (AC)

This part of IPAC, which is in the form of hardware, is responsible for interfacing the controlling PC (Server) with electrical devices. This part comprises of a 7X128 line decode that is used to address one out of 128 devices (7 input bits are taken from the 7 LSBs of parallel port). The MSB from the parallel port represents the data (status i.e. ON or OFF). With the addition of a relay of proper rating we can connect any device. The relay passes the AC signal as long as the device-status associated with this is ON. For maintaining the device continuously (even in power failure condition) we have to attach a memory element for storing the device status (We have used flip-flop) for that.

Electrical Appliances

These are home appliances which we are going to control. We will control the electrical device using this system but inclusion of transducer we can also control mechanical or electromechanical devices.

IV. SIMULATION

In Section-I, Section-II and Section-III we have laid out, designed our proposed system IPAC. In this section we will discuss a particular simulation of IPAC to analyze the results.

Simulation Requirements

Following table gives the detail requirements needed for running the simulation of IPAC :

Platform (Technology)	Purpose
Visual Basic 6.0	User Interface
HTML	Web Interface
ASP	For Controlling from Internet
Access 2000	Database Mappings
D-25 Type Parallel Port	Sending Control Signal to Appliance Controller
PC74HC154P	For Decoding Control Word
HD74LS04P	For Inverting Active Low Output of Decoder
12 LEDs	To display the Control Signals

Table.2 Simulation Requirements

Simulation Screen Shots

Following are some snaps taken from the simulation:

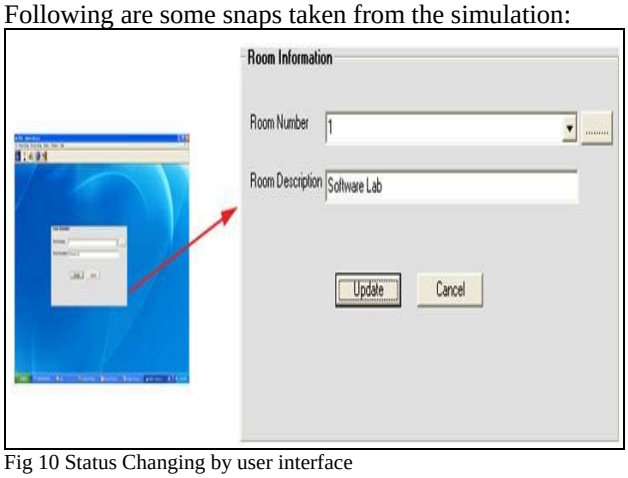


Fig 10 Status Changing by user interface

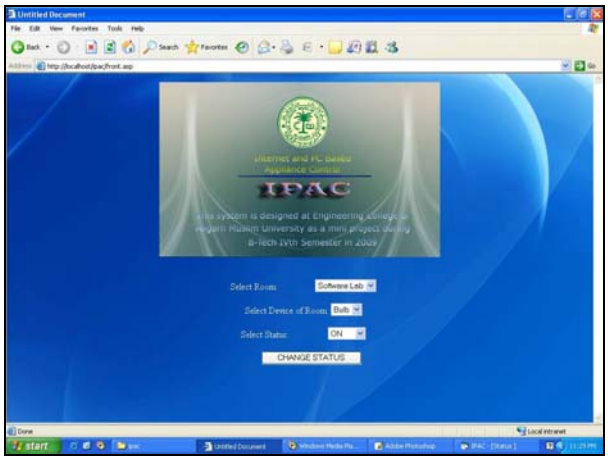


Fig 11 Status Changing by Web Interface

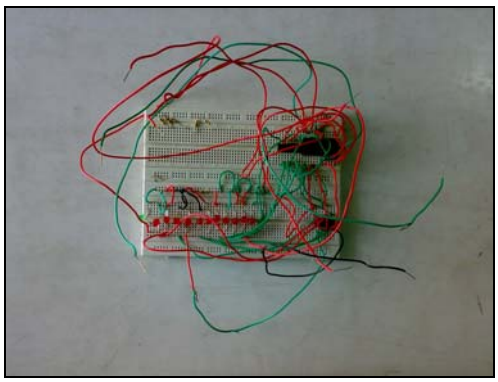


Fig 12 Appliance Controller for simulating controlling the LEDs

V. RESULT AND CONCLUSION

We have designed this system for controlling electrical devices but the design can be extended to control mechanical devices. We have simulated the system for controlling the LEDs and this is working properly. At last we will finish our system design discussion by concluding that using the *parallel programming concept of PC* is also a good tool for controlling home appliances. The major benefit is the cost. Since here the major cost factor is the PC which is generally present in every intermediate level family. Other major cost distribution factor is the cost of the software but this is a very long time asset. The hardware part (i.e. Appliance Controller) is just a decoder and a set of flip-flops so it is hardly of app. $10 USD. The remaining is the relays and the cost of the relay depends on the rating which depends on the device to be control. Except relays all the cost-distribution factors are one-time and fixed investment and does not depends on the number of devices we are going to control.

VI. ACKNOWLEDGEMENT

The authors would like to thank to Mr. M. Inamullah (Department of Computer Engineering), Mr. Izharuddin (Department of Computer Engineering) for their help in completing this work.

REFERENCES

[1] Gene J. Koprowski,"A Brilliant Future for the Smart Home", TechNewsWorld 08/04/03 8:11 AM PT
[2] "Solution that Gives you complex control of your House",SMARTHOME,VOL 118,2009
[3] Jennifer Recktenwald's "Home networks: Connecting more than computers", Tech Republic, November 10, 1999
[4] Piyush Varshney, "Remote Controlled Home Automation System", M Tech 3rd Sem, ZHCET, AMU, 2006
[5] G. Kannan and S. Vijayakumar, "Smart Home Tested for Disabled People",TIFAC CORE, Velammal Engineering College, Chennai, Mobile and Pervasive Computing-2008
[6] HyunRyong Lee, JongWon Kim, "UPnP Protocol Extension for Contents Sharing among Digital Home Networks", KISS, Vol.3 1, No.2, 2004
[7] Phil Kingery, "Digital X-10,"Advanced Control Technologies .Inc, 2002.
[8] M. Hashimoto, K. Teramoto, T. Saito, and T. Okamoto, "Home Network Architecture Considerin Digital home Appliance," IWNA98.
[9] Stefan Knauth, Rolf Kistler, Daniel K"aslin and Alexander Klapproth,"UPnP Compression Implementation for Building Automation Devices"
[10] Priyanka Varshney, Nikhil Kumar, Mohd. Rihan,"Implementation of Bluetooth Based Smart Home for Assisting the Physically Disabled"
[11] A. R. Al-Ali, M. AL-Rousan, "Java-Based Home Automation System," IEEE Transactions on Consumer Electronics, vol. 50, no. 2, pp. 498-504, May 2004.
[12] Peter M. Corcoran and Joe Desbonnet, "Browser-Style Interfaces To A Home Automation Network", EEE Transactions on Consumer Electronics, Vol. 43, No. 4, NOVEMBER 1997
[13] Kwang Ye01 Lee', Jae Weon Choi', "Remote-Controlled Home Automation System via Bluetooth Home Network ",SICE Annual Conference in Fukui, August 4-6,2003, Fukui University, Japan
[14] Intark Han, Hong-Shik Park, Youn-Kwae Jeong, and Kwang-Roh Park, "An Integrated Home Server for Communication, Broadcast Reception, and Home Automation" , IEEE Transactions on Consumer Electronics, Vol. 52, No. 1, FEBRUARY 2006
[15] John J. Greichen, "Value Based Home Automation For Today's Market", IEEE Transactions on Consumer Electronics, Vol. 38, No. 3, AUGUST 1992
[16] P.M. Cocoran, F. Papai and A. Zoldi, "User Interface Technologies for Home Appliances and Networks," IEEE Transactions on Consumer Electronics, Vol.44, No.3
[17] Andreas Rosendahl and J. Felix Hampe, Goetz Botterweck, "Mobile Home Automation – Merging Mobile Value Added Services and Home Automation Technologies", Sixth International Conference on the Management of Mobile Business (ICMB 2007)
[18] A. Alheraish, "Design and Implementation of Home Automation

System," IEEE Transactions on Consumer Electronics, vol. 50, no. 4, pp. 1087-1092, Nov. 2004.
[19] T. Uemukai, H. Hagino, T. Hara, M. Tsukamoto, K. Harumoto, and S. Nishio, "A WWW Browsing Method Using a Cellular Phone in a Remote Display Environment," Data Processing Conference (99-HI-86,99-MBL-l l), Vo1.99, No.97, p.51-56, 1999.
[20] Park Gwangro, "Trends of Home Network Technologies and Services," KRNET 2004, June, 2004.
[21] Zhang Yuejun, Wu Mingguang, "The Study and Development on Intelligent Lighting System Based on X-10 Protocol," China Illuminating Engineering Journal, vol. 15 No. 1, pp.22-26, Mar.2004.
[22] Bill Rose, WJR Consulting Inc. , "Home Networks : A Standards Perspective", IEEE Communication Magzine, December 2001

BIBLIOGRAPHY

Sandeep Kumar was born in India in July 1989. He has completed polytechnic in computer engineering from University Polytechnic, Aligarh Muslim University, Aligarh India in 2007. Currently he is pursuing a graduate degree in Computer Engineering. His area of interest includes automation, GPS services and location based services.

Mr. Kumar has two IEEE publications and one international publication. Mr. Kumar is also a member of CSI (Computer Society of India).

Saiful Islam was born in india in June 1978. He has completed his graduation and post graduation from Aligarh Muslim University, Aligarh India. Currently he is working as Senior Lecturer in the department of computer engineering, Aligarh Muslim University. He is also going to start his PhD from IIT Kanpur. His area of research includes Network Services and Security.

Archana Gupta was born in India in September 1987. She has completed polytechnic in computer engineering from University Polytechnic, Aligarh Muslim University, Aligarh India in 2007. Currently she is pursuing a graduate degree in Computer Engineering.
Ms. Gupta has one IEEE publication and one international publication. Her area of research includes web services, LBS services and device controlling.

Himanshu Bhardwaj was born in india in September 1989. After completing initial education in 2006 currently he is pursuing a graduate degree in Computer Engineering.
Mr. Bharddwaj is currently working in the area of security and web services.

GPS Talking For Blind People

Ameer H. Morad
Al Balqa Applied University, Faculty of Engineering Technology, Jordan
ameer_morad@yahoo.com

***Abstract*—In this paper, we design device to help the blind people to navigate the environment without asking anyone. The device based on GPS (Global Positioning System), the raw data for location coordinate where the blind people stands is detect by GPS receiver, processing these data by PIC microcontroller to calculate real coordinate related with current position, then translate it to specific voice message which are presorted in voice recorder, the blind person hears voice message through the headset. Our design aims are to produce device that is more cheap by using little number of components and easy to use so that the blind person not need to do any thing just hearing the voice message. The device be practically tested by some blind people who are members of Abdallah Bin Maktoom blinds school in Jordan, they gives good opinion about device.**

***Index Terms*—GPS receiver applications, navigate system, microcontroller applications, digital system design.**

I. INTRODUCTION

Imagine being blind and trying to find your way around a city you've never visited before -- that can be challenging for a sighted person. Researches design navigator system based on new development technology, tools, small and wearable devices to help people who are blind and visually impaired. The WHO (World Health Organization)[1] estimates in 2002 that there are 161 million; about 2.6 % of the world population; visually impaired people in the world, 124 million of them had low vision and 37 million are blind. These visual impairment people are distributed fair around the world as shown the Fig. 1. For that we try in this research to build a useful device to help the blind people to navigate through the city and receive in each step a voice massage tell them where he is now.

Blind people can obtain information from the unwilling contact with objects, persons or animals, by exploring the environment and using their hands to understand the shape of an object, more over, blind people can perceive other features of the objects as temperature, texture, weight… and though the tact has certain limits in confront of sight, it has a very important function in reveal to blind persons the world around them. Another sense is very important in the life of blind people: the heard, that has great qualities of global, longer range discernment, and on which they rely upon for identification of objects and spaces.

Many researchers and companies are developing technologies and device that emits sound to help blind peoples to navigate the world.

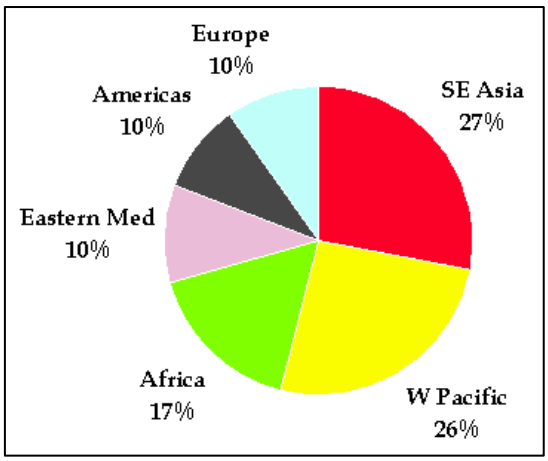


Figure 1: Global estimation of visual impairment by WHO region

The SWAN system [2] consists of a small laptop worn in a backpack, a proprietary tracking chip, Global Positioning System (GPS) sensors, a digital compass, a head tracker, four cameras, a light sensor, and special headphones. The sensors and tracking chip send data to the SWAN applications on the laptop, which computes the blind's location and in what direction he is looking. It then maps the travel route, and sends 3D audio cues to the bone phones to guide the traveler along a path to the destination. Researches at Wright State University [3] design a portable system; called Tyflos -- Greek for blind -- consists of a tiny camera mounted on a pair of glasses, a laptop carried in a backpack, a headset and a microphone. Tyflos converts the images to sound. Computer algorithms process the images and extract information from them to give the blind information about what they are looking. Scientists at the European Commission's Joint Research Centre [4] have developed a prototype system, SESAMONET (Secure and Safe Mobility Network), which uses RFID micro-chips embedded in the ground to guide a visually impaired person through a predefined area. The microchips can be recycled from the electronic tracking of cattle. Each micro-chip sends position signals via a dedicated walking stick to a smart phone containing information about the location and a recorded voice – via a Bluetooth headset - guides the visually impaired person along the route.

doi:10.4304/jetwi.2.3.239-243

The main objective of our project is to design a small and simple navigate device to help the blind people to get environmental information as voice message depend on GPS technology. GPS system is used wide in both civilian and military applications. GPS system offers various aids for the autonomous mobility of visually impaired persons:

1). Provides information about turns and obstacles on the path.
2). Checks the right direction (useful if user falls or is disoriented).
3). Provides general and specific environmental information.

II. DESIGN OF THE BLIND DEVICE

The block diagram for our designing of blind navigation device is shown in Fig. 2, which is consist of GPS receiver, PIC microcontroller, Voice recorder, LCD, Headset and other supplementary components.

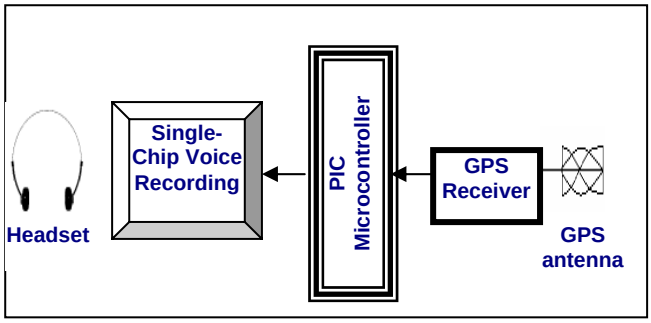


Figure 2: Block Diagram for Blind Device.

In this part we describe the features of all components consist in the designing system and explain how these components are connecting together, as shown in Fig. 3, to accomplishment the blind device concepts.

A. GPS Receiver

GPS is a satellite-based global navigation system [5, 6] that enables users to accurately determine 3-dimensional positions (x, y, z) worldwide.
GPS consists of 3 segments. They are:

1). The space segment consists of at least 24 satellites. The satellites circle the Earth once every 12 hours. The satellites transmit radio signals continuously to broadcast its changing position and time.
2). The control segment consists of ground stations that monitor and control the satellites.
3). The user segment consists of the user and the GPS receiver. The GPS receiver measures the signals from the satellites and identifies the user's position.

The satellites are spaced so that from any point on Earth, at least four satellites will be above the horizon. Each satellite continually transmits radio signals and broadcasts its position and time. GPS uses satellites in space as reference points to locate the positions on the Earth. The GPS receiver measures our distance from the satellites by measuring the travel time of the radio signals. The distance from the satellite to the GPS receiver is equal to the travel time from the satellite to GPS receiver multiplied by the speed of light. That is, Distance = Travel time x Speed of light. 0On the ground, any GPS receiver that contains a computer can locate its own position on the Earth by measuring accurately the distance from three satellites. The result is provided in the form of a geographic position – longitude and latitude.

We use Rockwell TU00-D200-401 as GPS receiver is single-board, high performance, low power, and 12-channel receiver. It is Dimension 50.98 x 71.12 x 10.2 mm, a +5V primary DC input power supply, a GPS antenna and a TTL to RS232 level to be converter able to communicate with serial port.

B. Voice Recorder

For voice information we use single chip voice recorder (APR9600), it is high quality voice recording with non volatile storage and a play back ability for 40 to 60 seconds with minimum external components so no external IC required, and it is easy to use so there is no need to program or develop it. The APR9600 device is able to reproduce voice signals in their natural form. It eliminates the need for encoding and compression, which often introduce distortion.

The device supports five message management modes Therefore, the designer must select the appropriate operating mode before beginning the design, and these modes are:

1). Random Access 2 fixed duration messages
2). Random Access 4 fixed duration messages
3). Random Access 8 fixed duration messages
4). Tape mode, Auto rewind operation
5). Tape mode, Normal operation

In our project we use Random Access 8 fixed duration messages mode, because Random Access mode provides easy indexing to massage segments and the recording or play back can be made randomly in any of the selected messages.

C. PIC Microcontroller 16F877

We used low cost 40 pin PIC Microcontroller 16F877, as controller part to the system, as shown in Fig. 3, for correct work, it needs:

1). Power supply: in range between 2-6 volts, we used a +5V as a power supply with the VDD, but the VSS we connected it with the ground.
2). Voltage stabilizer: to stable source voltage, so we used a voltage stabilizer which gives stable +5V on its output.
3). Reset: to return the device into a particular state. We connected the pin1 (MCLR) with 10k resistor.

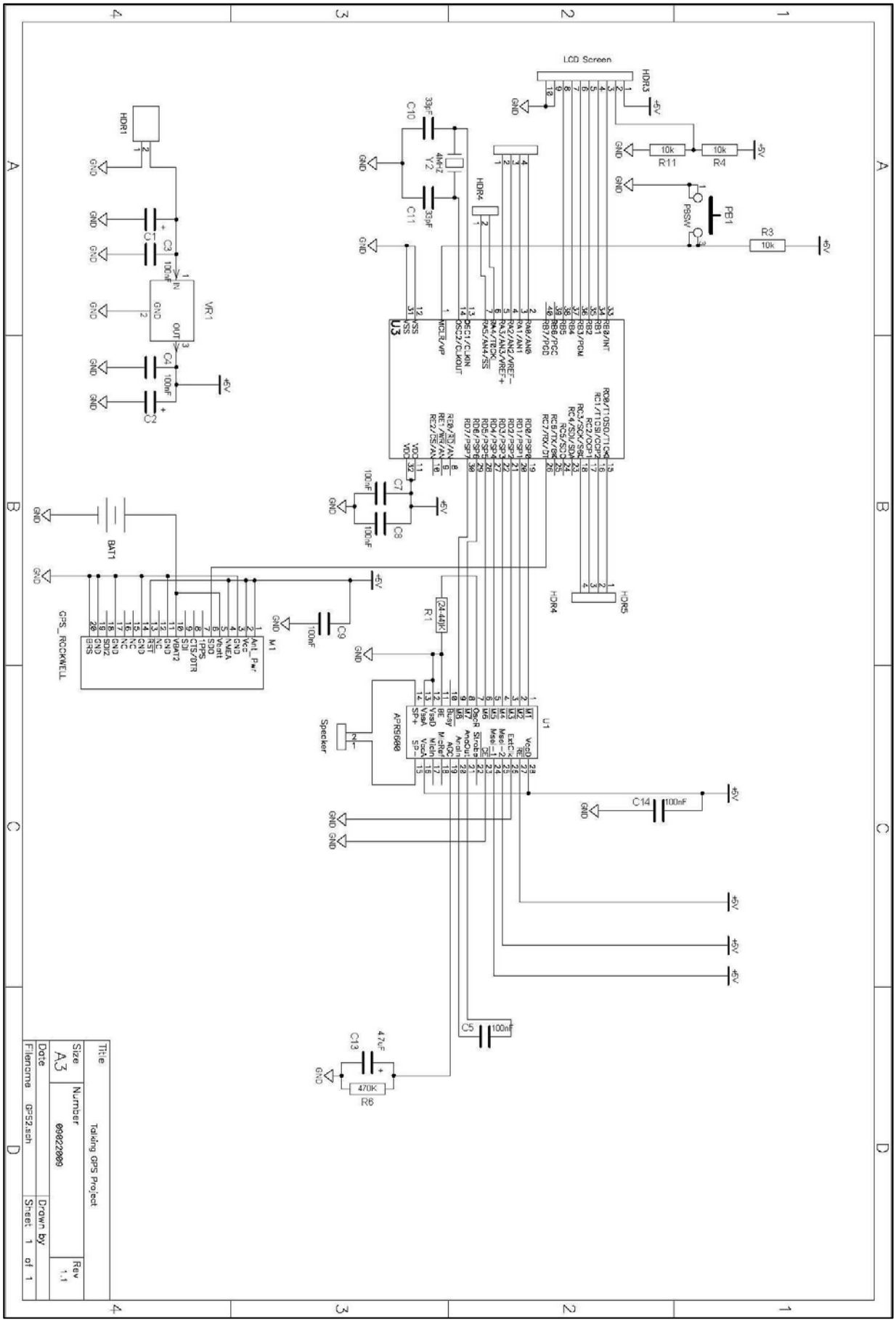


Figure 3: Connection of all device components

4). Oscillators: use 4MHz crystal oscillator to generate the device clock, connect pin 13 and 14 (OSC1 and OSC2) with two capacitors 33pf.

The microcontroller converts the current position coordinates; which be compute by GPS receiver; to address the specific voice message stored before in voice recorder chip related with the current position.

The functionality of microcontroller, as shown in Fig.4, is to receive GPS Fix Data (GPGGA) for a GPS receiver, which content the time, position and other fix related data, for detail of GPGGA format see reference [7], and from GPGGA message, the microcontroller extract the Latitude and Longitudinal components of position. Then the microcontroller sends this position information to both units, LCD to display it and to voice recorder chip to address and play the voice message related with current position.

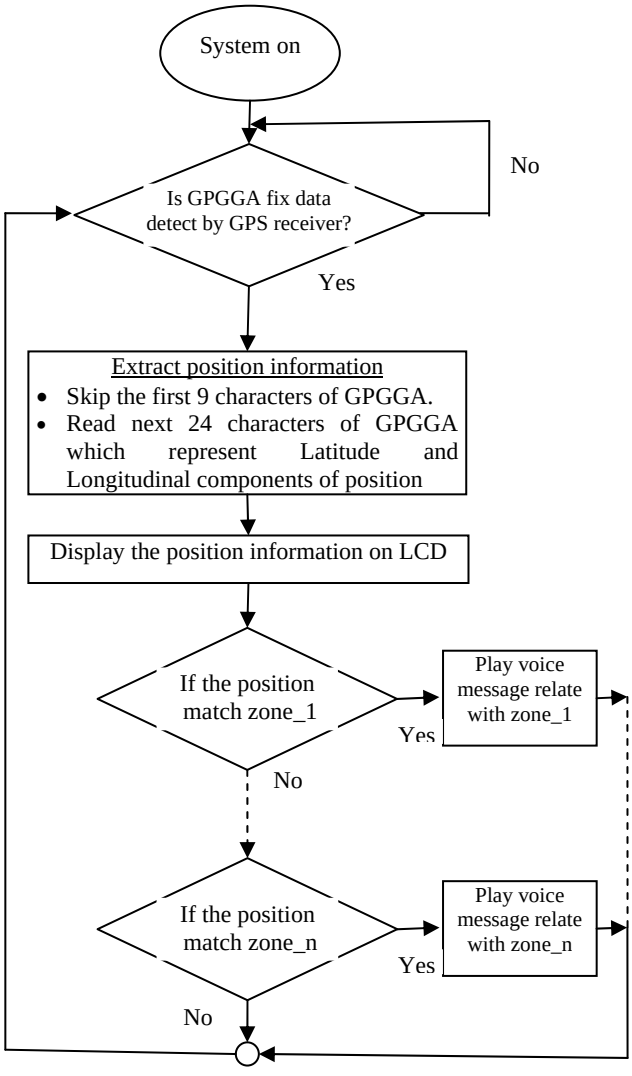


Figure 4: microcontroller function algorithm

D. Microphone and Headset

During learning and updating step of the system, the designer use microphone to store voice message into voice recorder chip. Also during running step of the system, when the blind people navigate the places, the headset is necessary for them to hear the voice message related with current position.

E. LCD Device

The LCD device is used by the designer in learning step and updating of the system, through LCD, the designer can read the current position coordinates that being calculate by microcontroller.

III. System Test and Implementation

The device is tested in faculty of Computer Engineering Technology - Al Balqa Applied University camp. During learning phase we navigate the camp and determine the latitude and longitudinal components of position for boundary are of 8 building prototype, each building known as zone and define it by two diagonal points, see table (1).

In each zone the position coordinates for it points are displayed on LCD screen, when the GPS receiver capture GPGGA messages at these points and then the microcontroller process it to extract latitude and longitudinal components. Also the microcontroller convert these latitude and longitudinal components for zone's points to specify address on voice recorder chip in which the designer store voice message tell what is the building in that zone.

During running phase we contact with Abdallah Bin Maktoom blinds school to test the device practically, the device be use by blind people and we get good opinion. "the device is very simple and it be more useful if it cover the whole Jordan, so it is easy form to navigate across Jordan without asking any body where I am", said Mr. Mohammad Dyab Khaid who lost sight, and he work as science teacher in the school. "I change my home and really I faced a big problem because I haven't any idea about places near my new home, but may be with my new friend, GPS Talking device, I can easily tour any where", said Mr. Mansoor Martooq, who is Arabic teacher in the same school.

IV. Conclusions

In this project we design a small device, very easy to use called, GPS Talking Blind People, help blind people to navigate around camps, cities and get voice messages tell him where he is now?

The device loaded with prototype information about some building in faculty of engineering technology camps and test successively by blind people.

Table 1: Position Coordinates for 8 prototype zones.

Locations	First point	End point
Building number 17	3159.604 N-3601.114 E	3159.647 N-3601.129 E
Building number 5	3159.601 N-3601.093 E	3159.583 N-3601.135 E
Building number 9	3159.787 N-3600.968 E	3159.539 N-3601.116 E
Book shop	3159.557 N-3601.055 E	3159.567 N-3601.134 E
IEEE student branch	3159.567 N-3601.134 E	3159.511 N-3601.119 E
Engineering Workshops	3159.616 N-3601.087 E	3159.428 N-3600.903 E
Library & Registration	3159.560 N-3600.996 E	3159.571 N-3601.009 E
Student society & Cafeteria	3159.428 N-3600.903 E	3159.557 N-3601.055 E

During designing of the device we face some problems such as:

1). Same times the position coordinates is not stable at the same point because the GPS system is not very accurate, it has a range of error about 2-3 meters, delay of GPS signals when pass atmosphere, also the accuracy effect on number of satellites be detect by GPS receiver.

2). GPS system can't use it indoor building or closed area, so we test the device outdoor only, for indoor we must use some additional equipments.

3). The storage capacity of APR9600 voice recorder is relative small, play back ability for only 40 to 60 seconds. So for extending and updating the device it must it with more storage capacity chips.

REFERENCES

[1]. http://www.who.int/topics/blindness/en/index.html
[2]. http://www.eurekalert.org/pub_releases/2006-08/giot-sst080706.php
[3]. http://www.wired.com/gadgets/miscellaneous/news/2004/03/62810?currentPage=all
[4]. Proceeding of 3d European conference on eAccessibility Forum, Mobile communications helping people with disabilities at work, 30 March 2009.
[5]. Elliott D Kaplan and Christopher Hegarty, “Understanding GPS: Principles and Applications”, Artech House,2nd ed., 2006.
[6]. James Bao-Yen Tsui, “Fundamentals of Global Positioning System Receivers: A Software Approach”, 2nd ed., Wiley, 2005.
[7]. http://aprs.gids.nl/nmea/

Towards Intelligent Ontology Alignment Systems for Question Answering: Challenges and Roadblocks

Maria Vargas-Vera, Miklos Nagy
Open University, Walton Hall, Milton Keynes, MK7 6AA, England, UK
Email: {M.Vargas-Vera, M.Nagy}@open.ac.uk

Abstract— **This paper introduces the main challenges and future research directions for the Ontology Alignment problem. To date a good number of ontology alignment solutions have been proposed. These solutions utilise a wide variety of techniques from machine learning to uncertain reasoning. However, none of the approaches have proved to be an integrated solution, which can be used by different communities. Since 2004, the Ontology Alignment Initiative (OAEI) has established an annual evaluation for systems that could be tested using the same datasets. This, of course, has helped to improve the work on ontology alignment, as the ontology community now has a set of common datasets to make comparisons on the performance of different algorithms for ontology alignment. In this paper we discuss the main challenges and roadblocks that need to be addressed in order to built successful mapping frameworks. Finally this paper presents DSSim and our results on the ontology evaluation 2008.**

Index Terms— **ontology mapping, uncertain reasoning**

I. Challenges on ontology alignment

The Ontology Alignment Community has identified several challenges [1], [2], which are considered as major roadblocks for successful future implementations of ontology mapping systems. In our context (ontology mapping for question answering), we have identified five critical and interrelated challenges that can be considered as roadblocks for future successful mapping frameworks. The first challenge relates to the representation, the second to the quality of the data and the third one to the efficient ontology alignment for large ontologies. The fourth and the fifth challenge goes beyond the information related concepts and tries to address the overall difficulties namely the problem of generic and intelligent systems. We will discuss each our suggested challenges in turn.

A. Representation and interpretation problems

The vision of the Semantic Web is to achieve machine-processable interoperability through the annotation of the content. This implies that computer programs can achieve a certain degree of understanding of such data and use it to reason about user specific tasks like question answering or data integration. Data on the semantic web is represented by ontologies, which typically consist of a number of classes, relations, instances and axioms. These elements are expressed using a logical language. The W3C has proposed RDF(S) [3] and OWL [4] as Web ontology language however OWL has three increasingly-expressive sublanguages (OWL Lite, OWL DL, OWL Full) with different expressiveness and language constructs. In addition to the existing Web ontology languages W3C has proposed other languages like SKOS [5], which is a standard to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. SKOS are based on the Resource Description Framework (RDF) and it allows information to be passed between computer applications in an interoperable way. Ontology designers can choose between these language variants depending on the intended purpose of the ontologies. The problem of interpreting semantic web data however stems not only from the different language representations [6] but the fact that ontologies especially OWL Full has been designed as a general framework to represent domain knowledge, which in turn can differ from designer to designer. Consider the following excerpts Fig. 1, 2 from different FAO (Food and Agricultural Organization of the United Nations) ontologies.

Assume we need to assess similarity between classes and individuals between the two ontologies. In fragment one a class *c_8375* is modelled as named OWL individuals. In the class description only the ID is indicated therefore to determine the properties of the class one needs to extract the necessary information from the actual named individual. In Fig. 2 the classes are represented as RDF individuals where the individual properties are defined as OWL data properties. One can note the difference how the class labels are represented on Fig. 1 through *rdfs:label* and Fig. 2 through *hasNameScientific* and *hasNameLongEN* tags. From the logical representation point of view both ontologies are valid separately and no logical reasoner would find inconsistency in them individually. However the problem occurs once we need to compare them in order to determine the similarities between classes and individuals. It is easy to see that once we need to compare the two ontologies a considerable amount of uncertainty arises over the classes and its properties and in a way they can be compared. This uncertainty can be contributed to the fact that due to the different representation certain elements will be missing for the comparison e.g. we have label in fragment Fig.

doi:10.4304/jetwi.2.3.244-257

```
...
<owl:Class rdf:ID="c_8375">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="c_7033"/>
  </rdfs:subClassOf>
</owl:Class>
...
< c_8375 rdf:ID="i_8375">
  <aos:hasScopeNote xml:lang="EN">Isscaap group
b-52</aos:hasScopeNote>
  <aos:hasScopeNote xml:lang="FR">Groupe b-52 de
la csitapa</aos:hasScopeNote>
...
  <rdfs:label xml:lang="en">Demospongiae</rdfs:label>
</c_8375>
...
```

Figure 1. Ontology fragment from the AGROVOC ontology

```
...
<owl:Class rdf:about="#species">
  <rdfs:subClassOf rdf:resource="#biological_entity"/>
  <owl:disjointWith rdf:resource="#family"/>
  <owl:disjointWith rdf:resource="#order"/>
  <owl:disjointWith rdf:resource="#group"/>
</owl:Class>
...
<rdf:Description rdf:about="http://www.fao.org/aims
/aos/fi/species_v1.0.owl#31005_17431">
  <j.0:hasNameLongEN>Barrel
sponge</j.0:hasNameLongEN>  <j.0:hasMeta>31005
  </j.0:hasMeta>  <j.0:hasNameScientific>
  DEMOSPONGIAE</j.0:hasNameScientific>
</rdf:Description> ...
```

Figure 2. Ontology fragment from the ASFA ontology

1 but is missing from fragment Fig. 2 but there is *hasNameLongEN* tag in fragment Fig. 2 but missing in fragment Fig. 1.

As a result of these representation differences ontology mapping systems will always need to consider the uncertain aspects of how the semantic web data can be interpreted.

B. Quality of the Semantic Web Data

Data quality problems [7] [8] in the context of database integration [9] have emerged long before the Semantic Web concept has been proposed. The major reason for this is the increase in interconnectivity among data producers and data consumers, mainly spurred through the development of the Internet and various Web-based technologies. For every organisation or individual the context of the data, which is published can be slightly different depending on how they want to use their data. Therefore from the exchange point of view incompleteness of a particular data is quite common. The problem is that fragmented data environments like the Semantic Web inevitably lead to data and information quality problems causing the applications that process this data deal with

```
...
<owl:Class rdf:about="http://matching.com/source
  /3887.owl#Windows_Vista">
  <rdfs:label xml:lang="en"> Windows Vista Home
Edition </rdfs:label>
  <j:hasSerialNumber>
    <rdfs:label >00043-683-036-658</rdfs:label>
  </j:hasSeriamNumber>
  <rdfs:subClassOf>
    <owl:Class rdf:about="http://matching.com
    /source/3887.owl#Operating_Systems">
    </owl:Class>
  </rdfs:subClassOf>
</owl:Class>
...
```

Figure 3. Ontology fragment from the Web directories ontology

ill-defined, inaccurate or inconsistent information on the domain. The incomplete data can mean different things to data consumer and data producer in a given application scenario. In traditional integration scenarios resolving these data quality issues represents a vast amount of time and resources for human experts before any integration can take place. Data quality has two aspects

- Data syntax covers the way data is formatted and gets represented
- Data semantics addresses the meaning of data

Data syntax is not the main reason of concern as it can be resolved independently from the context because it can be defined what changes must occur to make the data consistent and standardized for the application e.g. defining a separation rule of compound terms like "MSc-Thesis", "MSc_Thesis". The main problem what Semantic Web applications need to solve is how to resolve semantic data quality problems i.e. what is useful and meaningful because it would require more direct input from the users or creators of the ontologies. Clearly considering any kind of designer support in the Semantic Web environment is unrealistic therefore applications itself need to have built in mechanisms to decide and reason about whether the data is accurate, usable and useful in essence, whether it will deliver good information and function well for the required purpose. Consider the following example Fig. 3 from the directory ontologies.

As figure Fig. 3 shows we can interpret Windows Vista as the subclass of the operating systems however the designer has indicated that it has a specific serial number therefore it can be considered as an individual as well. At any case the semantic data quality is considered as low as the information is dubious therefore the Semantic Web application has to create its own hypothesises over the meaning of this data.

C. Efficient mapping with large scale ontologies

Ontologies can get quite complex and very large, causing difficulties in using them for any application [10] [11]. This is especially true for ontology mapping where

overcoming scalability issues becomes one of the decisive factors for determining the usefulness of a system. Nowadays with the rapid development of ontology applications, domain ontologies can become very large and complex. This can partly be contributed to the fact that a number of general knowledge bases or lexical databases have been and will be transformed into ontologies in order to support more applications on the Semantic Web. Consider for example WordNet. Since the project started in 1985 WordNet [1] has been used for a number of different purposes in information systems. It is popular general background knowledge for ontology mapping systems because it contains around 150.000 synsets and their semantic relations. Other efforts to represent common sense knowledge as ontology is the Cyc project [2], which consists of more than 300.000 concepts and nearly 3.000.000 assertions or the Suggested Upper Merged Ontology(SUMO)[3] with its 20.000 terms and 70.000 axioms when all domain ontologies are combined. However the far largest ontology so far (according to our knowledge) in terms of concept number is the DBPedia [4] , which contains over 2.18 million resources or "things", each tied to an article in the English language Wikipedia. Discovering correspondences between these large scale ontologies is an ongoing effort however only partial mappings have been established i.e. SUMO-Wordnet due to the vast amount of human and computational effort involved in these tasks. The Ontology Alignment Initiative 2008 [12] has also included a mapping track for very large cross lingual ontologies, which includes establishing mappings between Wordnet, DBPedia an GTAA (Dutch acronym for Common Thesaurus for Audiovisual Archives) [13], which is a domain specific thesaurus with approximately 160.000 terms. A good number of researchers might argue that the Semantic Web is not just about large ontologies created by the large organisations but more about individuals or domain experts who can create their own relatively small ontologies and publish it on the Web. Indeed might be true however from the scalability point of view it does not change anything if thousands of small ontologies or a small number of huge ontologies need to be processed. Consider that in 2007 Swoogle [14] has already indexed more than 10.000 ontologies, which were available on the Web. The large number of concepts and properties that is implied by the scale or number of these ontologies poses several scalability problems from the reasoning point of view. Any Semantic Web application not only from ontology mapping domain has to be designed to cope with these difficulties otherwise it is deemed to be a failure from the usability point of view.

D. Task specific vs. generic systems

Existing mapping systems can clearly be classified into two categories. First group includes domain specific systems, which are build around well defined domains e.g. medical, scientific etc. These systems use specific rules, heuristics or background knowledge. As a consequence domain specific systems perform well on their own domain but their performance deteriorate across different domains. As a result the practical applicability of these systems on the Semantic Web can easily be questioned. The second group includes systems that aim to perform equally well across different domains. These systems utilise generic methods e.g. uncertain reasoning, machine learning, similarity combination etc. These systems has the potential to support a wide variety of applications on the Semantic Web in the future.

Based on this classification it is clear that building generic systems that perform equally well on different domains and provide acceptable results is a considerable challenge for the future research.

E. Incorporating intelligence

To date the quality of the ontology mapping was considered to be an important factor for systems that need to produce mappings between different ontologies. However competitions organised on ontology mapping has demonstrated that even if systems use a wide variety techniques, it is difficult to push the mapping quality beyond certain limits. It has also been recognised [15] that in order to gain better user acceptance, systems need to introduce cognitive support for the users i.e. reduce the difficulty of understanding the presented mappings.

There are different aspects of this cognitive support i.e. how to present the end results, how to explain the reasoning behind the mapping, etc. Ongoing research focuses on how the end results can be represented in a way that end users can understand better the complex relations of large-scale ontologies. Consider for example a mapping representation between two ontologies with over 10.000 concepts each. The result file can contain thousands of mappings. To visualise this mapping existing interfaces will most likely present an unrecognizable web of connections between these properties. Even though this complex representation can be presented in a way that users could better understand the problem still arises once the users need to understand why actually these mappings have been selected. This aspect so far has totally been hidden from the end users and has formed an internal and unexpoitable part of mapping systems itself.

Nevertheless in order to further improve the quality of the mapping systems these intermediary details need to be exposed to the users who can actually judge if the certain reasoning process is flawed or not. This important feedback or the ability to introspect can then be exploited by the system designers or ultimately the system itself through improving the reasoning processes, which is carried out behind the scenes in order to produce the end results. This ability to introspect the internal reasoning steps is a fundamental component of how human beings reason, learn and adapt. However, many existing ontology mapping systems that use different

[1] http://wordnet.princeton.edu/
[2] http://www.cyc.com/
[3] http://www.ontologyportal.org/
[4] http://dbpedia.org/About

forms of reasoning exclude the possibility of introspection because their design does not allow a representation of their own reasoning procedures as data. Using a model of reasoning based on observable effect it is possible to test the ability of any given data structure to represent reasoning. Through such a model we present a minimal data structure [16] necessary to record a computable reasoning process and define the operations that can be performed on this representation to facilitate computer reasoning. This model facilitates the introduction and development of basic operations, which perform reasoning tasks using data recorded in this format. It is necessary that we define a formal description of the structures and operations to facilitate reasoning on the application of stored reasoning procedures. By the help of such framework provable assertions about the nature and the limits of numerical reasoning can be made.

II. Approach to Ontology alignment considering uncertainty

For ontology mapping in the context of Question Answering over heterogeneous sources we propose a multi agent system called DSSim [17] because as a particular domain becomes larger and more complex, open and distributed, a set of cooperating agents are necessary in order to address the ontology mapping task effectively. In real scenarios, ontology mapping can be carried out on domains with large number of classes and properties. Without the multi agent architecture the response time of the system can increase exponentially when the number of concepts to map increases. The main objective of DSSim architecture is to be able to use it in different domains for creating ontology mappings. These domains include Question Answering, Web services or any application that need to map database metadata e.g. Extract, Transform and Load (ETL) tools for data warehouses. Therefore DSSim is not designed to have its own user interface but to integrate with other systems through well defined interfaces. In our implementation we have used the AQUA Question Answering system, which is the user interface that creates First Order Logic(FOL) statements based on natural language queries posed by the user. As a consequence the inputs and outputs for the DSSim component are valid FOL formulas.

An overview of our system is depicted on Fig. 4 The two real word ontologies[5][6] describe BibTeX publications from the University of Maryland, Baltimore County (UMBC) and from the Massachusetts Institute of Technology (MIT) . The AQUA [18] system and the answer composition component are described just to provide the context of our work (our overall framework) but these are not our major target in this paper. The user poses a natural language query to the AQUA system, which converts it into FOL (First Order Logic) terms. The main components and its functions of the system are as follows:

[5]http://ebiquity.umbc.edu/ontology/publication.owl
[6]http://visus.mit.edu/bibtex/0.01/bibtex.owl

1) Broker agent receives FOL term, decomposes it(in case more than one concepts are in the query) and distributes the sub queries to the mapping agents.
2) Mapping agents retrieve sub query class and property hypernyms from WordNet.
3) Mapping agents retrieve ontology fragments from the external ontologies, which are candidate mappings to the received sub-queries. Mapping agents use WordNet as background knowledge in order to enhance their beliefs on the possible meaning of the concepts or properties in the particular context.
4) Mapping agents build up coherent beliefs by combining all possible beliefs over the similarities of the sub queries and ontology fragments. Mapping agents utilize both syntactic and semantic similarity algorithms build their beliefs over the correctness of the mapping.
5) Broker agent passes the possible mappings into the answer composition component for particular sub-query ontology fragment mapping in which the belief function has the highest value.
6) Answer composition component retrieves the concrete instances from the external ontologies or data sources, which is included into the answer.
7) Answer composition component creates an answer to the user's question.

The main novelty in our solution is that we propose solving the ontology mapping problem based on the principles of collective intelligence, where each mapping agent has its own individual belief over the solution. However before the final mapping is proposed the broker agent creates the result based on a consensus between the different mapping agents. This process reflects well how humans reach consensus over a difficult issue.

A. Example scenario

Based on the architecture depicted on Fig. 4 we present the following simplified example, which will be used in the following sections of the paper in order to demonstrate our algorithm. We consider the following user query and its FOL representation as an input to our mapping component framework: List all papers with keywords uncertain ontology mapping?

$$(\exists x)\, paper\,(x)\, and$$
$$hasKeywords\,(x, [uncertain, ontology\ mapping])$$

- Step 1: Broker agent distributes (no decomposition is necessary in this case) the FOL query to the mapping agents.
- Step 2: Mapping agents 1 and 2 consult WordNet in order to extend the concepts and properties with their inherited hypernym in the query. These hypernyms serve as variables in the hypothesis. For the concepts "paper" e.g. we have found that "article" and "communication" or "publication" are possible concepts that can appear in any of the external ontologies.
- Step 3: Mapping agents iterate through all concepts and properties from the ontologies and create sev-

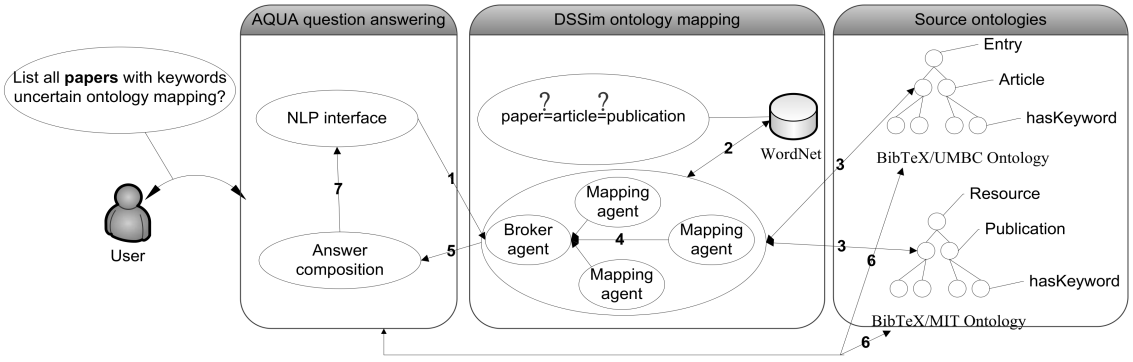


Figure 4. Overview of the mapping system

eral hypothesises that must be verified with finding evidences e.g.

$$Agent1 : H_n(mapping) =$$
$$Query\ \{paper,article,communation,publication\} \Longleftrightarrow$$
$$Ontology_{\text{MIT}}\ \{Article\}$$
$$and$$
$$Agent2 : H_n(mapping) =$$
$$Query\ \{paper,article,communation,publication\}$$
$$\Leftrightarrow\ Ontology_{\text{UMBC}}\ \{Publication\}$$

where Hn is the hypothesis for the mapping.
Further, we find supporting evidences for hypothesis. In this phase different syntactic and semantic similarity measures are used. These similarity measures are considered as different experts determining belief functions for the hypothesis. The last phase of this step is to combine the belief mass functions using Dempster's combination rule in order to form a coherent belief of the different experts on the hypothesises.

- Step 4: Mapping agents select the hypothesis in which they believe in most and sent it back to the broker agent. In our example the following mappings have been established:

$$Mapping_{\text{Query, MIT ontology}}(paper \leftrightarrow article)$$
$$Mapping_{\text{Query, UMBC ontology}}(paper \leftrightarrow publication)$$

- Step 5-6: The answer is composed for the user's query, which includes the relevant instances from the ontologies.

III. Uncertain reasoning and agent belief

Our proposed method works with two ontologies, which contain arbitrary number of concepts and their properties.

$$O_1 = \{C_1, .., C_n; P_1, .., P_n; I_1, .., I_n\}$$
$$O_2 = \{C_1, .., C_m; P_1, .., P_m; I_1, .., I_m\}$$

where O represents a particular ontology, C, P and I the set of concepts, properties and instances in the ontology.

In order to assess similarity we need to compare all concepts and properties from O_1 to all concepts and properties in O_2. Our similarity assessments, both syntactic and semantic produce a sparse similarity matrix where the similarity between C_n from O_1 and C_m in O_2 is represented by a particular similarity measure between the i and j elements of the matrix as follows:

$$SIM := (s_{i,j})_{n \times m}$$
$$1 \leq i \leq n\ and\ 1 \leq j \leq m$$

where SIM represents a particular similarity assessment matrix, s is a degree of similarity that has been determined by a particular similarity e.g. Jaccard or semantic similarity measure. We consider each measure as an "expert", which assess mapping precision based on its knowledge. Therefore we assume that each similarity matrix is a subjective assessment of the mapping what needs to be combined into a coherent view. If combined appropriately this combined view provides a more reliable and precise mapping than each separate mapping alone. However one similarity measure or some technique can perform particularly well for one pair of concepts or properties and particularly badly for another pair of concepts or properties, which has to be considered in any mapping algorithm.

In our ontology mapping framework each agent carries only partial knowledge of the domain and can observe it from its own perspective where available prior knowledge is generally uncertain. Our main argument is that knowledge cannot be viewed as a simple conceptualization of the world, but it has to represent some degree of interpretation. Such interpretation depends on the context of the entities involved in the process. This idea is rooted in the fact the different entities' interpretations are always subjective, since they occur according to an individual schema, which is than communicated to other individuals by a particular language. In order to represent these subjective probabilities in our system we use the Dempster-Shafer theory of evidence [19], which provides a mechanism for modelling and reasoning uncertain information in a numerical way, particularly when it is not possible to assign belief to a single element of a set of variables. Consequently the theory allows the user to represent uncertainty for knowledge representation,

because the interval between support and plausibility can be easily assessed for a set of hypothesizes. Missing data (ignorance) can also be modelled by Dempster-Shafer approach and additionally evidences from two or more sources can be combined using Dempster's rule of combination. The combined support, disbelief and uncertainty can each be separately evaluated. The main advantage of the Dempster-Shafer theory is that it provides a method for combining the effect of different learned evidences to establish a new belief by using Dempster's combination rule.

The following elements have been used in our system in order to model uncertainty:

Frame of Discernment(Θ) :finite set representing the space of hypothesizes. It contains all possible mutually exclusive context events of the same kind.

$$\Theta = \{H_1, ..., H_n, ...H_N\} \tag{1}$$

In our method Θ contains all possible mappings that have been assessed by the particular expert.

Evidence:available certain fact and is usually a result of observation. Used during the reasoning process to choose the best hypothesis in Θ. We observe evidence for the mapping if the expert detects that there is a similarity between C_n from O_1 and C_m in O_2.

Belief mass function (m): is a finite amount of support assigned to the subset of Θ. It represents the strength of some evidence and

$$\sum_{A \subseteq \Theta} m_i(A) = 1 \tag{2}$$

where m_i(A) is our exact belief in a proposition represented by A that belongs to expert *i*. The similarity algorithms itself produce these assignment based on different similarity measures. As an example consider that O_1 contains the concept "paper", which needs to be mapped to a concept "hasArticle" in O_2. Based on the WordNet we identify that the concept "article" is one of the inherited hypernyms of "paper", which according to both JaroWinkler(0.91) and Jaccard(0.85) measure [20] is highly similarity to "hasArticle" in O_2. Therefore after similarity assessment our variables will have the following belief mass value:

$$- \; m_{\text{expert1}}(O_1 \{paper, article, publication\}, \; O_2 \{hasArticle\}) = 0.85$$
$$- \; m_{\text{expert2}}(O_1 \{paper, article, publication\}, \; O_2 \{hasArticle\}) = 0.91$$

In practice we assess up to 8 inherited hypernyms similarities with different algorithms (considered as experts), which can be combined based on the combination rule in order to create a more reliable mapping. Once the combined belief mass functions have been assigned the following additional measures can be derived from the available information.

Belief: amount of justified support to A that is the lower probability function of Dempster, which accounts for all evidence E_k that supports the given proposition A.

$$belief_i(A) = \sum_{E_k \subseteq A} m_i(E_k) \tag{3}$$

An important aspect of the mapping is how one can make a decision over how different similarity measures can be combined and which nodes should be retained as best possible candidates for the match. To combine the qualitative similarity measures that have been converted into belief mass functions we use the Dempster's rule of combination and we retain the node where the belief function has the highest value.

Dempster's rule of combination:Suppose we have two mass functions $m_i(E_k)$ and $m_j(E_{k'})$ and we want to combine them into a global $m_{ij}(A)$. Following Dempster's combination rule

$$m_{ij}(A) = m_i \oplus m_j = \sum_{E_k E_{k'}} m_i(E_k) * m_j(E_{k'}) \tag{4}$$

where *i* and *j* represent two different agents.

The belief combination process is computationally very expensive and from an engineering point of view, this means that it not always convenient or possible to build systems in which the belief revision process is performed globally by a single unit. Therefore, applying multi agent architecture is an alternative and distributed approach to the single one, where the belief revision process is no longer assigned to a single agent but to a group of agents, in which each single agent is able to perform belief revision and communicate with the others. Our algorithm takes all the concepts and its properties from the different external ontologies and assesses similarity with all the concepts and properties in the query graph.

A. Voting and the best possible alternative

The idea of individual voting in order to resolve conflict and choose the best option available is not rooted in computer but political science. Democratic systems are based on voting as Condorcet jury theorem [21] [22] postulates that a group of voters using majority rule is more likely to choose the right action than an arbitrary single voter is. In these situations voters have a common goal, but do not know how to obtain this goal. Voters are informed differently about the performance of alternative ways of reaching it. If each member of a jury has only partial information, the majority decision is more likely to be correct than a decision arrived at by an individual juror. Moreover, the probability of a correct decision increases with the size of the jury. But things become more complicated when information is shared before a vote is taken. People then have to evaluate the information before making a collective decision. The same ideas apply for software agents especially if they need to reach a consensus on a particular issue. In case of ontology mapping where each agent can built up beliefs over the correctness of the mappings based on partial information we believe that voting can find the socially optimal choice. Software agents can use voting

to determine the best decision for agent society but in case voters make mistakes in their judgments, then the majority alternative (if it exists) is statistically most likely to be the best choice. The application of voting for software agents is a possible way to make systems more intelligent i.e. mimic the decision making how humans reach consensus decision on a problematic issue.

B. Fuzzy voting model

In ontology mapping the conflicting results of the different beliefs in similarity can be resolved if the mapping algorithm can produce an agreed solution, even though the individual opinions about the available alternatives may vary. We propose a solution for reaching this agreement by evaluating trust between established beliefs through voting, which is a general method of reconciling differences. Voting is a mechanism where the opinions from a set of votes are evaluated in order to select the alternatives that best represent the collective preferences. Unfortunately deriving binary trust like trustful or not trustful from the difference of belief functions is not so straightforward since the different voters express their opinion as subjective probability over the similarities. For a particular mapping this always involves a certain degree of vagueness hence the threshold between the trust and distrust cannot be set definitely for all cases that can occur during the process. Additionally there is no clear transition between characterising a particular belief highly or less trustful. Therefore our argument is that the trust membership or belief difference values, which are expressed by different voters can be modeled properly by using fuzzy representation. Before each agent evaluates the trust in other agent's belief over the correctness of the mapping it calculates the difference between its own and the other agent's belief. Depending on the difference it can choose the available trust levels e.g. if the difference in beliefs is 0.2 then the available trust level can be high and medium. We model these trust levels as fuzzy membership functions. In fuzzy logic the membership function $\mu(x)$ is defined on the universe of discourse U and represents a particular input value as a member of the fuzzy set i.e. $\mu(x)$ is a curve that defines how each point in the U is mapped to a membership value (or degree of membership) between 0 and 1. Our ontology mapping system models the conflict resolution as a fuzzy system where the system components are as follows:

1) Fuzzification of input and output variables: Fuzzification is the process of decomposing a system input and/or output into one or more fuzzy sets. We have experimented different types of curves namely the triangular, trapezoidal and gauss shaped membership functions. Each fuzzy set spans a region of input (or output) value graphed with the membership. Our selected membership functions overlap to allow smooth mapping of the system. The process of fuzzification allows the system inputs and outputs to be expressed in linguistic terms so that rules can be applied in a simple manner to express a complex system.

Definition 1: Belief difference is an input variable, which represents the agents own belief over the correctness of a mapping in order to establish mappings between concepts and properties in the ontology. During conflict resolution we need to be able to determine the level of difference. We propose three values for the fuzzy membership value $\mu(x) = \{small, average, large\}$

Definition 2: belief is an input variable, which described the amount of justified support to A that is the lower probability function of Dempster, which accounts for all evidence E_k that supports the given proposition A.

$$belief_i(A) = \sum_{E_k \subseteq A} m_i(E_k) \tag{5}$$

where m Demster's belief mass function represents the strength of some evidence i.e. $m(A)$ is our exact belief in a proposition represented by A. The similarity algorithms itself produce these assignment based on different similarity measures. We propose two values for the fuzzy membership value $\nu(x) = \{weak, strong\}$

Definition 3: Similarity is an input variable and is the result of some syntactic or semantic similarity measure. We propose three values for the fuzzy membership value $\xi(x) = \{low, average, high\}$

Definition 4: Low, medium and high trusts are output variables and represent the level of trust we can assign to the combination of our input variables. We propose three values for the fuzzy membership value $\tau(x) = \{low, medium, high\}$

2) Rule set: Fuzzy sets are used to quantify the information in the rule-base, and the inference mechanism operates on fuzzy sets to produce defuzzified values. Fuzzy systems map the inputs to the outputs by a set of $condition \rightarrow action$ rules i.e. rules that can be expressed in $If - Then$ form. For our conflict resolution problem we have defined four simple rules that ensure that each combination of the input variables produce output on more than one output i.e. there is always more than one initial trust level is assigned to any input variables. As an example consider a rule for cases when the trust level is defined as low:

"IF (beliefdifference IS large OR beliefdifference IS average) AND belief IS weak AND (similarity IS low OR similarity IS average) THEN trust IS low"

The rules we have initially defined are the most general ones. In our future research we intend to investigate the impact of more fine grained rules (i.e. more rules could be defined to cover overlapping areas of our fuzzy sets) on our conflict resolution.

3) Defuzzification method: After fuzzy reasoning we have the linguistic output variables, which need to be translated into a crisp (i.e. real numbers, not fuzzy sets) value. The objective is to derive a single crisp numeric value that best represents the inferred fuzzy values of the linguistic output variable. Defuzzification is such inverse transformation, which maps the output from the fuzzy domain back into the crisp domain. In our ontology mapping system we have selected the Center-of-Area (C-o-A) defuzzification method. The C-o-A method is often

referred to as the Center-of-Gravity method because it computes the centroid of the composite area representing the output fuzzy term. In our system the trust levels are proportional with the area of the membership functions therefore other defuzzification methods like Center-of-Maximum (C-o-M) or Mean-of-Maximum (M-o-M) does not correspond well to our requirements.

Definition 5: For representing trust in beliefs over similarities we have defined three membership functions, $\tau(x) = \{low, average, high\}$

in the beliefs over concept and property similarities in our ontology mapping system. Our main objective is to be able to resolve conflict between two beliefs in Dempster-Shafer theory, which can be interpreted qualitatively as one source strongly supports one hypothesis and the other strongly supports another hypothesis, where the two hypotheses are not compatible. Consider for example a situation where three agents have used WordNet as background knowledge and build their beliefs considering different concepts context, which was derived from the background knowledge e.g. agent 1 used the direct hypernyms, agent 2 the sister terms and agent 3 the inherited hypernyms. Based on string similarity measures a numerical belief value is calculated, which represent a strength of the confidence that the two terms are related to each other. The scenario is depicted in Table I.

TABLE I.
BELIEF CONFLICT DETECTION

Conflict detection	Belief 1	Belief 2	Belief 3
Obvious	0.85	0.80	0.1
Difficult	0.85	0.65	0.45

The values given in Table I are demonstrative numbers just for the purpose of providing an example. In our ontology mapping framework DSSim, the similarities are considered as subjective beliefs, which is represented by belief mass functions that can be combined using the Dempster's combination rule. This subjective belief is the outcome of a similarity algorithm, which is applied by a software agent for creating mapping between two concepts in different ontologies. In our ontology mapping framework different agents assess similarities and their beliefs in the similarities need to be combined into a more coherent result. However these individual beliefs in practice are often conflicting. In this scenario applying Dempster's combination rule to conflicting beliefs can lead to an almost impossible choice because the combination rule strongly emphasizes the agreement between multiple sources and ignores all the conflicting evidence through a normalization factor. The counter-intuitive results that can occur with Dempsters rule of combination are well known and have generated a great deal of debate within the uncertainty reasoning community. Different variants of the combination rule [23] have been proposed to achieve more realistic combined belief. Instead of proposing an additional combination rule we turned our attention to the root cause of the conflict itself namely how the uncertain information was produced in our model.

The fuzzy voting model was developed by Baldwin [24] and has been used in fuzzy logic applications. However, to our knowledge it has not been introduced in the context of trust management on the Semantic Web. In this section, we will briefly introduce the fuzzy voting model theory using a simple example of 10 voters voting against or in favour of the trustfulness of an another agent's belief over the correctness of mapping. In our ontology mapping framework each mapping agent can request a number of voting agents to help assessing how trustful the other mapping agent's belief is.

According to Baldwin [24] a linguistic variable is a quintuple $(L, T(L), U, G, \mu)$ in which L is the name of the variable, $T(L)$ is the term set of labels or words (i.e. the linguistic values), U is a universe of discourse, G is a syntactic rule and μ is a semantic rule or membership function. We also assume for this work that G corresponds to a null syntactic rule so that $T(L)$ consists of a finite set of words. A formalization of the fuzzy voting model can be found in [25].

Consider the set of words { Low_trust (L_t), Medium_trust (M_t) and High_trust (H_t) } as labels of a linguistic variable trust with values in $U = [0, 1]$. Given a set "m" of voters where each voter is asked to provide the subset of words from the finite set $T(L)$, which are appropriate as labels for the value u. The membership value $\chi_{\mu_(w)(u)}$ is taking the proportion of voters who include u in their set of labels which is represented by w.

The main objective when resolving conflict is to have sufficient number of independent opinions that can be consolidated. To achieve our objective we need to introduce more opinions into the system i.e. we need to add the opinion of the other agents in order to vote for the best possible outcome. Therefore we assume for the purpose of our example that we have 10 voters (agents). Formally, let us define

$$V = \{A1, A2, A3, A4, A5, A6, A7, A8, A9, A10\} \tag{6}$$

$$T(L) = \{L_t, M_t, H_t\}$$

The number of voters can differ however assuming 10 voters can ensure that

1) The overlap between the membership functions can proportionally be distributed on the possible scale of the belief difference [0..1]
2) The work load of the voters does not slow the mapping process down

Let us start illustrating the previous ideas with a small example - By definition consider three linguistic output variables L representing trust levels and T(L) the set of linguistic values as $T(L) = \{Low\_trust, Medium\_trust, High\_trust\}$. The universe of discourse is U, which is defined as $U = [0, 1]$. Then, we define the fuzzy sets per output variables $\mu(Low\_trust), \mu(Medium\_trust)$ and $\mu(High\_trust)$ for the voters where each voter has different overlapping trapezoidal, triangular or gauss

membership functions. The difference in the membership functions represented by the different vertices of the membership functions, which ensures that voters can introduce different opinions as they pick the possible trust levels for the same difference in belief.
The possible set of trust levels $L = TRUST$ is defined by the Table II. Note that in the table we use a short notation L_t means Low_trust, M_t means Medium_trust and H_t means High_trust. Once the input fuzzy sets (membership functions) have been defined the system is ready to assess the output trust memberships for the input values. Both input and output variables are real numbers on the range between [0..1]. Based on the difference of beliefs represented by a real number, own belief and similarity of the different voters the system evaluates the scenario. The evaluation includes the fuzzication, which converts the crisp inputs to fuzzy sets, the inference mechanism, which uses the fuzzy rules in the rule-base to produce fuzzy conclusions (e.g. the implied fuzzy sets), and the defuzzication block, which converts these fuzzy conclusions into the crisp outputs. Therefore each input (belief difference, belief and similarity) produces a possible defuzzified output (low, medium or high trust) for the possible output variables. Each defuzzified value can be interpreted as a possible trust level where the linguistic variable with the highest defuzzified value is retained in case more than one output variable is selected. As an example consider a case where the defuzzified output for belief difference between agent 1 and agent 2 with a value 0.67 has resulted in the situation described in Table II. Note that each voter has its own membership function where the level of overlap is different for each voter. Based on a concrete input the first voting agent could map the defuzzified variables into high, medium and low trust whereas tenth voting agent to only low trust.

TABLE II.
POSSIBLE VALUES FOR THE VOTING

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
L_t	L_t	L_t	L_t	L_t	L_t	L_t	L_t	L_t	L_t
M_t	M_t	M_t	M_t	M_t	M_t				
H_t	H_t	H_t							

Note that behind each trust level there is a real number, which represents the defuzzified value. These values are used to reduce the number of possible linguistic variables in order to obtain the vote for each voting agent. Each agent retains the lingiustic variable that represents the highest value and is depicted in Table III.

TABLE III.
VOTING

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
H_t	M_t	L_t	L_t	M_t	M_t	L_t	L_t	L_t	L_t

Taken as a function of x these probabilities form probability functions. They should therefore satisfy:

$$\sum_{w \in T(L)} Pr(L = w|x) = 1 \tag{7}$$

which gives a probability distribution on words:

$$\sum Pr(L = Low\_trust|x) = 0.6 \tag{8}$$

$$\sum Pr(L = Medium\_trust|x) = 0.3 \tag{9}$$

$$\sum Pr(L = High\_trust|x) = 0.1 \tag{10}$$

As a result of voting we can conclude that given the particular difference in beliefs (represented by a real number 0.67 in this example) the combination should not consider this belief of agent 2. This is because based on its difference compared to belief of agent 1 it turns out to be a distrustful assessment. The before mentioned process of taking the ”probability distributions on words” is then repeated as many times as needed. In fact, the process is repeated as many different beliefs we have for the similarity (i.e. as many as different similarity measures exist in the ontology mapping system).

C. Possible membership functions for conflict resolution

Membership functions in fuzzy systems are subjectively specied in an ad hoc (heuristic) manner from experience or intuition. This might be possible for a real time control system, however in our case it is difficult to find intuitive choice for the membership function or the combination of the membership functions. For our conflict resolution problem we have carried out experiments in order to select the best possible membership function combination that fit well to our problem.

We have chosen the trapezodial, triangular and gauss membership function and their combinations to represent our input and output variables. For each test have generated 300 scenarios, which contain random input variables (belief difference, belief and similarity) that maps to a single trust level i.e. output variable(high, medium or low trust). In addition we have defined nine combination of membership functions that describes our input and output variables. We repeated our experiment 1000 times regenerating the 300 scenarios in each iteration.

D. Results on the use of different memberships functions

Our experiments have shown that the the fuzzy conflict resolution is really sensitive on the input membership function. The best results can be achieved using triangular membership functions. In each experiments the average wrong answers are 121 and the minimum wrong answers are 109 whereas the maximum are 134 when choosing triangular input functions. The results are promising as we are able to resolve conflict in nearly 2/3 of the cases. In practice the real improvements in the ontology mapping quality can be foreseen where the number of conflict for the candidate mapping set is high. These

situations of course likely to occur where both source and target ontologies contain large number (up to 10.000) of concepts and properties. The selection of the output function does not influence the end result of the conflict resolution.

IV. Case study

Experimental comparison of ontology mapping systems is not a straightforward task as each system is usually designed to address a particular need from a specific domain. Authors have the freedom to hand pick some specific set of ontologies and demonstrate the strengths and weaknesses of their system carrying out some experiments with these ontologies. The problem is however that it is difficult to run the same experiments with another system and compare the two results. This problem has been acknowledged by the Ontology Mapping community and as a response to this need the Ontology Alignment Evaluation Initiative [7] has been set up in 2004. The evaluation was measured with recall, precision and F-Measure, which are useful measures that have a fixed range and meaningful from the mapping point of view. Recall is 100% when every relevant entity is retrieved. However it is possible to achieve 100% by simply returning every entity in the collection for every query. Therefore, recall by itself is not a good measure of the quality of a search engine. Precision is a measure of how well the engine performs in not returning non relevant documents. Precision is 100% when every entity returned to the user is relevant to the query. There is no easy way to achieve 100% precision other than in the trivial case where no document is ever returned for any query. Both precision and recall has a fixed range: 0.0 to 1.0 (or 0% to 100%). A good mapping algorithm must have a high recall to be acceptable for most applications. The most important factor in building better mapping algorithms is to increase precision without worsening the recall. In order to compare our system with other solutions we have participated in the OAEI competitions since 2006. Each year we have been involved in more tracks than the previous year. This gave us the possibility to test our mapping system on different domains including medical, agriculture, scientific publications, web directories, food and agricultural products and multimedia descriptions. The experiments were carried out to assess the efficiency of the mapping algorithms themselves. The experiments of the question answering (AQUA) using our mappings algorithms are out of the scope of this paper. Our main objective was to compare our system and algorithms to existing approaches on the same basis and to allow drawing constructive conclusions.

A. Benchmarks

The OAEI benchmark contains tests, which were systematically generated starting from some reference ontology and discarding a number of information in order to evaluate how the algorithm behave when this information is lacking. The bibliographic reference ontology (different classifications of publications) contained 33 named classes, 24 object properties, 40 data properties. Further each generated ontology was aligned with the reference ontology. The benchmark tests were created and grouped by the following criteria:

- Group 1xx: simple tests such as comparing the reference ontology with itself, with another not related (food domain) ontology or the same ontology in its restriction to OWL-Lite
- Group 2xx: systematic tests that were obtained by discarding some features from some reference ontology e.g. name of entities replaced by random strings, synonyms, name with different conventions, strings in another language than English, comments that can be suppressed or translated in another language, hierarchy that can be suppressed, expanded or flattened.
- Group 3xx: four real-life ontologies of bibliographic references that were found on the web e.g. BibTeX/MIT, BibTeX/UMBC

Figure 5 shows the 6 best performing systems out of 13 participants. We have ordered the systems based on the their the F-Value of the H-means because the H-mean unifies all results for the test and F-Value represents both precision and recall.

In the benchmark test we have performed in the upper mid range compared to other systems. Depending on the group of tests our system compares differently to other solutions:

- Group 1xx: Our results are nearly identical to the other systems.
- Group 2xx: For the tests where syntactic similarity can determine the mapping outcome our system is comparable to other systems. However where semantic similarity is the only way to provide mappings our systems provides less mappings compared to the other systems in the best six.
- Group 3xx: Considering the F-value for this group only 3 systems SAMBO, RIMOM and Lily are ahead.

The weakness of our system to provide good mappings when only semantic similarity can be exploited is the direct consequence of our mapping architecture. At the moment we are using four mapping agents where 3 carries our syntactic similarity comparisons and only 1 is specialised in semantics. However it is worth to note that our approach seems to be stable compared to our last years performance, as our precision recall values were similar in spite of the fact that more and more difficult tests have been introduced in 2008. As our architecture is easily expandable with adding more mapping agents it is possible to enhance our semantic mapping performance in the future.

B. Directory

The purpose of this track was to evaluate performance of existing alignment tools in real world taxonomy inte-

[7] http://oaei.ontologymatching.org/

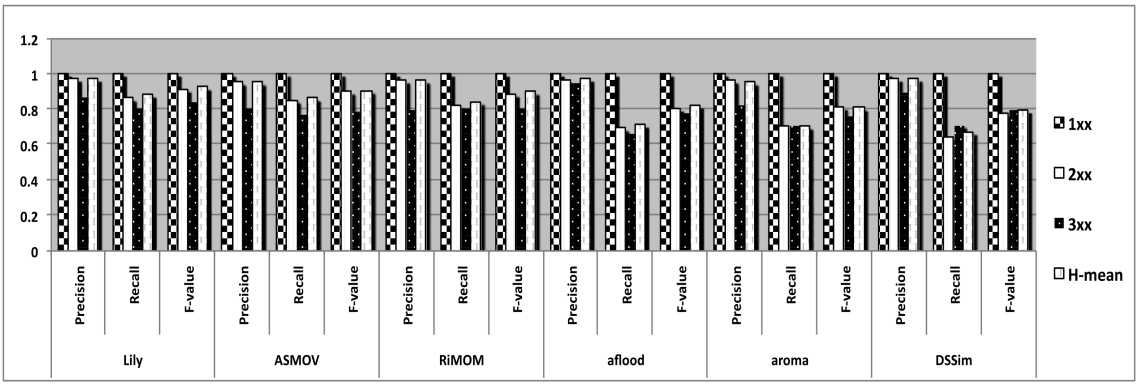


Figure 5. Best performing systems in the benchmarks track based on H-mean and F-value

gration scenario. Our aim is to show whether ontology alignment tools can effectively be applied to integration of "shallow ontologies". The evaluation dataset was extracted from Google, Yahoo and Looksmart web directories. The specific characteristics of the dataset are:

- More than 4500 of node matching tasks, where each node matching task is composed from the paths to root of the nodes in the web directories. Expert mappings for all the matching tasks.
- Simple relationships: Basically web directories contain only one type of relationship so called "classification relation".
- Vague terminology and modelling principles: The matching tasks incorporate the typical "real world" modelling and terminological errors.

These node matching tasks were represented by pairs of OWL ontologies, where classification relation is modelled as OWL *subClassOf* construct. Therefore all OWL ontologies are taxonomies (i.e. they contain only classes (without Object and Data properties) connected with subclass relation.

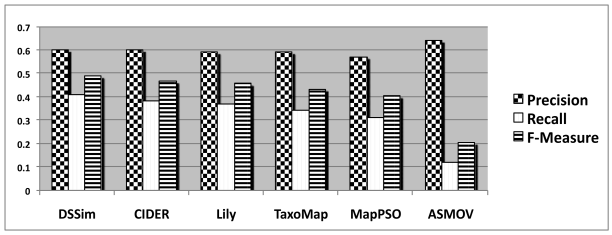


Figure 6. All participating systems in the directories track ordered by F-value

In the library track only 6 systems have participated in 2008. In terms of F-value DSSim has performed the best however the difference is marginal compared to the CIDER [26] or Lily systems. The concepts in the directory ontologies mostly can mostly be characterised as compound nouns e.g. "News_and_Media" and we need to process(split) them properly before consulting background knowledge in order to provide better mappings in the future.

C. Library

The objective of this track was to align two Dutch thesauri used to index books from two collections held by the National Library of the Netherlands. Each collection is described according to its own indexing system and conceptual vocabulary. On the one hand, the Scientific Collection is described using the GTT, a huge vocabulary containing 35.000 general concepts ranging from "Wolkenkrabbers (Sky-scrapers)" to "Verzorging (Care)". On the other hand, the books contained in the Deposit Collection are mainly indexed against the Brinkman thesaurus, containing a large set of headings (more than 5.000) that are expected to serve as global subjects of books. Both thesauri have similar coverage (there are more than 2.000 concepts having exactly the same label) but differ in granularity. For each concept, the thesauri provide the usual lexical and semantic information: preferred labels, synonyms and notes, broader and related concepts, etc. The language of both thesauri is Dutch, but a quite substantial part of Brinkman concepts (around 60%) come with English labels. For the purpose of the alignment, the two thesauri have been represented according to the SKOS model, which provides with all these features.

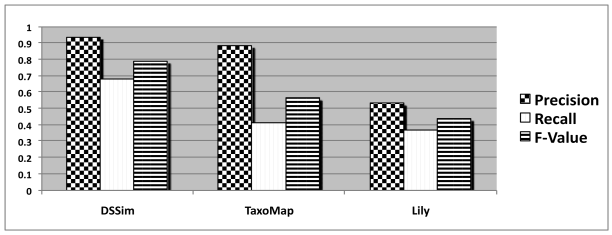


Figure 7. All participating systems in the library track ordered by F-value

In the library track DSSim has performed the best out of the 3 participating systems. The track is difficult partly because of its relative large size and because of its multilingual representation. However these ontologies contain related and broader terms therefore the mapping can be carried out without consulting multi lingual background knowledge. This year the organisers have provided instances as separate ontology as well however we did

not make use of it for creating our final mappings. For further improvements in recall and precision we will need to consider these additional instances in the future.

V. Strengths and weaknesses of our solution

Based on the OAEI experiments, we can conclude that our solution compares and scales well to other well established ontology mapping systems. Nevertheless it is clear (OAEI seems to share our opinion) that it is not possible to clearly define a "winner" on these yearly competitions. Each system has its strengths and weaknesses and they tend to perform differently on different domains. However we can define some criteria to determine where we perform well and on which areas do we need to make further progress.

1) Domain independence: This is a definite strength of our system. Our solution does not rely on predefined thresholds or parameters that needs to be changed from domain to domain. Several mapping systems utilise machine learning in order to determine these parameters however these solutions are likely to be dependent on the training set. DSSim uses WordNet as the background knowledge. This ensures that we can provide equivalent mappings on different domains. Nevertheless domain specific background knowledge can influence the results positively. The anatomy track has proved that systems that use domain specific background knowledge are far superior compared to the systems with general background knowledge. Nevertheless the drawback of these systems is that they cannot produce equally good results once the domain is changing. For example the AOAS system [27] performed the best on the anatomy track on the OAEI 2007 but they did not produce result in any other track as their system was fine tuned for the medical domain.
2) Conflict management: This area needs to be improved in our system. DSSim do manage conflicting beliefs over a particular mapping, which can occur when different agents have built up conflicting beliefs for the correctness of a mapping candidate. The problem occurs when we have already selected a mapping candidate and later on in the mapping process we add an another mapping that contradicts the previous one. Systems e.g. ASMOV, which try to detect conflicting mappings in the result-set can provide better overall results compared to our solution.
3) Mapping quality: DSSim does not produce always the best precision and recall for each track however our mapping quality is stable throughout different domains. We consider this as a strength of our system because we foresee different application domains where our solution can be used. In this context it is more important that we can produce equally good enough mappings.
4) Mapping system scalability: Due to our multi-agent architecture our solution scales well with medium and large domains alike. For example in the OAEI 2008 the largest ontologies were in the Very Large Cross-Lingual Resources track. DSSim was the only system that has participated in this track. Our solution can scale well for large domains because as the domain increases we can distribute the problem space between an increasing number of agents. Additionally our solution fits well to current hardware development trends, which predicts an increasing number of processor core in order to increase the computing power.
5) Traceability of the reasoning: Unfortunately this is a weakness of our system as we cannot guarantee that running the algorithm twice on the same domain we will always get exactly the same results. The reason is that our belief conflict resolution approach [28] uses fuzzy voting for resolving belief conflicts which can vary from case to case. Additionally beliefs are based on similarities between a set of source and target variables. The set of variables are deducted from the background knowledge, which can differ depending on the actual context of our query. Therefore it is not feasible to trace exactly why a particular mapping has been selected as good mapping compared to another candidate mappings.

VI. Related work

Several ontology mapping systems have been proposed to address the semantic data integration problem of different domains independently. In this paper we consider only those systems, which have participated in the OAEI (Ontology Alignment Evaluation Initiative) competitions and has been participated more than two tracks. There are other proposed systems as well however as the experimental comparison cannot be achieved we do not include them in the scope of our analysis. Lily [29] is an ontology mapping system with different purpose ranging from generic ontology matching to mapping debugging. It uses different syntactic and semantic similarity measures and combines them with the experiential weights. Further it applies similarity propagation matcher with strong propagation condition and the matching algorithm utilises the results of literal matching to produce more alignments. In order to assess when to use similarity propagation Lily uses different strategies, which prevents the algorithm from producing more incorrect alignments. ASMOV [30] has been proposed as a domain specific mapping tool in order to facilitate the integration of heterogeneous systems, using their data source ontologies. It uses different matchers and generates similarity matrices between concepts, properties, and individuals, including mappings from object properties to datatype properties. It does not combine the similarities but uses the best values to create a pre-alignment, which are then being semantically re-validated by the system. Mappings, which pass the semantic validation will be added to the

final alignment. ASMOV can use different background knowledge e.g. Wordnet or UMLS Metathesaurus(medical background knowledge) for the assessment of the similarity measures. RiMOM [31] is an automatic ontology mapping system, which models the ontology mapping problem as making decisions over entities with minimal risk. It uses the Bayesian theory to model decision making under uncertainty where observations are all entities in the two ontologies. Further it implements different matching strategies where each defined strategy is based on one kind of ontological information. RiMOM includes different methods for choosing appropriate strategies (or strategy combination) according to the available information in the ontologies. The strategy combination is conducted by a linear-interpolation method. In addition to the different strategies RiMOM uses similarity propagation process to refine the existing alignments and to find new alignments that cannot be found using other strategies. RiMOM is the only system other than DSSim in the OAEI contest that considers the uncertain nature of the mapping process however it models uncertainty differently from DSSim. RiMOM appeared for first time in the OAEI-2007 whilst DSSim appeared in the OAEI-2006. MapPSO [32] is a research prototype, which has been designed to address the need for highly scalable, massively parallel tool for both large scale and numerous ontology alignments. MapPSO method models the ontology alignment problem as an optimisation problem. It employs a population based optimisation paradigm based on social interaction between swarming animals, which provides the best answer being available at that time. Therefore it is especially suitable for providing answers under time constraint like the ontology mapping. MapPSO employs different syntactic and semantic similarity measures and combines the available base distances by applying the Ordered Weighted Average(OWA) [33] aggregation of the base distances. It aggregates the components by ordering the base distances and applying a fixed weight vector. The motivation of the MapPSO system is identical with one of the motivations of the DSSim namely to address the need of scalable mapping solutions for large scale ontologies. Surprisingly MapPSO did not participate in the Very Large Cross Lingual Resources track (especially designed for large scale thesauri) therefore experimental comparison cannot be achieved from this point of view. TaxoMap [34] is an alignment tool, which aims is to discover rich correspondences between concepts with performing oriented alignment from a source to a target ontology taking into account labels and sub-class descriptions. It uses a part-of-speech [35] and lemma information, which enables to take into account the language, lemma and an use word categories in an efficient way. TaxoMap performs a linguistic similarity measure between labels and description of concepts and it has been designed to process large scale ontologies by using partitioning techniques. TaxoMap however does not process instances, which can be a drawback in several situations. SAMBO and SAMBOdtf [36] is a general framework for ontology matching. The methods and techniques used in the framework are general and applicable to different areas nevertheless SAMBO has been designed to align biomedical ontologies. Their algorithm includes one or several matchers, which calculate similarity values between the terms from the different source ontologies. These similarities are then filtered and combined as a weighted sum of the similarity values computed by different matchers.

VII. Conclusions

This paper presented the main challenges for an alignment system in the context of question-answering. The challenges related to the data or information representation, quality and volume are addressed with introducing uncertain reasoning and representation when the available information is interpreted by our system. Our approach tries to establish an interpretation of the available information and avoids the usage of heuristics or any domain specific rules. To achieve this interpretation we have utilised Dempster-Shafer theory for managing the reasoning with vague information and have introduced fuzzy voting model for resolving conflicts during the interpretation of the Semantic Web data. Concerning the challenges related to the nature of the systems from the generic and intelligence point of view our proposed architecture is conceived to be able to exhibit a kind of "machine intelligence" through the multi-agent architecture, which is a form of collective intelligence that can emerge from the collaboration and competition of many software agents. Further we have also introduced DSSim and our performance in the benchmarks, directory and library tracks of the OAEI-2008 evaluation. The performance of our DSSim was the best among participants in the library track in 2008. Our system is conceived to be a generic mapping tool and such the performance still varies slightly across different domains. These variations and comparisons with other mappings systems are accessible from the OAEI workshop proceedings. Our participation in the Ontology Alignment Evaluation Initiative was an excellent opportunity to test and compare our system with other solutions and helped a great deal in identifying the future possibilities that needs to be investigated further.

References

[1] P. Shvaiko and J. Euzenat, "Ten challenges for ontology matching," University of Trento, Tech. Rep. DISI-08-042, 2008.
[2] J. Euzenat and P. Shvaiko, *Ontology matching*. Heidelberg (DE): Springer-Verlag, 2007.
[3] D. Beckett, "Rdf/xml syntax specification." [Online]. Available: http://www.w3.org/RDF/
[4] D. L. McGuinness and F. van Harmelen, "Owl web ontology language." [Online]. Available: http://www.w3.org/TR/owl-features/
[5] A. Miles and S. Bechhofer, "Skos simple knowledge organization system." [Online]. Available: http://www.w3.org/TR/skos-reference/
[6] M. Lenzerini, D. Milano, and A. Poggi, "Ontology representation & reasoning," Universit di Roma La Sapienza, Roma, Italy, Tech. Rep. NoE InterOp (IST-508011), 2004.

[7] R. Y. Wang, H. B. Kon, and S. E. Madnick, "Data quality requirements analysis and modeling," in *Proceedings of the Ninth International Conference on Data Engineering*, 1993, pp. 670–677.

[8] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, pp. 86–95, 1996.

[9] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," *ACM Computing Surveys*, vol. 18, no. 4, pp. 323–364, 1986.

[10] W. Carlo, D. Tharam, R. Wenny, and C. Elizabeth, "Large scale ontology visualisation using ontology extraction," *International Journal of Web and Grid Services*, vol. 1, no. 1, pp. 113–135, 2005.

[11] A. Flahive, B. O. Apduhan, J. W. Rahayu, and D. Taniar, "Large scale ontology tailoring and simulation in the semantic grid environment," *International Journal of Metadata, Semantics and Ontologies*, vol. 1, no. 4, pp. 265–281, 2006.

[12] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Šváb-Zamazal, and V. Svátek, "First results of the ontology alignment evaluation initiative 2008," in *The 3rd International Workshop on Ontology Matching*, 2008.

[13] H. Brugman, V. Malaisé, and L. Gazendam, "A web based general thesaurus browser to support indexing of television and radio programs," in *the 5th international conference on Language Resources and Evaluation (LREC 2006)*, 2006.

[14] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, and J. Sachs, "Swoogle: A search and metadata engine for the semantic web," in *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.

[15] S. M. Falconer and M.-A. D. Storey, "A cognitive support framework for ontology mapping," in *Proceedings of 6th International Semantic Web Conference (ISWC2007)*, 2007, pp. 114–127.

[16] M. Nagy and M. Vargas-Vera, "Reasoning representation and visualisation framework for ontology mapping using 3d modeling," in *proceedings of the 4th International Conference on Interdisciplinarity in Engineering (INTERENG 2009)*, 2009.

[17] M. Nagy, M. Vargas-Vera, and E. Motta, "Multi agent ontology mapping framework in the aqua question answering system," in *MICAI 2005: Advances in Artificial Intelligence, 4th Mexican International Conference on Artificial Intelligence*, 2005, pp. 70–79.

[18] M. Vargas-Vera and E. Motta, "Aqua - ontology-based question answering system," in *Third International Mexican Conference on Artificial Intelligence (MICAI-2004)*, 2004.

[19] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[20] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of Information Integration on the Web (IIWeb 2003)*, 2003.

[21] D. Austen-Smith and J. S. Banks, "Information aggregation, rationality, and the condorcet jury theorem," *The American Political Science Review*, vol. 90, no. 1, pp. 34–45, 1996.

[22] H. P. Young, "Condorcet's theory of voting," *The American Political Science Review*, vol. 82, no. 4, pp. 1231–1244, 1988.

[23] K. Sentz and S. Ferson, "Combination of evidence in dempster-shafer theory," Systems Science and Industrial Engineering Department, Binghamton University, Tech. Rep., 2002.

[24] J. F. Baldwin, *Mass assignment Fundamentals for computing with words*, ser. Selected and Invited Papers from the Workshop on Fuzzy Logic in Artificial Intelligence, Lecture Notes In Computer Science. Springer-Verlag, 1999, vol. 1566, pp. 22–44.

[25] J. Lawry, "A voting mechanism for fuzzy logic," *International Journal of Approximate Reasoning*, vol. 19, pp. 315–333, 1998.

[26] J. Gracia and E. Mena, "Ontology matching with cider:evaluation report for the oaei 2008," in *The 3rd International Workshop on Ontology Matching*, 2008.

[27] S. Zhang and O. Bodenreide, "Hybrid alignment strategy for anatomical ontologies results of the 2007 ontology alignment contest," in *The Second International Workshop on Ontology Matching*, 2007.

[28] M. Nagy, M. Vargas-Vera, and E. Motta, "Introducing fuzzy trust for managing belief conflict over semantic web data," in *Fourth International Workshop on Uncertain Reasoning on the Semantic Web*, 2008.

[29] P. Wang and B. Xu, "Lily: Ontology alignment results for oaei 2008," in *The 3rd International Workshop on Ontology Matching*, 2008.

[30] Y. R. Jean-Mary and M. R. Kabuka, "Asmov: Ontology alignment with semantic validation," in *Joint SWDB-ODBIS Workshop*, 2007.

[31] J. Tang, J. Li, B. Liang, X. Huang, Y. Li, and K. Wang, "Using bayesian decision for ontology mapping," *Web Semantics*, 2006.

[32] J. Bock and J. Hettenhausen, "Mappso results for oaei 2008," in *The 3rd International Workshop on Ontology Matching*, 2008.

[33] Q. Ji, P. Haase, and G. Qi, "Combination of similarity measures in ontology matching by owa operator," in *The 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'08)*, 2008.

[34] F. Hamdi, H. Zargayouna, B. Safar, and C. Reynaud, "Taxomap in the oaei 2008 alignment contest," in *The 3rd International Workshop on Ontology Matching*, 2008.

[35] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*, 1994.

[36] P. Lambrix and H. Tan, "Sambo - a system for aligning and merging biomedical ontologies," *Journal of Web Semantics, Special issue on Semantic Web for the Life Sciences*, 2006.

Maria Vargas-Vera received her PhD from the Artificial Intelligence Department at Edinburgh University in 1995 and she was awarded Fellow of the British Computer Society (FBCS) from November 2009. Dr. Vargas-Vera is a Lecturer in Computing at the Open University, England UK. Her current research focuses on Automatic Construction of Ontologies from Text, Ontology Mapping and E-Learning Applications using Semantic Web Technologies. Dr Vargas-Vera has published many research papers in international conferences and journals and she is also a member of program committees of international conferences and workshops. Dr Vargas-Vera is an Associated Editor of the Revista Iberoamericana Computacion y Sistemas, the International Journal of Knowledge and Learning (IJKL) and the Journal of Emerging Technologies in Web Intelligence (JETWI).

Miklos Nagy is a Ph.D. candidate at the Open University's Knowledge Media Institute. His research interests are Uncertain Reasoning, Ontology Mapping, Multi-agent systems and information integration using Semantic Web technologies. His current research focuses on the development of intelligent multi-agent systems that can exploit the emerging Semantic Web's large-scale data. Miklos Nagy received his MSc in Information Engineering from the University of Miskolc, Hungary.

A Survey of Text Summarization Extractive Techniques

Vishal Gupta
University Institute of Engineering & Technology,
Computer Science & Engineering, Panjab University Chandigarh, India,
Email: vishal@pu.ac.in

Gurpreet Singh Lehal
Department of Computer Science,
Punjabi University Patiala, Punjab, India,
Email: gslehal@yahoo.com

***Abstract*— Text Summarization is condensing the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. In this paper, a Survey of Text Summarization Extractive techniques has been presented.**

***Index Terms*—Text Summarization, extractive summary, abstractive summary**

I. INTRODUCTION

Text summarization [1] has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age. It is very difficult for human beings to manually summarize large documents of text. There is an abundance of text material available on the internet. However, usually the Internet provides more information than is needed. Therefore, a twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning.

A summary [4] can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences.

An Abstractive summarization [32][33] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. This paper focuses on extractive text summarization methods.

Extractive summaries [2]are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favorably positioned" content. Such an approach thus avoids any efforts on deep text understanding. They are conceptually simple, easy to implement.

Extractive text summarization process [31] can be divided into two steps: 1) Pre Processing step and 2) Processing step.

Pre Processing is structured representation of the original text. It usually includes: a) Sentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence. b) Stop-Word Elimination—Common words with no semantics and

Manuscript received January 12, 2010; revised March 22, 2010; accepted April 29, 2010.
Corresponding author: Vishal Gupta

doi:10.4304/jetwi.2.3.258-268

which do not aggregate relevant information to the task are eliminated. c) Stemming—The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics.

In Processing step, features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary.

Problems with the extractive summary [46] [47] are:

1. Extracted sentences usually tend to be longer than average. Due to this, parts of the segments that are not essential for summary also get included, consuming space.
2. Important or relevant information is usually spread across sentences, and extractive summaries cannot capture this (unless the summary is long enough to hold all those sentences).
3. Conflicting information may not be presented accurately.
4. Pure extraction often leads to problems in overall coherence of the summary—a frequent issue concerns "dangling" anaphora. Sentences often contain pronouns, which lose their referents when extracted out of context. Worse yet, stitching together decontextualized extracts may lead to a misleading interpretation of anaphors (resulting in an inaccurate representation of source information, i.e., low fidelity). Similar issues exist with temporal expressions. These problems become more severe in the multi-document case, since extracts are drawn from different sources. A general approach to addressing these issues involves post-processing extracts, for example, replacing pronouns with their antecedents, replacing relative temporal expression with actual dates, etc.

Problems with the abstractive summary [46] are:

The biggest challenge for abstractive summary is the representation problem. Systems' capabilities are constrained by the richness of their representations and their ability to generate such structures—systems cannot summarize what their representations cannot capture. In limited domains, it may be feasible to devise appropriate structures, but a general-purpose solution depends on open-domain semantic analysis. Systems that can truly "understand" natural language are beyond the capabilities of today's technology.

Summary evaluation [34][36][37] is a very important aspect for text summarization. Generally, summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation and extrinsic methods measure the same through a task-based [35] performance measure such the information retrieval-oriented task.

Newsblaster is a good example of a text summarizer, that helps users find the news that is of the most interest to them. The system automatically collects, clusters, categorizes, and summarizes news from several sites on the web (CNN, Reuters, Fox News, etc.) on a daily basis, and it provides users a user-friendly interface to browse the results.

II. TEXT SUMMARIZATION EARLY HISTORY

Interest in automatic text summarization, arose as early as the fifties. An important paper of these days is the one in 1958, suggested to weight the sentences of a document as a function of high frequency words[7], disregarding the very high frequency common words. Automatic text summarization system [8] in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights:

1. Cue Method: This is based on the hypothesis that the relevance of a sentence is computed by the presence or absence of certain cue words in the cue dictionary.
2. Title Method: Here, the sentence weight is computed as a sum of all the content words appearing in the title and (sub-) headings of a text.
3. Location Method: This method is based on the assumption that sentences occurring in initial position of both text and individual paragraphs have a higher probability of being relevant. The results showed, that the best correlation between the automatic and human-made extracts was achieved using a combination of these three latter methods.

The Trainable Document Summarizer [9] in 1995 performs sentence extracting task, based on a number of weighting heuristics. Following features were used and evaluated:

1. Sentence Length Cut-O Feature: sentences containing less than a pre-specified number of words are not included in the abstract
2. Fixed-Phrase Feature: sentences containing certain cue words and phrases are included
3. Paragraph Feature: this is basically equivalent to Location Method feature in [8]
4. Thematic Word Feature: the most frequent words are defined as thematic words. Sentence scores are functions of the thematic words' frequencies
5. Uppercase Word Feature: upper-case words (with certain obvious exceptions) are treated as thematic words, as well.

A Corpus was used in this method, which contained 188 document/summary pairs from 21 publications in a scientific/technical domain. The summaries were produced by professional experts and the sentences occurring in the summaries were aligned to the original document texts, indicating also the degree of similarity as mentioned earlier, the vast majority (about 80%) of the summary sentences could be classified as direct sentence matches.

The ANES text extraction system [10] in 1995 is a system that performs automatic, domain-independent condensation of news data. The process of summary generation has four major constituents:

1. Corpus analysis: this is mainly a calculation of the tf*idf -weights for all terms

2. Statistical selection of signature words: terms with a high tf*idf-weight
plus headline-words
3. Sentence weighting: summing over all signature word weights, modifying the weights by some other factors, such as relative location
4. Sentence selection: Selecting high scored sentences.

Hidden Markov Models (HMMs) [11]: As prove to be a mathematically sound frame-work for document retrieval. If one approaches the task of text abstracting from such a probabilistic modeling perspective, it might well be possible that HMMs could be employed for this purpose, as well.

Clustering: Building links [12] and/or clusters between index terms, phrases and/or other subparts of the documents has been employed by standard information retrieval. Although this is not an issue in any of the above mentioned abstracting systems, it seems to be worth of consideration when building such systems.

III. FEATURES FOR EXTRACTIVE TEXT SUMMARIZATION

Some features [2][5][29] to be considered for including a sentence in final summary are:

A. Content word (Keyword) feature:

Content words or Keywords are usually nouns and determined using tf × idf measure. Sentences having keywords are of greater chances to be included in summary. Another keyword extraction method [23][31] is given below, having three modules:
1) Morphological Analysis
2) Noun Phrase (NP) Extraction and Scoring
3) Noun Phrase (NP) Clustering and Scoring
Figure1 shows a pictorial representation of the keyword extraction method.

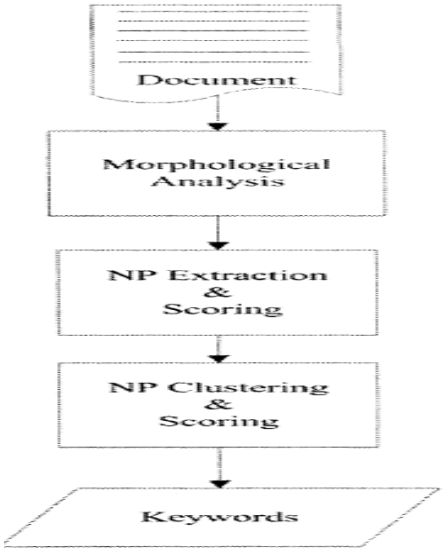


Figure 1. Keyword extraction method

B. Title word feature:

Sentences containing words that appear in the title are also indicative of the theme of the document. These sentences are having greater chances for including in summary.

C. Sentence location feature:

Usually first and last sentence of first and last paragraph of a text document are more important and are having greater chances to be included in summary.

D. Sentence Length feature:

Very large and very short sentences are usually not included in summary.

E. Proper Noun feature:

Proper noun is name of a person, place and concept etc. Sentences containing proper nouns are having greater chances for including in summary.

F. Upper-case word feature:

Sentences containing acronyms or proper names are included.

G. Cue-Phrase Feature:

Sentences containing any cue phrase (e.g. “in conclusion”, “this letter”, “this report”, “summary”, “argue”, “purpose”, “develop”, “attempt” etc.) are most likely to be in summaries.

H. Biased Word Feature:

If a word appearing in a sentence is from biased word list, then that sentence is important. Biased word list is previously defined and may contain domain specific words.

I. Font based feature:

Sentences containing words appearing in upper case, bold, italics or Underlined fonts are usually more important.

J. Pronouns:

Pronouns such as “she, they, it” cannot be included in summary unless they are expanded into corresponding nouns.

K. Sentence-to-Sentence Cohesion:

For each sentence s compute the similarity between s and each other sentence s’ of the document, then add up those similarity values, obtaining the raw value of this feature for s. The process is repeated for all sentences.

L. Sentence-to-Centroid Cohesion:

For each sentence s as compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence.

M. Occurrence of non-essential information:

Some words are indicators of non-essential information. These words are speech markers such as "because", "furthermore", and "additionally", and typically occur in the beginning of a sentence. This is also a binary feature, taking on the value "true" if the sentence contains at least one of these discourse markers, and "false" otherwise.

N. Discourse analysis:

Discourse level information [38], in a text is one of good feature for text summarization. In order to produce a coherent, fluent summary, and to determine the flow of the author's argument, it is necessary to determine the overall discourse structure of the text and then removing sentences peripheral to the main message of the text.

These features are important as, a number of methods of text summarization are using them. These features are covering statistical and linguistic characteristics of a language.

IV. EXTRACTIVE SUMMARIZATION METHODS

Extractive summarizers [13][14][30] aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary.

A. Term Frequency-Inverse Document Frequency (TF-IDF) method:

Bag-of-words model is built at sentence level, with the usual weighted term-frequency and inverse sentence-frequency paradigm [16], where sentence-frequency is the number of sentences in the document that contain that term. These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary. This is a direct adaptation of Information Retrieval paradigm to summarization. Summarization is query-specific, but can be adapted to be generic as described below.

To generate a generic summary, non stop-words that occur most frequently in the document(s) may be taken as the query words. Since these words represent the theme of the document, they generate generic summaries. Term-frequency is usually 0 or 1 for sentences—since normally the same content-word does not appear many times in a given sentence. If users create query words the way they create for information retrieval, then the query based summary generation would become generic summarization.

B. Cluster based method:

Documents are usually written such that they address different topics one after the other in an organized manner. They are normally broken up explicitly or implicitly into sections. This organization applies even to summaries of documents. It is intuitive to think that summaries should address different "themes" appearing in the documents. Some summarizers incorporate this aspect through clustering. If the document collection for which summary is being produced is of totally different topics, document clustering becomes almost essential to generate a meaningful summary.

Documents are represented using term frequency-inverse document frequency (TF-IDF) [17] of scores of words. Term frequency used in this context is the average number of occurrences (per document) over the cluster. IDF value is computed based on the entire corpus. The summarizer takes already clustered documents as input. Each cluster is considered a theme. The theme is represented by words with top ranking term frequency, inverse document frequency (TF-IDF) scores in that cluster.

Sentence selection is based on similarity of the sentences to the theme of the cluster C_i .The next factor that is considered for sentence selection is the location of the sentence in the document (L_i). In the context of newswire articles, the closer to the beginning a sentence appears, the higher its weight age for inclusion in summary. The last factor that increases the score of a sentence is its similarity to the first sentence in the document to which it belongs (F_i).

The overall score (Si) of a sentence i is a weighted sum of the above three factors:

$$S_i = W_1 * C_i + W_2 * F_i + W_3 * L_i \quad \text{.....(2)}$$

where S_i is the score of sentence $C_{i,,}$ F_i are the scores of the sentence i based on the similarity to theme of cluster and first sentence of the document it belongs to, respectively. L_i is the score of the sentence based on its location in the document. $w_{1,}$ w_2 andw_3 are the weights for linear combination of the three scores. Note the similarity between the sentence score in equations (1) and (2). The role of F in (2) is similar to that of T in (1). The difference however, is that $S_{i,}$ in (2) is further re-scored using a redundancy factor. Once the documents are clustered, sentence selection from within the cluster to form its summary is local to the documents in the cluster. The IDF value based on the corpus statistics seems counter-intuitive. A better choice may be to take the Average-TF alone to determine the theme of the cluster, and then rely on the "anti redundancy" factor to cover the important 'themes' within the cluster.

C. Graph theoretic approach:

As seen in the previous methods, the first step involved in the process of summarizing one or more documents is identifying the issues or topics addressed in the document. Graph theoretic representation [18] of passages provides a method of identification of these themes. After the common preprocessing steps, namely, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph.

There is a node for every sentence. Two sentences are connected with an edge if the two sentences share some common words, or in other words, their (cosine, or such) similarity is above some threshold. This representation yields two results: The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub graphs), form distinct topics covered in the documents. This allows a choice of coverage in the summary. For query-specific summaries, sentences may be selected only from the pertinent sub graph, while for generic summaries, representative sentences may be chosen from each of the sub-graphs.

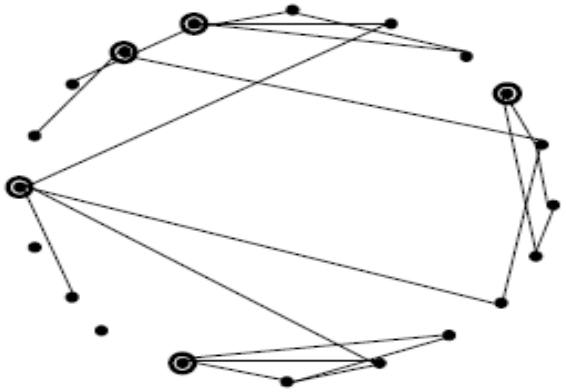


Figure 2. Graph theoretic approach

The second result yielded by the graph-theoretic method is the identification of the important sentences in the document. The nodes with high cardinality (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference to be included in the summary. Figure2 shows an example graph for a document. It can be seen that there are about 3-4 topics in the document; the nodes that are encircled can be seen to be informative sentences in the document, since they share information with many other sentences in the document. The graph theoretic method may also be adapted easily for visualization of inter- and intra-document similarity.

D. Machine Learning approach

Given a set of training document and their extractive summaries, the summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess. The classification probabilities are learnt statistically [3] from the training data, using Bayes' rule:

$$P(s \in S \mid F1, F2, ..., FN) = P(F1, F2, ..., FN \mid s \in S) * P(s \in S) / P(F1, F2, ..., FN)$$

where s is a sentence from the document collection, F1, F2…FN are features used in classification. S is the summary to be generated, and P (s∈< S | F1, F2, ..., FN) is the probability that sentence s will be chosen to form the summary given that it possesses features F1,F2…FN.

E. LSA Method

Singular Value Decomposition (SVD) [13] is a very powerful mathematical tool that can find principal orthogonal dimensions of multidimensional data. It has applications in many areas and is known by different names: Karhunen-Loeve Transform in image processing, Principal Component Analysis (PCA) in signal processes and Latent Semantic Analysis (LSA) in text processing. It gets this name LSA because SVD applied to document-word matrices, groups documents that are semantically related to each other, even when they do not share common words.

Words that usually occur in related contexts are also related in the same singular space. This method can be applied to extract the topic-words and content-sentences from documents. The advantage of using LSA vectors for summarization rather than the word vectors is that conceptual (or semantic) relations as represented in human brain are automatically captured in the LSA, while using word vectors without the LSA transformation requires design of explicit methods to derive conceptual relations. Since SVD finds principal and mutually orthogonal dimensions of the sentence vectors, picking out a representative sentence from each of the dimensions ensures relevance to the document, and orthogonality ensures non-redundancy. It is to be noted that this property applies only to data that has principal dimensions inherently—however, LSA would probably work since most of the text data has such principal dimensions owing to the variety of topics it addresses.

F. An approach to concept-obtained text summarization

The idea of this approach is to obtain concepts of words based on HowNet [19][20], and use concept as feature, instead of word. This approach uses conceptual vector space model to form a rough summarization, and then calculate degree of semantic similarity of sentence for reducing its redundancy. A good summary system should extract the diverse topics of the document while keeping redundancy to a minimum. This method consists of the following three main stages:

Stage 1: Using Hownet as tool to obtain concept of text, and establishing conceptual vector space model.

Stage 2: Calculate importance of concept based on conceptual vector space model.

Stage 3: Generate the final summary by calculating importance of sentence and reducing the redundancy of summarization.

G. Text summarization with neural networks

This method involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each

sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network [21] learns the patterns inherent in sentences that should be included in the summary and those that should not be included. It uses three-layered Feed forward neural network, which has been proven to be a universal function approximator.

The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included. The Neural Network [27] after Training is shown in figure3.

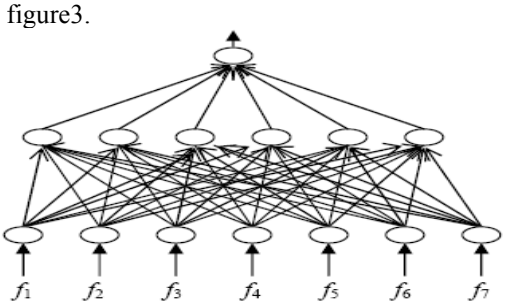


Figure 3. Neural Network after Training

Once the network has learned the features that must exist in summary sentences, we need to discover the trends and relationships among the features that are inherent in the majority of sentences. This is accomplished by the feature fusion phase, which consists of two steps: 1) eliminating uncommon features; and 2) collapsing the effects of common features. The connections having very small weights after training can be pruned without affecting the performance of the network. As a result, any input or hidden layer neuron having no emanating connections can be safely removed from the network. In addition, any hidden layer neuron having no abutting connections can be removed. This corresponds to eliminating uncommon features from the network [27] as shown in figure4.

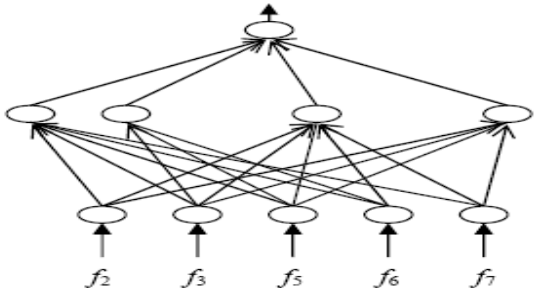


Figure 4. Neural Network after Pruning

The hidden layer activation values for each hidden layer neuron are clustered utilizing an adaptive clustering technique. Each cluster is identified by its centroid and frequency. The activation value of each hidden layer neuron is replaced by the centroid of the cluster, which the activation value belongs to. This corresponds to collapsing the effects of common features. The combination of these two steps corresponds to

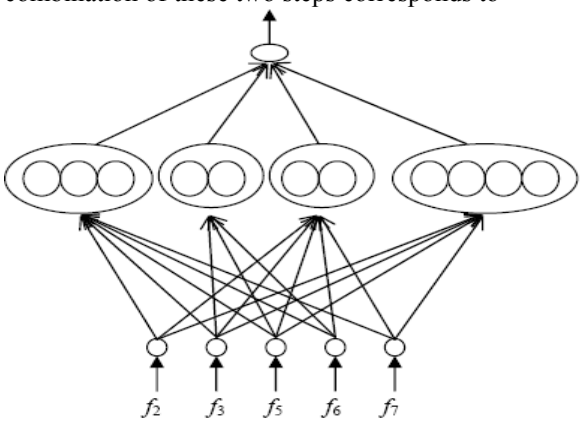


Figure 5. Neural Network after feature fusion

generalizing the effects of features, as a whole, and providing control parameters for sentence ranking. The Neural Network [27] after feature fusion is shown in figure5.

H. Automatic text summarization based on fuzzy logic

This method considers each characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system[2][22]. Then, it enters all the rules needed for summarization, in the knowledge base of system. After ward, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using IF-THEN rules according to the feature criteria.

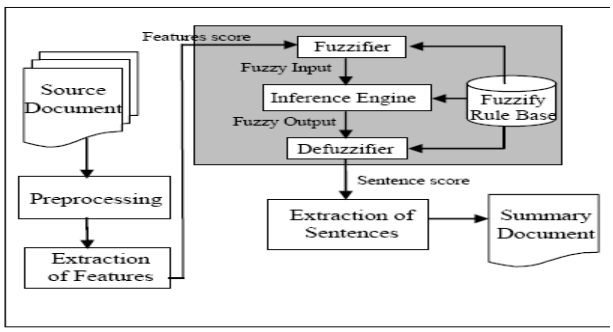


Figure 6. Text summarization based on fuzzy logic system architecture

Text summarization based on fuzzy logic system architecture [28] is shown in figure6. Fuzzy logic system design usually implicates selecting fuzzy rules and

membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system. The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, crisp inputs are translated into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IFTHEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score.

I. Text summarization using regression for estimating feature weights

Mathematical regression [6] is a good model to estimate the text feature weights. In this model, a mathematical function can relate output to input. The feature parameters of many manually summarized English documents are used as independent input variables and corresponding dependent outputs are specified in training phase. A relation between inputs and outputs is established. Then testing data are introduced to the system model for evaluation of its efficiency. In matrix notation we can represent regression as follow:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ . \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & .... & X_{010} \\ . & . & .... & . \\ . & . & .... & . \\ X_{m1} & X_{m2} & .... & X_{m10} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ . \\ w_m \end{bmatrix}$$

Where

$[Y]$ is output vector.

$[X]$ is the input matrix (feature parameters)

$[w]$ is linear statistical model of system (the weights $w_1, w_2 \ldots\ldots\ldots w_{10}$ in the equation)

m is total number of sentences in the training corpus

J. Multi-document extractive summarization

Multi document extractive summarization deals with extraction of summarized information from multiple texts written about the same topic. Resulting summary report allows individual users, so as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. Multi-document summarization creates information reports that are both concise and comprehensive. With different opinions being put together & outlined, every topic is described from multiple perspectives within a single document.

NeATS [15] is a multi-document summarization system that attempts to extract relevant or interesting portions from a set of documents about some topic and present them in coherent order. It is an extraction-based multi-document summarization system. Given an input of a collection of sets of newspaper articles, NeATS generates summaries in three stages: content selection, filtering, and presentation.

The goal of content selection is to identify important concepts mentioned in a document collection. In a key step for locating important sentences, NeATS computes the likelihood ratio to identify key concepts in unigrams, bigrams, and trigrams, using the on- topic document collection as the relevant set and the off-topic document collection as the irrelevant set. With the individual key concepts available, these concepts are clustered in order to identify major subtopics within the main topic. Clusters are formed through strict lexical connection. Each sentence in the document set is then ranked, using the key concept structures.

NeATS uses three different filters: sentence position, stigma words, and maximum marginal relevancy. Sentence position is a good content filter, that only retains the leading 10 sentences. Some sentences start with stigma words like:

- Conjunctions (e.g., but, although, however)
- The verb *say* and its derivatives
- Quotation marks
- Pronouns such as he, she, and they

usually cause discontinuity in summaries. The scores of these sentences are reduced to avoid including them in short summaries. Redundancy issue is addressed in maximum marginal relevancy filter. A sentence is added to the summary if and only if its content has less than *X* percent overlap with the summary. The overlap ratio is computed using simple stemmed word overlap and the threshold *X* is set empirically.

Hub/Authority [39] framework is multi document summarization system which, firstly detect the sub-topics in multi-documents by sentence clustering and extract the feature words (or phrase) of different sub-topics. Secondly, all feature words and the cue phrases are used as the vertex of Hub and all sentences are regarded as the vertex of Authority. If the sentence contains the words in Hub, there is an edge between the Hub word and the Authority sentence. The initial weight of each vertex considers both the content and the cues such as cue phrase and first sentence. Through the mutual reinforcement mechanism of the Hub-Authority algorithm, we can rank the importance of the sentences within the multi-documents. The assumption behind this cue-based Hub/Authority approach is that a good Hub word (or phrase) is the content that points to many good authorities sentences and a good authority sentence is a vertex that is pointed to by many good hub words. Thirdly, It has used the Markov Model to order the sub-topics that the final summarization should contain and output the text summarization according to the sentence

ranking score of all sentences within one sub-topic as user' requirement.

Generic relation extraction (GRE) [40] is a novel multi document text summarization approach, which aims to build systems for relation identification and characterization that can be transferred across domains and tasks without modification of model parameters.

K. Query based extractive text summarization

In query based text summarization [42] system, the sentences in a given document are scored based on the frequency counts of terms (words or phrases). The sentences containing the query phrases are given higher scores than the ones containing single query words. Then, the sentences with highest scores are incorporated into the output summary together with their structural context. Portions of text may be extracted from different sections or subsections. The resulting summary is the union of such extracts. The number of extracted sentences and the extent to which their context is displayed depends on the summary frame size which is fixed to the size of the screen that can be seen without scrolling. In the sentence extraction algorithm, whenever a sentence is selected for the inclusion in the summary, some of the headings in that context are also selected.
The query based sentence extraction algorithm is as follows:

Algorithm:
1: Rank all the sentences according to their score.
2: Add the main title of the document to the summary.
3: Add the first level-1 heading to the summary.
4: While (summary size limit not exceeded)
5: Add the next highest scored sentence.
6: Add the structural context of the sentence:
(if any and not already included in the summary)
7: Add the highest level heading above the
extracted text (call this heading h).
8: Add the heading before h in the same level.
9: Add the heading after h in the same level.
10: Repeat steps 7, 8 and 9 for the next highest level
headings.
11: End while

An another query-specific summarization [43] method views a document as a set of interconnected text fragments (passages) and focuses on keyword queries, since keyword search is the most popular information discovery method on documents, because of its power and ease of use. Firstly, at the preprocessing stage, it adds structure to every document, which can then be viewed as a labeled, weighted graph, called the document graph. Then, at query time, given a set of keywords, it performs keyword proximity search on the document graphs to discover how the keywords are associated in the document graphs. For each document its summary is the minimum spanning tree on the corresponding document graph that contains all the keywords.

In query-specific opinion summarization system [44] (QOS), When input an opinion question, the system returns a summary with relevance to the opinion and target described by the question. The system has several modules to be able to do this: a question analysis and query reformulation module, a latent semantic indexing based sentence scoring module, a sentence polarity detection module, and a redundancy removal module.

Bayesian summarization [45] (BAYESUM) is a model for sentence extraction in query-focused summarization. BAYESUM leverages the common case in which multiple documents are relevant to a single query. Using these documents as reinforcement for query terms, BAYESUM is not afflicted by the paucity of information in short queries. For a collection of D documents and Q queries, assume a D × Q binary matrix r, where $r_{dq} = 1$ if an only if document d is relevant to query q. In multi document summarization, r_{dq} will be 1 exactly when d is in the document set corresponding to query q.

L. Multilingual Extractive Text summarization

Multilingual text summarization is to summarize the source text in different language to the target language final summary. SimFinderML [24] identifies similar pieces of text by computing similarity over multiple features. There are two types of features, composite features, and unary features. All features are computed over primitives, syntactic, linguistic, or knowledge-based information units extracted from the sentences. Both composite and unary features are constructed over the primitives. The primitives used and features computed can be set at run-time, allowing for easy experimentation with different settings, and making it easy to add new features and primitives. Support for new languages is added to the system by developing modules conforming to interfaces for text pre-processing and primitive extraction for the language, and using existing dictionary-based translation methods, or adding other language-specific translation methods.

MINDS [25] integrates multi-lingual summarization and multi document summarization capabilities using a multiengine, core summarization system and provides fast, interactive document access through hypertext summaries. Core summarization problem of MINDS is taking a single text and producing a shorter text in the same language that contains all the main points in the input text. It is using a robust, graded approach for building the core engine by incorporating statistical, syntactic and documents structure analyses among other techniques. This approach is less expensive and more robust than a summarization technique based entirely on a single method. The core engine is being designed in such a way that as additional resources, such as lexical and other knowledge bases or text processing and MT engines, become available from other ongoing research efforts they can be incorporated into the overall multi-engine MINDS system. Ideally the core engine itself will remain language independent. A prototype core engine has been built for English, Spanish, Russian, and Japanese documents.

MEAD [26] is the multi-lingual summarization and evaluation method. MEAD's architecture consists of four stages. First, documents in a cluster are converted to MEAD's internal (XML-based) format. Second, given a configuration file or command-line options, a number of features are extracted for each sentence of the cluster. Third, these features are combined into a composite score for each sentence. Fourth, these scores can be further refined after considering possible cross-sentence dependencies (e.g., repeated sentences, chronological ordering, source preferences, etc.) In addition to a number of command-line utilities, MEAD provides a Perl API which lets external programs access its internal libraries.

V. Conclusions

This survey paper is concentrating on extractive summarization methods. An extractive summary is selection of important sentences from the original text. The importance of sentences is decided based on statistical and linguistic features of sentences.

Many variations of the extractive approach [41] have been tried in the last ten years. However, it is hard to say how much greater interpretive sophistication, at sentence or text level, contributes to performance. Without the use of NLP, the generated summary may suffer from lack of cohesion and semantics. If texts containing multiple topics, the generated summary might not be balanced. Deciding proper weights of individual features is very important as quality of final summary is depending on it. We should devote more time in deciding feature weights.

The biggest challenge for text summarization is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user. The text summarization software should produce the effective summary in less time and with least redundancy. Summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation and extrinsic methods measure the same through a task based performance measure [35] such the information retrieval-oriented task.

VI. References

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[2] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[3] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach", Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.

[4] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.

[5] Fang Chen, Kesong Han and Guilin Chen, "An Approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489-493, 2002.

[6] Mohamed Abdel Fattah and Fuji Ren, "Automatic Text Summarization", Proceedings of World Academy of Science, Engineering and Technology, Vol 27,ISSN 1307-6884, 192-195, Feb 2008.

[7] H. P. Luhn**, "**The Automatic Creation of Literature Abstracts", Presented at IRE National Convention, New York, 159-165, 1958.

[8] H. P. Edmundson.," New methods in automatic extracting", Journal of the ACM, 16(2):264-285, April 1969.

[9] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer", In Proceedings of the 18th ACM-SIGIR Conference, pages 68-73, 1995

[10] Ronald Brandow, Karl Mitze, and Lisa F. Rau. "Automatic condensation of electronic publications by sentence selection. Information Processing and Management", 31(5):675-685,1995.

[11] E. Mittendorf and P. Schauble, " Document and passage retrieval based on hidden markov models", In Proceedings of the 17th ACM-SIGIR Conference, pages 318-327,1994.

[12] A. Bookstein, S. T. Klein, and T. Raita, "Detecting content-bearing words by serial clustering", In Proceedings of the 18th ACM-SIGIR Conference, pages 319-327, 1995.

[13] Madhavi K. Ganapathiraju, "Overview of summarization methods", 11-742: Self-paced lab in Information Retrieval, November 26, 2002.

[14] Klaus Zechner, "A Literature Survey on Information Extraction and Text Summarization", Computational Linguistics Program, Carnegie Mellon University, April 14, 1997.

[15] Chin-Yew Lin and Eduard Hovy," From Single to Multi-document Summarization: A Prototype System and its Evaluation", Proceedings of the ACL conference, pp. 457–464. Philadelphia, PA. 2002.

[16] Rene Arnulfo Garcia-Herandez and Yulia Ledeneva, "Word Sequence Models for Single Text Summarization", IEEE,44-48, 2009.

[17] Yongzheng, Nur and Evangelos, "Narrative Text Classification for Automatic Key Phrase Extraction in Web Document Corpora", WIDM'5, 51-57, Bremen Germany,2005.

[18] Canasai Kruengkari and Chuleerat Jaruskulchai, "Generic Text Summarization Using Local and Global Properties of Sentences", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03) , 2003.

[19] Meng Wang, Xiaorong Wang and Chao Xu, "An Approach to Concept Oriented Text Summarization", in Proceedings of ISCIT'05, IEEE international conference, China,1290-1293, 2005.

[20] Azadeh Zamanifar, Behrouz minaei-Bidgoli and Mohsen Sharifi," A New Hybrid Farsi Text Summarization Technique Based on Term Co-Occurrence and Conceptual Property of Text ", In Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, 635-639, Iran, 2008.

[21] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", in Proceedings of second international Conference on intelligent systems, IEEE, 40-44, Texas, USA, June 2004.
[22] Ladda Suanmali, Mohammed Salem, Binwahlan and Naomie Salim, "Sentence Features Fusion for Text summarization using Fuzzy Logic, IEEE, 142-145, 2009
[23] David B. Bracewell, Fuji REN and Shingo Kuriowa, "Multilingual Single Document Keyword Extraction for Information Retrieval", Proceedings of NLP-KE'05, IEEE, Tokushima, 2005.
[24] David Kirk Evans, "Identifying Similarity in Text: Multi Lingual Analysis for Summarization", PhD thesis, Graduate School of Arts and Sciences, Columbia University, 2005.
[25] Cowie, J., Mahesh, K., Nirenburg, S., and Zajaz, R., "MINDS-Multilingual INteractive document summarization", In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 131–132). Menlo Park, CA: AAAI, 1998.
[26] Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda C, elebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhang Zhu., "MEAD - a platform for multi document multilingual text summarization", In Proceedings of LREC 2004, Lisbon, Portugal, May 2004.
[27] Khosrow Kaikhah "Text Summarization using Neural Networks", Department of Faculty Publications-Computer Science, Texas State University, eCommons,2004.
[28] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.
[29] Rasim M. Alguliev and Ramiz M. Aliguliyev, "Effective Summarization Method of Text Documents", in Proceedings of IEEE/WIC/ACM international conference on Web Intelligence (WI'05), 1-8, 2005.
[30] Berry Michael W., "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43, 2004.
[31] Vishal Gupta, G.Sl Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009.
[32] G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.
[33] Udo Hahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator", Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics , ACM, Morristown, NJ, USA , 2001.
[34] Ani Nenkova and ,Rebecca Passonneau, "Evaluating content selction in summarization: The Pyramid method", in HLT-NAACL, 145-152, 2004.
[35] Kathleen Mackeown, Ani Nenkova, David Elson, Rebecca Passonneau, and Julia Hirschberg "A task based evaluation of multidocument system",in SIGIR'05, ACM, 2005.
[36] Chin-yew Lin, "A package for automatic evaluation of summaries",in Proc. ACL workshop on text summarization branches out,2004.
[37] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto, "Automated Summarization Evaluation with Basic Elements", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006.
[38] Samuel W K. Chan, Tom B. Y. Lai, W.J. Gao and Benjamin K. T'sou, "Mining discourse structures for chinese textual summarization", NAACL-ANLP workshop on Automatic Summarization, ACM, Seattle, Washington, 11-20, 2000.
[39] Junlin Zhanq, Le Sun and Quan Zhou, "A Cue-based Hub-Authority Approach for Multi-Document Text Summarization", in Proceeding of NLP-KE'05, IEEE,642-645, 2005
[40] Ben Hachey, "Multi-document summarization using generic relation extraction", Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing: Volume 1, 420-429, 2009.
[41] K. S. Jones, "Automatic summarizing: the state of the art," Information Processing and Management, Elsevier, Vol. 43, No. 6, pp. 1449–1481, 2007.
[42] F. Canan Pembe and Tunga Güngör, "Automated Query-biased and Structure-preserving Text Summarization on Web Documents", Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, İstanbul, June 2007.
[43] Ramakrishna Varadarajan and Vagelis Hristidis, "Structure-Based Query-Specific Document Summarization", in proceedings of CIKM'05, ACM, Bremen, Germany, 2005.
[44] Feng Jin, Minlie Huang and Xiaoyan Zhu, " A Query-specific Opinion Summarization System", in proceedings ofICCI '09, 8th IEEE international conference on cognitive informatics, Kowloon, Hong Kong, 428-433, 2009.
[45] Hal Daum´e III and Daniel Marcu, "Bayesian Query-Focused Summarization", Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, 305–312, Sydney, July 2006.
[46] Jimmy Lin., "Summarization.", Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009.
[47] Jackie CK Cheung, "Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection", B. Sc. (Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia, 2008.

VISHAL GUPTA

Vishal Gupta is Assistant Professor in Computer Science & Engineering at University Institute of Engineering & Technology, Panjab university Chandigarh. He has done MTech. in computer science & engineering from Punjabi University Patiala in 2005. He is among University toppers. He has done BTech. in Computer Science & Engineering from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Science &

Engineering from University College of Engineering, Punjabi University Patiala, under the supervision of Dr. Gurpreet Singh Lehal. He is state merit holder in 10^{th} and 12^{th} classes of Punjab School education board. He is devoting his research work in field of Natural Language Processing. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference.

DR. GURPREET SINGH LEHAL

Professor Gurpreet Singh Lehal received undergraduate degree in Mathematics in 1988 from Panjab University, Chandigarh, India, and Post Graduate degree in Computer Science in 1995 from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, in 2002. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and research. His current areas of research are- Natural Language Processing and Optical Character recognition. He has published more than 25 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions- Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration Solution for Networking.

Call for Papers and Special Issues

Aims and Scope

Journal of Emerging Technologies in Web Intelligence (JETWI, ISSN 1798-0461) is a peer reviewed and indexed international journal, aims at gathering the latest advances of various topics in web intelligence and reporting how organizations can gain competitive advantages by applying the different emergent techniques in the real-world scenarios. Papers and studies which couple the intelligence techniques and theories with specific web technology problems are mainly targeted. Survey and tutorial articles that emphasize the research and application of web intelligence in a particular domain are also welcomed. These areas include, but are not limited to, the following:

- Web 3.0
- Enterprise Mashup
- Ambient Intelligence (AmI)
- Situational Applications
- Emerging Web-based Systems
- Ambient Awareness
- Ambient and Ubiquitous Learning
- Ambient Assisted Living
- Telepresence
- Lifelong Integrated Learning
- Smart Environments
- Web 2.0 and Social intelligence
- Context Aware Ubiquitous Computing
- Intelligent Brokers and Mediators
- Web Mining and Farming
- Wisdom Web
- Web Security
- Web Information Filtering and Access Control Models
- Web Services and Semantic Web
- Human-Web Interaction
- Web Technologies and Protocols
- Web Agents and Agent-based Systems
- Agent Self-organization, Learning, and Adaptation
- Agent-based Knowledge Discovery
- Agent-mediated Markets
- Knowledge Grid and Grid intelligence
- Knowledge Management, Networks, and Communities
- Agent Infrastructure and Architecture
- Agent-mediated Markets
- Cooperative Problem Solving
- Distributed Intelligence and Emergent Behavior
- Information Ecology
- Mediators and Middlewares
- Granular Computing for the Web
- Ontology Engineering
- Personalization Techniques
- Semantic Web
- Web based Support Systems
- Web based Information Retrieval Support Systems
- Web Services, Services Discovery & Composition
- Ubiquitous Imaging and Multimedia
- Wearable, Wireless and Mobile e-interfacing
- E-Applications
- Cloud Computing
- Web-Oriented Architectrues

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers” to be included on the Journal’s Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. “Special Issue: Selected Best Papers of XYZ Conference”.
- Sending us a formal “Letter of Intent” for the Special Issue.
- Creating a “Call for Papers” for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jetwi/.

(Contents Continued from Back Cover)