THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

November 2007 Vol. 8 No. 1 (ISSN 1727-5997)

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

The Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society deals with tools and systems using biologically and linguistically motivated computational paradigms such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality. If you are a member of the IEEE Computer Society, you may join the TCII without cost. Just fill out the form at http://computer.org/tcsignup/.

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published twice a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee

2) Feature Articles

3) R&D Profiles (R&D organizations, Interview profile on individuals, and projects etc.)

4) Book Reviews

5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and **DBLP** Computer Science Bibliography (www.informatik.uni-trier.de/»ley/db/index.html).*

Message from the TCII Chair

The Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society deals with tools and systems using cognitive and intelligent paradigms such as knowledge engineering, artificial neural networks, fuzzy logic, evolutionary computing, and rough sets, with research and applications in data mining, Web intelligence, brain informatics, intelligent agent technology, parallel and distributed information processing, and virtual reality.

The TCII has been playing a vital role in the IEEE Computer Society's activities, based on its success story for the past 4 years under great leadership of the founding chairman, Dr. Xindong Wu. As the new chair, I will promote more activities sponsored by the TCII and advance the TCII's role both within the IEEE Computer Society and in collaboration with related Special Interest Groups in ACM and other organizations.

I have formed a new Executive Committee to manage TCII activities with the following members:

- Jeffrey M. Bradshaw (Industry Connections), Institute for Human and Machine Cognition, USA
- Nick J. Cercone (Student Affairs), York University, Canada.
- Boi Faltings (Curriculum Issues), Swiss Federal Institute of Technology, Switzerland.
- Vipin Kumar (Bulletin Editor), University of Minnesota, USA.
- Vice Chair: Jiming Liu (Conferences and Membership), Hong Kong Baptist University, Hong Kong.
- Benjamin W. Wah (Awards), University of Illinois, Urbana-Champaign, USA
- Past Chair: Xindong Wu, University of Vermont, USA.
- Chengqi Zhang (Cooperation with Sister Societies/TCs), University of Technology, Sydney, Australia.

The TCII currently co-sponsors the IEEE International Conference on Data Mining (ICDM), the IEEE/WIC/ACM International Conference on Web Intelligence (WI), the IEEE/WIC/ACM International Conference on Intelligence Agent Technology (IAT), the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), and the IEEE International Conference on BioInformation and BioMedicine (BIBM). I am confident that

we can sponsor these conferences to promote TCII activities in these areas.

In addition to conference activities, I will also work hard to advocate the following activities:

a) Enhancing the TCII ability by increasing membership, promoting truly international technical activities, and improving the TCII website and the TCII Bulletin publication.

b) Close cooperation with other TC's in the IEEE Computer Society, the IEEE sister societies, Special Interest Groups in ACM, and other related organizations, such as Web Intelligence Consortium (WIC), International Rough Set Society, and AAAI.

c) Student participations in the TCII and TCII-sponsored conferences, in collaboration with the IEEE Computer Society's Chapters Activities Board.

d) Curriculum and text books development in intelligent informatics, in collaboration with the IEEE Computer Society's Educational Activities Board.

e) Joint activities with the IEEE Computer Society's Publications Board, possibly in TKDE and TPAMI in the form of special issue promotion and editorial appointments.

f) The bridge between the research community and the industry practitioners.

I hope you will enjoy reading this Bulletin. If you need any more information about TCII, please visit the TCII website at http://www.maebashi-it.org/cyberchair/tcii/index.html. If you are a member of the IEEE Computer Society, you may join the TCII without cost. Just fill out the registration form available on the TCII website mentioned above.

Ning Zhong
(Maebashi Institute of Technology, Japan)
Chair, TCII –
IEEE Computer Society Technical Committee on Intelligent Informatics

Message from the Editors

The IEEE Intelligent Informatics Bulletin is the official publication of the IEEE Computer Society Technical Committee on Intelligent Informatics. The bulletin is intended to provide the community of researchers and practitioners in this vibrant field timely information on latest research, as well as educational and professional activities in areas related to Intelligent Informatics. The bulletin is also intended to serve as a forum for quick dissemination of ideas and experiences for the research community.

The issue contains three exciting feature articles: one on Web search personalization based on an ontology learning approach, one on conversational informatics for human-centered Web intelligence, and one more on ontology discovery for business knowledge management. The R&D profile section highlights the research being done in the intelligent systems group at Florida Institute of Technology. The book review section provides review of the timely book by Zhang, Zhang, and Wu on knowledge discovery in multiple databases. The announcement section contains call for papers for several international conferences of interest to the field of intelligent informatics.

The Bulletin is the result of hard work by the members of the Editorial Board. We would like to express our gratitude to their efforts in working closely with contributing authors to get this issue to you in a timely fashion. We hope that you will enjoy reading the Bulletin and find it informative. We also invite you to send us your comments and suggestions for improving the bulletin as well as your contributions for the various sections of the bulletin for the next issue.

Vipin Kumar
(University of Minnesota)
Editor-in-Chief, IEEE Intelligent Informatics Bulletin

William Cheung
(Hong Kong Baptist University, Hong Kong)
Managing Editor IEEE Intelligent Informatics Bulletin

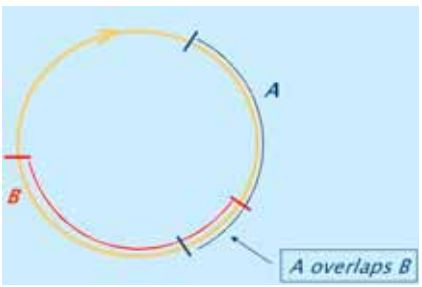
Intelligent Systems at Florida Tech

cci.cs.fit.edu

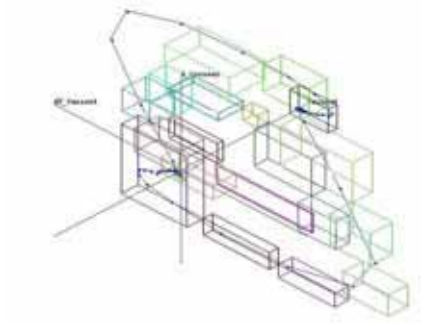
At Florida Institute of Technology, intelligent systems is one of the main areas of research in the Computer Sciences department. The focus in general is on *(i)* how to make computers more intelligent as well as *(ii)* how intelligence can change the ways we compute. Speci cally , one investigates algorithms that can help computers reason (constraint reasoning, spatio-temporal reasoning), learn (machine learning), and see (computer vision). Moreover, we examine how distributed intelligent agents can interact (distributed constraint reasoning and coordination). Our research also includes approaches on looking at how simplistic animal behavior can provide a novel way to solve problems (swarm intelligence).

I. Research

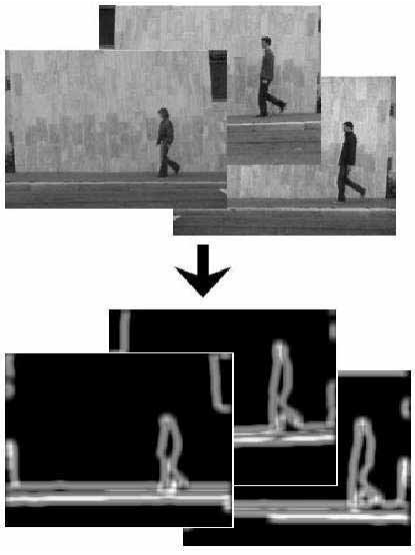
The ability to reason is fundamental to human intelligence for making decisions. The constraint reasoning group focuses on spatio-temporal constraint reasoning. The group has developed new calculi for qualitative spatial reasoning that has applications in geographical information systems. The group has also developed techniques in detecting the culprit constraints in an unsatis able temporal reasoning problem. Recently, we launched a project on understanding creativity from a constraint reasoning viewpoint. The long term vision of this initiative is to develop an intelligent workbench for helping physicists in their creative activities.



Another fundamental aspect of human intelligence is learning—the ability to generate new knowledge and adapt to the changing environment. The machine learning (data mining) group focuses on investigating techniques for anomaly detection and web personalization. Unlike the typical machine learning problem of building a classi er from training examples from two or more classes, the anomaly detection problem necessitates constructing a classi er from training examples from only one class—the "normal" class. The learned classi er is an anomaly detector that identi es and scores anomalies. For intrusion detection, anomaly detection has the potential of detecting novel attacks, which cannot be detected by identifying signatures of existing attacks. For device monitoring with time series data, we extract features, plot them, and generalize them into a sequence of "boxes," which form a model for anomaly detection. For web personalization, we learn a user pro le from a user's bookmarks and use the pro le to re-rank results returned by a search engine so that the top results are closer to the user's interests.

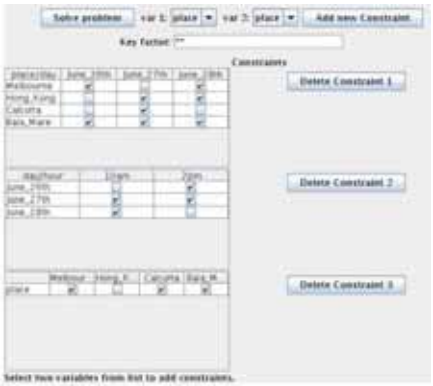


As eyes are windows to the world, the ability to interpret what we see is crucial in providing information for reasoning and learning. The computer vision group concentrates on object recognition, texture analysis, and human motion recognition and understanding. Our texture analysis work has focused on building algorithms for extracting 3D surface shape descriptions from textured surfaces as well as on the analysis and recognition of deforming textures. More recently, research in the group has been concentrated in developing probabilistic models for the recognition of objects and human motion. Object recognition and human motion analysis are key to many important real-world applications such as video surveillance and database content-based retrieval. We also perform research on human-agent interfaces in the area of speech recognition algorithms.

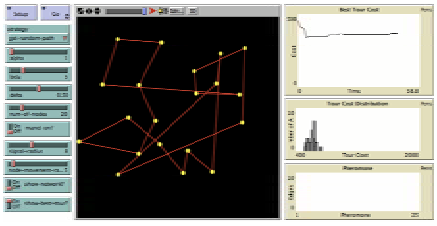


Humans do not live in isolation. Cooperation among individuals allows us to achieve goals that none of us can achieve individually. The ability to coordinate and cooperate is another key aspect of human intelligence. Hence, the agent systems group conducts research in distributed problem solving. The group created and maintains the Secure Multi-party Computation (SMC) language. SMC can be used for building problem solving agents that can run as applications and applets placed on

different machines. The group develops various algorithms for distributed problem solving with privacy requirements, and has a particular focus on distributed constraint reasoning.



Though humans are arguably the most intelligent organism, we have much to learn from others. Hence, the bio-inspired computing group strives to solve real-world problems using techniques based on our understanding of the biological world. Particularly, we study swarm intelligence—how group behavior of simple organisms can produce complex and intelligent behavior. Problem areas that we apply swarm intelligence include wireless sensor networks, distributed data organization, crime activity modeling, social networks, and software engineering.



This research has led to quite a number of publications over the years. Recently, within the past year, the group has published over a dozen articles at peer-reviewed conferences and journals. Also, two research grants were awarded within the past year.

II. Educational and Conference Activities

The intelligent systems group has ve faculty members. Dr. Philip Chan joined Florida Tech in 1995 coming from Columbia University. During 2000-2003, he was joined by Dr. Ronaldo Menezes, Dr. Debasis Mitra, Dr. Marius Silaghi, and Dr. Eraldo Ribeiro. In 2004 the ve faculty members founded the Center for Computation and Intelligence (CCI).

To help students explore various areas in intelligent systems, the group offers courses in arti cial intelligence, bio-inspired computing. computer vision, constraint reasoning, machine learning, and multi-agent systems. Over fteen graduate and undergraduate students actively participate in our research projects. Within the past ten years, at least ten master's and two PhD students have graduated.

In the past few years, we have helped organize and hosted four research conferences held in Melbourne, Florida:

- ACM Symposium in Applied Computing (SAC), March, 2003;
- IEEE International Conference on Data Mining (ICDM), November, 2003;
- ACM Southeast Conference (ACMSE), March, 2006; and
- International FLAIRS Conference (FLAIRS), May, 2006.

Moreover, we have helped organize workshops and conducted tutorials at research conferences, which include:

- Workshop on Integrating Multiple Learned Models, AAAI-96;
- Workshop on Distributed Data Mining, KDD-98;
- Workshop on Data Mining for Computer Security, ICDM-03, CCS-04;
- Workshop on Data Mining Methods for Anomaly Detection, KDD-05;
- Tutorial on Data Mining for Computer Security, KDD-03, SDM-04; and
- Tutorial on Distributed Constraint Reasoning, IJCAI-03, IJCAI-05.

III. Concluding Remarks

The research of CCI members has partially been funded by the Brazilian Funding Agency, Defense Advanced Research Project Agency (DARPA), Department of Homeland Security (DHS), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), and Of ce of Naval Research (ONR). For more information about the research of the group, please use the following contact address.

Contact Information

Department of Computer Sciences
Florida Institute of Technology
150 West University Boulevard
Melbourne, FL 32901, USA

Website: cci.cs.fit.edu

Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search

Ahu Sieg, Bamshad Mobasher, Robin Burke
Center for Web Intelligence
School of Computer Science, Telecommunication and Information Systems
DePaul University, Chicago, Illinois, USA
asieg, mobasher, rburke@cs.depaul.edu

***Abstract*—Every user has a distinct background and a specific goal when searching for information on the Web. The goal of Web search personalization is to tailor search results to a particular user based on that user's interests and preferences. Effective personalization of information access involves two important challenges: accurately identifying the user context, and organizing the information in such a way that matches the particular context. We present an approach to personalized search that involves modeling the user context as ontological profiles by assigning implicitly derived interest scores to existing concepts in a domain ontology. A spreading activation algorithm is used to maintain and incrementally update the interest scores based on the user's ongoing behavior. Our experiments show that re-ranking the search results based on the interest scores and the semantic evidence captured in an ontological user profile enables an adaptive system to present the most relevant results to the user.**

***Index Terms*—Search Personalization, Ontological User Profiles, User Context, Web Mining, Information Retrieval**

I. Introduction

Web personalization alleviates the burden of information overload by tailoring the information presented based on an individual user's needs. Every user has a specific goal when searching for information through entering keyword queries into a search engine. Keyword queries are inherently ambiguous but often formulated while the user is engaged in some larger task [1]. For example, an historian looking for early Renaissance Christian paintings may enter the query *Madonna and child* while browsing Web pages about art history, while a music fan may issue the same query to look for news about the famous pop star.

In recent years, personalized search has attracted interest in the research community as a means to decrease search ambiguity and return results that are more likely to be interesting to a particular user and thus providing more effective and efficient information access [2], [3], [4]. One of the key factors for accurate personalized information access is user context.

Researchers have long been interested in the role of context in a variety of fields including artificial intelligence, context-aware applications, and information retrieval. While there are many factors that may contribute to the delineation of the user context, here we consider three essential elements that collectively play a critical role in personalized Web information access. These three independent but related elements are the user's short-term information need, such as a query or localized context of current activity, semantic knowledge about the domain being investigated, and the user's profile that captures long-term interests. Each of these elements are considered to be critical sources of contextual evidence, a piece of knowledge that supports the disambiguation of the user's context for information access.

In this paper, we present a novel approach for building ontological user profiles by assigning interest scores to existing concepts in a domain ontology. These profiles are maintained and updated as annotated specializations of a pre-existing reference domain ontology. We propose a spreading activation algorithm for maintaining the interest scores in the user profile based on the user's ongoing behavior. Our experimental results show that re-ranking the search results based on the interest scores and the semantic evidence in an ontological user profile successfully provides the user with a personalized view of the search results by bringing results closer to the top when they are most relevant to the user.

We begin by discussing the related work and the motivational background behind this work. We then present our approach for building the ontological user profiles. Finally, we discuss the application of our contextual user model to *Search Personalization* and present the results of an extensive experimental evaluation.

II. Background and Motivation

A. Related Work

Web search engines are essential "one size fits all" applications [5]. In order to meet the demands of extremely high query volume, search engines tend to avoid any kind of representation of user preferences, search context, or the task context [6]. Allan et al. [5] define the problem of *contextual retrieval* as follows: "Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs."

Effective personalization of information access involves two important challenges: accurately identifying the user context, and organizing the information in such a way that matches the particular context. Since the acquisition of user interests and preferences is an essential element in identifying the user context, most personalized search systems employ a *user modeling* component.

Recent studies show that users often settle for the results returned by imprecise queries, picking through them for relevant information, rather than expending the cognitive effort required to formulate more accurate queries. Since the users are reluctant to specify their underlying intent and search goals, personalization must pursue techniques that leverage implicit information about the user's interests [7], [8].

Google Personalized Search[1] builds a user profile by means of implicit feedback where the system adapts the results according to the search history of the user. Many systems employ search personalization on the client-side by re-ranking documents that are suggested by an external search engine [9], [10] such as Google. Since the analysis of the pages in the result list is a time consuming process, these systems often take into account only the top ranked results. Also, only the snippets associated with each page in the search results is considered as opposed to the entire page content.

Many personalization approaches are based on some type of a user profile which is a data instance of a user model that is captured based on the user's interaction. User profiles may include demographic information as well as representing the interests and preferences of a specific user. User profiles that are maintained over time can be categorized into short-term and long-term profiles. Short-term profiles can be utilized to keep track of the user's more recent, faster-changing interests. Long-term profiles represent user interests that are relatively stable over time.

Personal browsing agents such as WebMate [11] and WebWatcher [12] perform tasks such as highlighting hyperlinks and refining search keywords to satisfy the user's short-term interests. These approaches focus on collecting information about the users as they browse or perform other activities.

InfoWeb [13] builds semantic network based profiles that represents long-term user interests. The user model is utilized for filtering online digital library documents. Gasperetti and Micarelli [14] propose a user model which tries to represent human memory. Each profile essentially consists of two keyword vectors, one vector represents the short-term interests whereas the other represents long-term interests. Our work differs from these approaches since we utilize a concept based model as opposed to representing the profiles as keyword vectors.

One increasingly popular method to mediate information access is through the use of ontologies [15]. Researchers have attempted to utilize ontologies for improving navigation effectiveness as well as personalized Web search and browsing, specifically when combined with the notion of automatically generating semantically enriched ontology-based user profiles [16]. Our research [17] follows recent ontology-based personalized search approaches [18], [19] in utilizing the *Open Directory Project (ODP)*[2] taxonomy as the Web topic ontology. The ODP is the largest and most comprehensive Web directory, which is maintained by a global community of volunteer editors. The ODP taxonomy is used as the basis for various research projects in the area of Web personalization [20], [21].

Liu et al. [22] utilize the first three levels of the ODP for learning profiles as bags of words associated with each category. The user's query is mapped into a small set of categories as a means to disambiguate the words in the query. The Web search is then conducted based on the user's original query and the set of categories. As opposed to using a set of categories, Chirita et al. [23] utilize the documents stored locally on a desktop PC for personalized query expansion. The query terms are selected for Web search by adapting summarization and natural language processing techniques to extract keywords from locally stored desktop documents.

Hyperlink-based approaches have also been explored as a means to personalize Web search. In Persona [24], the well-known *Hyperlink Induced Topic Selection (HITS)* algorithm [25] is enhanced with an interactive query scheme utilizing the Web taxonomy provided by the ODP to resolve the meaning of a user query.

Considerable amount of Web personalization research has been aimed at enhancing the original PageRank algorithm introduced in Google. In *Personalized Page Rank* [26], a set of personalized hub pages with high PageRank is needed to drive the personalized rank values. In order to automate the hub selection in *Personalized Page Rank*, a set of user collected bookmarks is utilized in a ranking platform called *PROS* [27].

Instead of computing a single global PageRank value for every page, the *Topic-Sensitive PageRank* [28] approach tailors the PageRank values based on the 16 main topics listed in the Open Directory. Multiple *Topic-Sensitive PageRank* values are computed off-line. Using the similarity of the topics to the query, a linear combination of the topic-sensitive ranks are employed at run-time to determine more accurately which pages are truly the most important with respect to a particular query. This approach is effective only if the search engine can estimate the suitable topic for the query and the user. Thus, Qui and Cho [29] extend the topic-sensitive method to address the problem of automatic identification of user preferences and interests.

B. Terminology

The notion of *context* may refer to a diverse range of ideas depending on the nature of the work being performed. Previous work defines context by using a fixed set of attributes such as location, time or identities of nearby individuals or objects, as is commonly done in ubiquitous computing [30]. In this section, we define more precisely what we mean by *context* and other related terminology used in the paper.

Context: The representation of a user's intent for information seeking. We propose to model a user's information access context by seamlessly integrating knowledge from the immediate and past user activity as well as knowledge from a pre-existing ontology as an explicit representation of the domain of interest. In our framework [31], *context* is implicitly defined through the notion of ontological user profiles, which are updated over time to reflect changes in user interests. This

[1] http://www.google.com/psearch
[2] http://www.dmoz.org

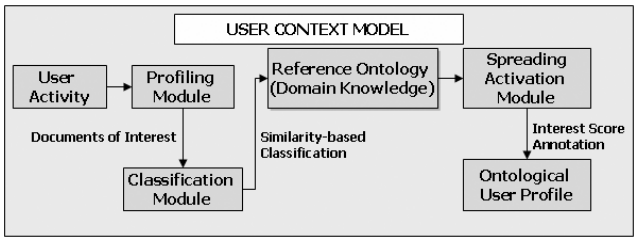


Fig. 1. Ontological User Profile as the Context Model

representation distinguishes our approach from previous work which depends on the *context* information to be explicitly defined.

Ontology: An ontology is an explicit specification of concepts and relationships that can exist between them. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge [32]. The set of relations such as subsumption *is-a* and meronymy *part-of* describe the semantics of the domain. Rather than creating our own ontology, we choose to base our reference ontology on an existing hierarchical taxonomy; a tree-like structure that organizes Web content into pre-defined topics.

Query: A search query consisting of one or more keywords and is the representation of a user's short-term or immediate information need.

III. Ontologies for Web Personalization

Our goal is to utilize the user context to personalize search results for a given query. The personalization is achieved by re-ranking the results returned from a search engine. Our unified context model for a user is represented as an instance of a pre-existing reference domain ontology in which concepts are annotated by *interest scores* derived and updated implicitly based on the user's information access behavior. We call this representation an *ontological user profile*.

Our assumption is that semantic knowledge is an essential part of the user context. Thus, we use a domain ontology as the fundamental source of semantic knowledge in our framework. An ontological approach to user profiling has proven to be successful in addressing the *cold-start problem* in recommender systems where no initial information is available early on upon which to base recommendations [33]. When initially learning user interests, systems perform poorly until enough information has been collected for user profiling. Using ontologies as the basis of the profile allows the initial user behavior to be matched with existing concepts in the domain ontology and relationships between these concepts.

Trajkova and Gauch [16] calculate the similarity between the Web pages visited by a user and the concepts in a domain ontology. After annotating each concept with a weight based on an accumulated similarity score, a user profile is created consisting of all concepts with non-zero weights.

In our approach, the purpose of using an ontology is to identify topics that might be of interest to a specific Web user. Therefore, we define our ontology as a hierarchy of topics, where the topics are utilized for the classification and categorization of Web pages. The hierarchical relationship among the concepts is taken into consideration for building the ontological user profile as we update the annotations for existing concepts using spreading activation.

A. Ontological User Profiles

The Web search personalization aspect of our research is built on the previous work in ARCH [34]. In ARCH, the initial query is modified based on the user's interaction with a concept hierarchy which captures the domain knowledge. This domain knowledge is utilized to disambiguate the user context.

In the present framework, the *user context* is represented using an *ontological user profile*, which is an annotated instance of a reference ontology. Figure 1 depicts a high-level picture of our proposed context model based on an *ontological user profile*. When disambiguating the context, the domain knowledge inherent in an existing reference ontology is called upon as a source of key domain concepts.

Each ontological user profile is initially an instance of the reference ontology. Each concept in the user profile is annotated with an *interest score* which has an initial value of one. As the user interacts with the system by selecting or viewing new documents, the ontological user profile is updated and the annotations for existing concepts are modified by spreading activation. Thus, the *user context* is maintained and updated incrementally based on user's ongoing behavior.

Accurate information about the user's interests must be collected and represented with minimal user intervention. This can be done by passively observing the user's browsing behavior over time and collecting Web pages in which the user has shown interest. Several factors, including the frequency of visits to a page, the amount of time spent on the page, and other user actions such as bookmarking a page can be used as bases for heuristics to automatically collect these documents [35].

B. Representation of Reference Ontology

Our current implementation uses the *Open Directory Project*, which is organized into a hierarchy of topics and Web pages that belong to these topics. We utilize the Web pages as training data for the representation of the concepts in the reference ontology. The textual information that can get extracted from Web pages explain the semantics of the concepts and is learned as we build a term vector representation for the concepts.

We create an aggregate representation of the reference ontology by computing a term vector $\vec{n}$ for each concept n in the concept hierarchy. Each concept vector represents, in aggregate form, all individual training documents indexed under that concept, as well as all of its subconcepts.

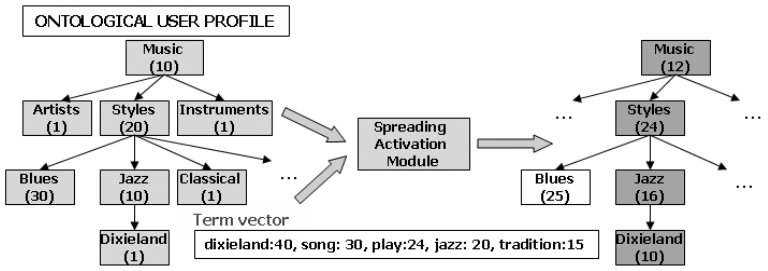


Fig. 2. Portion of an Ontological User Profile where Interest Scores are updated based on Spreading Activation

We begin by constructing a global dictionary of terms extracted from the training documents indexed under each concept. A stop list is used to remove high frequency, but semantically non-relevant terms from the content. Porter stemming [36] is utilized to reduce words to their stems. Each document d in the training data is represented as a term vector $\vec{d} = \langle w_1, w_2, ..., w_k \rangle$, where each term weight, w_i, is computed using term frequency and inverse document frequency [37]. Specifically, $w_i = tf_i * \log(N/n_i)$, where tf_i is the frequency of term i in document d, N is the total number of documents in the training set, and n_i is the number of documents that contain term i. We further normalize each document vector, so that $\vec{d}$ represents a term vector with unit length.

The aggregate representation of the concept hierarchy can be described more formally as follows. Let $S(n)$ be the set of subconcepts under concept n as non-leaf nodes. Also, let $\{d_1^n, d_2^n, ..., d_{k_n}^n\}$ be the individual documents indexed under concept n as leaf nodes. $Docs(n)$, which includes of all of the documents indexed under concept n along with all of the documents indexed under all of the subconcepts of n is defined as:

$$Docs(n) = [\bigcup_{n' \in S(n)} Docs(n')] \cup \{d_1^n, d_2^n, ..., d_{k_n}^n\}$$

The concept term vector $\vec{n}$ is then computed as:

$$\vec{n} = \left[\sum_{d \in Docs(n)} \vec{d} \right] / |Docs(n)|$$

Thus, $\vec{n}$ represents the centroid of the documents indexed under concept n along with the subconcepts of n. The resulting term vector is normalized into a unit term vector.

C. Context Model

Figure 2 depicts a portion an ontological user profile corresponding to the node *Music*. The interest scores for the concepts are updated with spreading activation using an input term vector.

Each node in the ontological user profile is a pair, $\langle C_j, IS(C_j) \rangle$, where C_j is a concept in the reference ontology and $IS(C_j)$ is the interest score annotation for that concept. The input term vector represents the active interaction of the user, such as a query or localized context of current activity.

Based on the user's information access behavior, let's assume the user has shown interest in *Dixieland Jazz*. Since the input term vector contains terms that appear in the term vector for the *Dixieland* concept, as a result of spreading activation, the interest scores for the *Dixieland*, *Jazz*, *Styles*, and *Music* concepts get incremented whereas the interest score for *Blues* gets decreased. The *Spreading Activation* algorithm and the process of updating the interest scores are discussed in detail in the next section.

D. Incrementally Learning Profiles by Spreading Activation

We use *Spreading Activation* to incrementally update the *interest score* of the concepts in the user profiles. Therefore, the ontological user profile is treated as the semantic network and the interest scores are updated based on activation values.

Traditionally, the spreading activation methods used in information retrieval are based on the existence of maps specifying the existence of particular relations between terms or concepts [38]. Alani et al. [39] use spreading activation to search ontologies in Ontocopi, which attempts to identify communities of practice in a particular domain. Spreading activation has also been utilized to find related concepts in an ontology given an initial set of concepts and corresponding initial activation values [40].

In our approach, we use a very specific configuration of spreading activation, depicted in Algorithm 1, for the sole purpose of maintaining *interest scores* within a user profile. We assume a model of user behavior can be learned through the passive observation of user's information access activity and Web pages in which the user has shown interest can automatically be collected for user profiling.

The algorithm has an initial set of concepts from the ontological user profile. These concepts are assigned an initial activation value. The main idea is to activate other concepts following a set of weighted relations during propagation and at the end obtain a set of concepts and their respective activations.

As any given concept propagates its activation to its neighbors, the weight of the relation between the origin concept and the destination concept plays an important role in the amount of activation that is passed through the network. Thus, a one-time computation of the weights for the relations in the network is needed. Since the nodes are organized into a concept hierarchy derived from the domain ontology, we compute the weights for the relations between each concept and all of its subconcepts using a measure of containment. The

containment weight produces a range of values between zero and one such that a value of zero indicates no overlap between the two nodes whereas a value of one indicates complete overlap.

The weight of the relation w_{is} for concept i and one of its subconcepts s is computed as $w_{is} = \frac{\vec{n}_i . \vec{n}_s}{\vec{n}_i . \vec{n}_i}$, where $\vec{n}_i$ is the term vector for concept i and $\vec{n}_s$ is the term vector for subconcept s. Once the weights are computed, we process the weights again to ensure the total sum of the weights of the relations between a concept and all of its subconcepts equals to 1.

Input: Ontological user profile with interest scores and a set of documents
Output: Ontological user profile concepts with updated activation values

CON = $\{C_1, ..., C_n\}$, *concepts with interest scores*
$IS(C_j)$, *interest score*
$IS(C_j) = 1$, *no interest information available*
$I = \{d_1, ..., d_n\}$, *user is interested in these documents*

foreach $d_i \in I$ **do**
  Initialize $priorityQueue$;
  foreach $C_j \in CON$ **do**
    $C_j.Activation = 0$; // Reset activation value
  end
  foreach $C_j \in CON$ **do**
    Calculate $sim(d_i, C_j)$;
    if $sim(d_i, C_j) > 0$ **then**
      $C_j.Activation = IS(C_j) * sim(d_i, C_j)$;
      $priorityQueue.Add(C_j)$;
    else
      $C_j.Activation = 0$;
    end
  end
  while $priorityQueue.Count > 0$ **do**
    Sort $priorityQueue$; // activation values(descending)
    $C_s = priorityQueue[0]$; // first item(spreading concept)
    $priorityQueue.Dequeue(C_s)$; // remove item
    if *passRestrictions*(C_s) **then**
      $linkedConcepts = GetLinkedConcepts(C_s)$;
      foreach C_l *in linkedConcepts* **do**
        $C_l.Activation += C_s.Activation * C_l.Weight$;
        $priorityQueue.Add(C_l)$;
      end
    end
  end
end

Algorithm 1: Spreading Activation Algorithm

The algorithm considers in turn each of the documents assumed to represent the current context. For each iteration of the algorithm, the initial activation value for each concept in the user profile is reset to zero. We compute a term vector for each document d_i and compare the term vector for d_i with the term vectors for each concept C_j in the user profile using a cosine similarity measure. Those concepts with a similarity score, $sim(d_i, C_j)$, greater than zero are added in a priority queue, which is in a non-increasing order with respect to the concepts' activation values.

The activation value for concept C_j is assigned to $IS(C_j) * sim(d_i, C_j)$, where $IS(C_j)$ is the existing interest score for the specific concept. The concept with the highest activation value is then removed from the queue and processed. If the current concept passes through restrictions, it propagates its activation to its neighbors. The amount of activation that is propagated to each neighbor is proportional to the weight of the relation. The neighboring concepts which are activated and are not currently in the priority queue are added to queue, which is then reordered. The process repeats itself until there

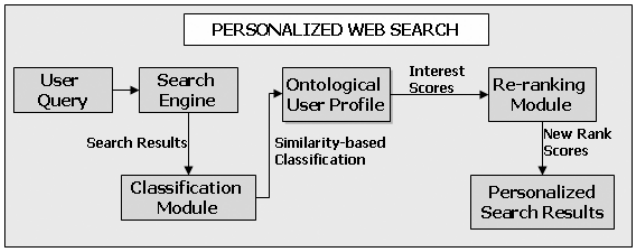


Fig. 3. Personalized Web Search based on Ontological User Profiles

are no further concepts to be processed in the priority queue.

The neighbors for the spreading concept are considered to be the linked concepts. For a given spreading concept, we can ensure the algorithm processes each edge only once by iterating over the linked concepts only one time. The order of the iteration over the linked concepts does not affect the results of activation. The linked concepts that are activated are added to the existing priority queue, which is then sorted with respect to activation values.

Input: Ontological user profile concepts with updated activation values
Output: Ontological user profile concepts with updated interest scores

CON = $\{C_1, ..., C_n\}$, *concepts with interest scores*
$IS(C_j)$, *interest score*
$C_j.Activation$, *activation value resulting from Spreading Activation*
k, *constant*

$n = 0$;
foreach $C_j \in CON$ **do**
  $IS(C_j) = IS(C_j) + C_j.Activation$;
  $n = n + (IS(C_j))^2$; // sum of squared interest scores
  $n = \sqrt{n}$; // square root of sum of squared interest scores
end
foreach $C_j \in CON$ **do**
  $IS(C_j) = (IS(C_j) * k)/n$; // normalize to constant length
end

Algorithm 2: Algorithm for the Normalization and Updating of Interest Scores in the Ontological User Profile

The interest score for each concept in the ontological user profile is then updated using Algorithm 2. First the resulting activation value is added to the existing interest score. The interest scores for all concepts are then treated as a vector, which is normalized to pre-defined constant length, k. The effect of normalization is to prevent the interest scores from continuously escalating throughout the network. As the user expresses interests in one set of concepts, the score for other concepts have to decrease. The concepts in the ontological user profile are updated with the normalized interest scores.

IV. Search Personalization

Our goal is to utilize the user context to personalize search results by re-ranking the results returned from a search engine for a given query. Figure 3 displays our approach for search personalization based on ontological user profiles. Assuming an ontological user profile with interest scores exists and we have a set of search results, Algorithm 3 is utilized to re-rank the search results based on the interest scores and the semantic evidence in the user profile.

A term vector $\vec{r}$ is computed for each document $r \in R$, where R is the set of search results for a given query. The

Input: Ontological user profile with interest scores and a set of search results
Output: Re-ranked search results

CON = $\{C_1, ..., C_n\}$, *concepts with interest scores*
$IS(C_j)$, *interest score*
$R = \{d_1, ..., d_n\}$, *search results from query q*

foreach $d_i \in R$ **do**
  Calculate $sim(d_i, q)$;
  maxSim = 0;
  foreach $C_j \in CON$ **do**
    Calculate $sim(d_i, C_j)$;
    if $sim(d_i, C_j) \geq maxSim$ **then**
      (Concept)c = C_j;
      maxSim = $sim(d_i, C_j)$;
    end
  end
  Calculate $sim(q, c)$;
  if $IS(c) > 1$ **then**
    rankScore(d_i) = $IS(c) * \alpha * sim(d_i, q) * sim(q, c)$;
  else
    rankScore(d_i) = $IS(c) * sim(d_i, q) * sim(q, c)$;
  end
end

Sort R based on rankScore;

Algorithm 3: Re-ranking Algorithm

term weights are obtained using the *tf.idf* formula described earlier. To calculate the rank score for each document, first the similarity of the document and the query is computed using a cosine similarity measure. Then, we compute the similarity of the document with each concept in the user profile to identify the best matching concept.

Once the best matching concept is identified, a rank score is assigned to the document by multiplying the interest score for the concept, the similarity of the document to the query, and the similarity of the specific concept to the query. If the interest score for the best matching concept is greater than one, it is further boosted by a tuning parameter α. Once all documents have been processed, the search results are sorted in descending order with respect to this new rank score.

V. Experimental Evaluation

Our experimental evaluation is designed to address three particular questions:

- Do the interest scores for individual concepts in the ontological profile converge?
- Do the changes in interest scores accurately reflect user interest in specific topics?
- Can the semantic evidence provided by the ontological profiles be used to effectively re-rank Web search results to present the user with a personalized view?

Since the queries of average Web users tend to be short and ambiguous [41], our goal is to demonstrate that re-ranking based on ontological user profiles can help in disambiguating the user's intent particularly when such queries are used.

A. Experimental Metrics

For the user profile convergence experiments, we employ two statistical measures; the arithmetic mean (average) and variance. We compute the average interest scores so that we can demonstrate the average rate of increase converges as a result of updating the ontological user profiles over time. Also, we utilize variance in order to measure how the interest scores are spread around the mean as a result of incremental updates. Our results are discussed in Section 5.3.

For the personalized search experiments, we measure the effectiveness of re-ranking in terms of *Top-n Recall* and *Top-n Precision*. For example, at $n = 100$, the top 100 search results are included in the computation of recall and precision, whereas at $n = 90$, only the top 90 results are taken into consideration.

Starting with the top one hundred results and going down to top ten search results, the values for *n* include $n = \{100, 90, 80, 70, ..., 10\}$. The *Top-n Recall* is computed by dividing the number of relevant documents that appear within the top *n* search results at each interval with the total number of relevant documents for the given concept.

$$\textit{Top-n Recall} = \frac{\textit{\# of relevant retrieved within n}}{\textit{total \# of relevant documents}}$$

We also compute the *Top-n Precision* at each interval by dividing the number of relevant documents that appear within the top *n* results with *n*.

$$\textit{Top-n Precision} = \frac{\textit{\# of relevant retrieved within n}}{n}$$

B. Experimental Data Sets

As of December 2006, the *Open Directory* contained more than 590,000 concepts. For experimental purposes, we use a branching factor of four with a depth of six levels in the hierarchy. Our experimental data set contained 563 concepts in the hierarchy and a total of 10,226 documents that were indexed under various concepts.

The indexed documents were pre-processed and divided into three separate sets including a *training set*, a *test set*, and a *profile set*. For all of the data sets, we kept track of which concepts these documents were originally indexed under in the hierarchy. The *training set* was utilized for the representation of the reference ontology, the *profile set* was used for spreading activation, and the *test set* was utilized as the document collection for searching.

The *training set* consisted of 5041 documents which were used for the one-time learning of the reference ontology. The concept terms and corresponding term weights were computed using the formula described in the Representation of Reference Ontology section.

A total of 3067 documents were included in the *test set*, which were used as the document collection for performing our search experiments. Depending on the search query, each document in our collection can be treated as a signal or a noise document. The signal documents are those documents relevant to a particular concept that should be ranked high in the search results for queries related to that concept. The noise documents are those documents that should be ranked low or excluded from the search results.

The *test set* documents that were originally indexed under a specific concept and all of its subconcepts were treated as signal documents for that concept whereas all other test set documents were treated as noise. In order to create an index for the signal and noise documents, a *tf.idf* weight was computed

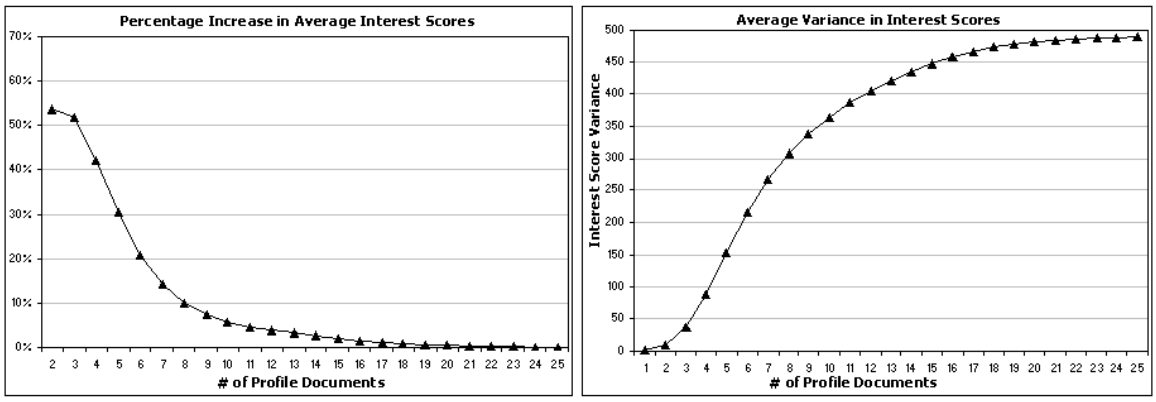


Fig. 4. The average rate of increase and average variance in *Interest Scores* as a result of incremental updates.

for each term in the document collection using the global dictionary of the reference ontology.

The *profile set* consisted of 2118 documents, which were treated as a representation of specific user interest for a given concept to simulate ontological user profiles. As we performed the automated experiments for each concept/query, only the profile documents that were originally indexed under that specific concept were utilized to build an ontological user profile by updating the interest scores with the spreading activation algorithm.

C. Experimental Methodology and Results

In this section, we provide our methodology and results for two independent but related aspects of our experimental evaluation. One aspect is to demonstrate user profile convergence. The second aspect of our evaluation is to design experiments to demonstrate the effectiveness of our approach for search personalization.

1) User Profile Convergence: With the user profile convergence experiments, our goal is to demonstrate that the rate of increase in interest scores stabilizes over incremental updates. Every time a new Web page, which the user has shown interest in, is processed via spreading activation, the interest scores for the concepts in the ontological user profile are updated.

Initially, the interest scores for the concepts in the profile will continue to change. However, once enough information has been processed for profiling, the amount of change in interest scores should decrease. Our expectation is that eventually the concepts with the highest interest scores should become relatively stable. Therefore, these concepts will reflect the user's primary interests.

To evaluate the user profile convergence, we used a single profile document for each concept and utilized that document as the input for the spreading activation algorithm for 25 rounds. We utilized the documents in the *profile set* for this experiment. For each concept, we used a profile document that was originally indexed under that specific concept, which we refer to as the signal concept.

Our methodology was as follows. We started with a given signal document and used a profile document to spread activation. As described in Section 3.4, after the propagation through the entire network is completed, the interest scores are normalized and updated. We recorded the interest scores for all concepts as well as the average interest score and variance across all concepts. This was considered round 1. For the same signal concept, we repeated the process for 25 rounds which is equivalent to updating the ontological user profile using 25 profile documents.

We ran the above experiment for 50 distinct signal concepts. The interest scores in the user profile were reset to one prior to processing each signal concept. Our goal was to measure the change in interest scores for the signal concept as well as the other concepts in the user profile.

As depicted in Figure 4, the average rate of increase for the interest scores for the signal concepts did converge. However, monitoring the interest scores for the signal concepts was not sufficient by itself. We needed to guarantee that the interest scores for all of the other concepts were not increasing at the same rate as the signal concept. Therefore, we computed the variance in interest scores after each round for a given signal concept.

Our expectation was that additional evidence in favor of a signal concept should result in discrimination of the signal concept from other concepts in the user profile. Figure 4 displays the average variance as a result of incremental updates. While the experimental conditions (repeated use of the same signal document) are somewhat artificial, the evaluation did confirm that the spreading activation mechanism is working correctly to focus the learned profile in the desired way.

2) User Profile Accuracy: With the user profile accuracy experiments, our goal is to demonstrate that the interest scores are maintained correctly with the incremental updates, especially in the case of mixed interests. Similar to the profile convergence experiments, we utilized the documents in the *profile set* for this experiment. We used a single profile document for each concept and utilized that document as the input for the spreading activation.

Our methodology was as follows. We identified a specific signal concept within the reference ontology. We used a profile document which belongs to the signal concept to spread activation. Same as the above experiments, the interest scores are normalized and updated after the propagation through the entire network is completed. We recorded the interest scores for all concepts for each round. For the same signal

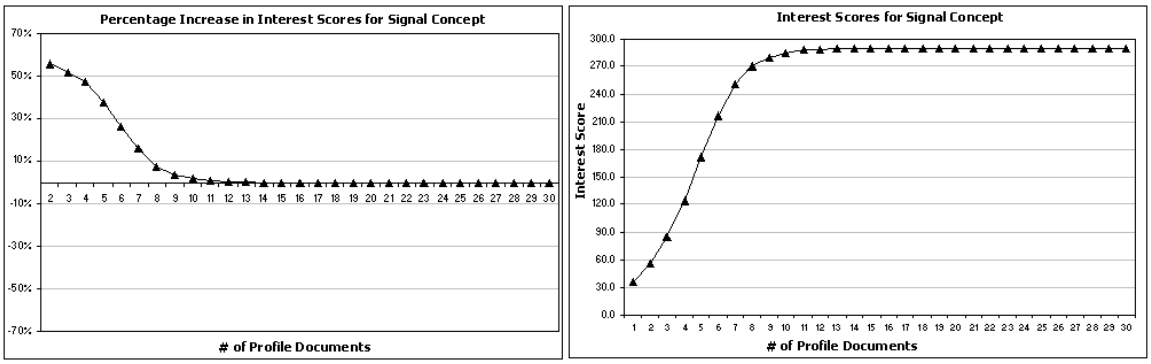


Fig. 5. Increase in *Interest Scores* for Signal concept, *Top/Science/Instruments and Supplies/Laboratory Equipment*

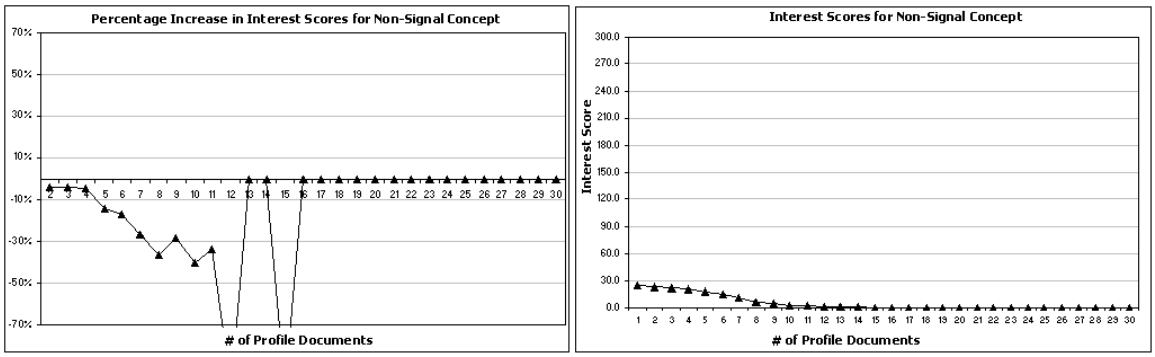


Fig. 6. Decrease in *Interest Scores* for Non-Signal concept, *Top/Computers/Artificial Intelligence/Vision*

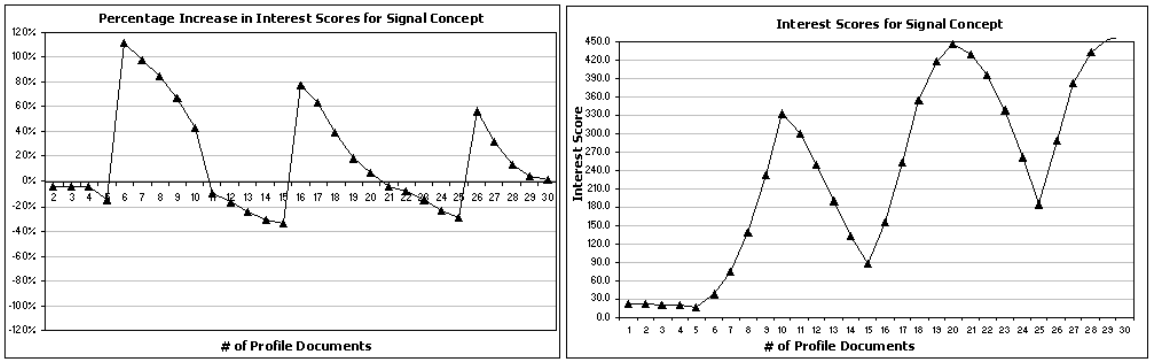


Fig. 7. Change in *Interest Scores* for Signal concept, *Top/Computers/Artificial Intelligence/Vision*

concept, we repeated the process for 30 rounds which is equivalent to updating the ontological user profile using 30 profile documents.

Again, the purpose of this somewhat artificial experiment was to ensure that the distribution of interest scores converged towards the signal concept and away from non-signal concepts, and that this effect was not significantly different between concepts in different parts of the ontology. Figures 5 and 6 show one such evaluation with the ”Laboratory Equipment” concept as signal. The interest scores for the signal increase uniformly. The non-signal concept ”Computer Vision” drops to zero interest after approximately 15 rounds.

We also performed another set of experiments where we treated a pair of concepts as signal. We used a separate profile document for each signal concept. We performed the spreading activation using the profile document for one of the signal concepts for the first 5 rounds and then using the profile document for the second signal concept for the next 5 rounds. We repeated the process for 30 rounds to monitor the change in interest scores for both concepts. Figure 7 displays the change in interest scores for one of the signal concepts as the profile documents are alternated every 5 rounds. The question here is whether the user model would converge to a bi-modal distribution of interest shared by both signal concepts. Although the actual interest score swings substantially, we can see that the overall trend is upward. The other concept in the pair has a similar shape. However, not every pair of concepts exhibited this form of stability. We are still investigating the behavior of the spreading activation model under these conditions.

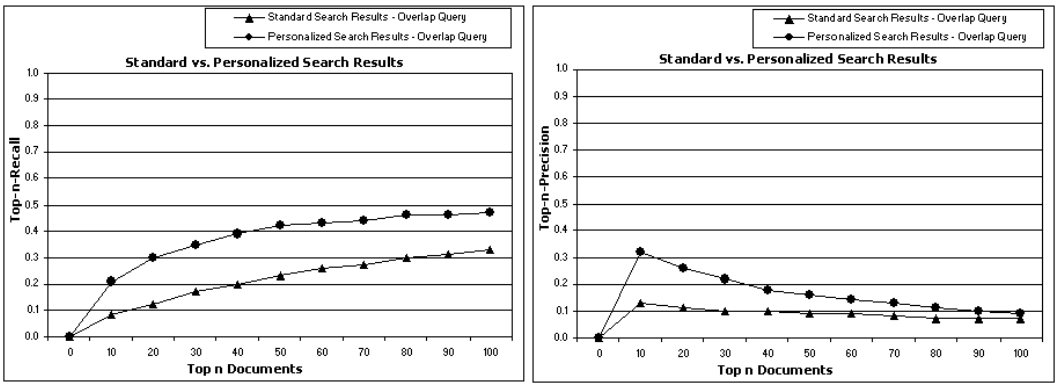


Fig. 8. Average *Top-n Recall* and *Top-n Precision* comparisons between the personalized search and standard search using "overlap queries".

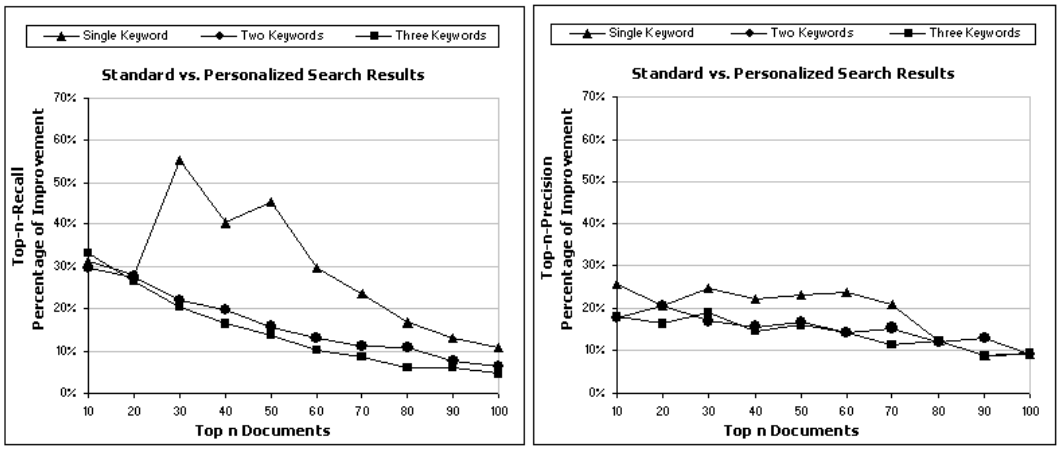


Fig. 9. Percentage of improvement in *Top-n Recall* and *Top-n Precision* achieved by personalized search relative to standard search with various query sizes.

3) Re-ranking Web Search Results: We constructed keyword queries to run our automated experiments. We decided to extract the query terms from the concept term vectors in the ontology. Each concept term vector was sorted in descending order with respect to term weights.

TABLE I
SET OF KEYWORD QUERIES

Query	# of Terms	Criteria
Set 1	1	highest weighted term in concept term vector
Set 2	2	two highest weighted terms in concept term vector
Set 3	3	three highest weighted terms in concept term vector
Set 4	2 or more	overlapping terms within highest weighted 10 terms

Table I depicts the four query sets that were automatically generated for evaluation purposes. Our keyword queries were used to run a number of automated search scenarios for each concept in our reference ontology. The first set of keyword queries contained only one term and included the highest weighing term for each concept. In order to evaluate the search results when a single keyword was typed by the user as the search query, the assumption was that the user was interested in the given concept.

The second set of queries contained two terms including the two highest weighing terms for each concept. The third set of queries were generated using the three highest weighing terms for each concept. As the number of keywords in a query increase, the search query becomes less ambiguous.

Even though one to two keyword queries tend to be vague, we intentionally came up with a fourth query set to focus specifically on ambiguous queries. Each concept term vector was sorted with respect to term weights. We compared the highest weighing ten terms in each concept with all other concepts in the ontology. A given concept was considered to be overlapping with another concept if a specific term appeared in the term vectors of both concepts.

The parents, children, and siblings of the concept were excluded when identifying the overlapping concepts for a given concept. Only the overlapping concepts were included in the experimental set with each query consisting of two or more overlapping terms within these concepts.

Our evaluation methodology was as follows. We used the system to perform a standard search for each query. As mentioned above, each query was designed for running our experiments for a specific concept. In the case of standard search, a term vector was built using the original keyword(s) in the query text. Removal of stop words and stemming was utilized. Each term in the original query was assigned a weight of 1.0.

The search results were retrieved from the test set, the signal and noise document collection, by using a cosine similarity measure for matching. Using an interval of ten, we calculated the *Top-n Recall* and *Top-n Precision* for the search results.

Next, documents from the profile set were utilized to simulate user interest for the specific concept. For each query, we started with a new instance of the ontological user profile with all interest scores initialized to one. Such a user profile represents a situation where no initial user interest information is available. We performed our spreading activation algorithm to update interest scores in the ontological user profile.

After building the ontological user profile, we sorted the original search results based on our re-ranking algorithm and computed the *Top-n Recall* and *Top-n Precision* with the personalized results.

In order to compare the standard search results with the personalized search results, we computed the average *Top-n Recall* and *Top-n Precision*, depicted in Figure 8.

We have also computed the percentage of improvement between standard and personalized search for *Top-n Recall* and *Top-n Precision*, depicted in Figure 9.

D. Discussion of Experimental Results

Personalized search provides the user with results that accurately satisfy their specific goal and intent for the search. The queries used in our experiments were intentionally designed to be short to demonstrate the effectiveness of our Web search personalization approach, especially in the typical case of Web users who tend to use very short queries.

Simulating user behavior allowed us to run automated experiments with a larger data set. In the worst case scenario, the user would enter only a single keyword. The evaluation results show significant improvement in recall and precision for single keyword queries as well as gradual enhancement for two-term and three-term queries. As the number of keywords in a query increase, the search query becomes more clear.

In addition to the one, two, and three keyword queries, we ran experiments with the overlap query set to focus on ambiguous queries. Two users may use the exact same keyword to express their search interest even though each user has a completely distinct intent for the search. For example, the keyword *Python* may refer to *python as a snake* as well as the *Python programming language* sense.

The purpose of the overlap queries is to simulate real user behavior where the user enters a vague keyword query as the search criteria. Our evaluation results verify that using the ontological user profiles for personalizing search results is an effective approach. Especially with the overlap queries, our evaluation results confirm that the ambiguous query terms are disambiguated by the semantic evidence in the ontological user profiles.

With the user profile and accuracy experiments, we have evaluated the stability of our approach separately from its performance in terms of Web search personalization. We have validated the interest propagation within the user profiles and demonstrated the effectiveness of profile normalization, especially in the case of mixed interests.

VI. Conclusion

We have presented a framework for contextual information access using ontologies and demonstrated that the semantic knowledge embedded in an ontology combined with long-term user profiles can be used to effectively tailor search results based on users' interests and preferences.

In our future work, we plan to continue evaluating the stability and convergence properties of the ontological profiles as interest scores are updated over consecutive interactions with the system. Since we focus on implicit methods for constructing the user profiles, the profiles need to adapt over time. Our future work will involve designing experiments that will allow us to monitor user profiles over time to ensure the incremental updates to the interest scores accurately reflect changes in user interests.

References

[1] R. Kraft, F. Maghoul, and C. C. Chang, "Y!q: contextual search at the point of inspiration," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005*, Bremen, Germany, November 2005, pp. 816–823.

[2] A. Singh and K. Nakata, "Hierarchical classification of web search results using personalized ontologies," in *Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction, HCI International 2005*, Las Vegas, NV, July 2005.

[3] M. Aktas, M. Nacar, and F. Menczer, "Using hyperlink features to personalize web search," in *Advances in Web Mining and Web Usage Analysis, Proceedings of the 6th International Workshop on Knowledge Discovery from the Web, WebKDD 2004*, Seattle, WA, August 2004.

[4] O. Boydell and B. Smyth, "Capturing community search expertise for personalized web search using snippet-indexes," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006*, Arlington, VA, November 2006, pp. 277–286.

[5] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harman, D. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai, "Challenges in information retrieval and language modeling," *ACM SIGIR Forum*, vol. 37, no. 1, pp. 31–47, 2003.

[6] S. Lawrence, "Context in web search," *IEEE Data Engineering Bulletin*, vol. 23, no. 3, pp. 25–32, 2000.

[7] X. Shen, B. Tan, and C. Zhai, "Ucair: Capturing and exploiting context for personalized search," in *Proceedings of the Information Retrieval in Context Workshop, SIGIR IRiX 2005*, Salvador, Brazil, August 2005.

[8] J. Teevan, S. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005*, Salvador, Brazil, August 2005, pp. 449–456.

[9] M. Speretta and S. Gauch, "Personalized search based on user search histories," in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005*, Compigne, France, September 2005, pp. 622–628.

[10] A. Micarelli and F. Sciarrone, "Anatomy and empirical evaluation of an adaptive web-based information filtering system," *User Modeling and User-Adapted Interaction*, vol. 14, no. 2-3, pp. 159–200, 2004.

[11] F. Gasparetti and A. Micarellli, "A personal agent for browsing and searching," in *Proceedings of the 2nd International Conference on Autonomous Agents*, St. Paul, MN, May 1998, pp. 132–139.

[12] D. Mladenic, "Personal webwatcher: Design and implementation," *Technical Report IJS-DP-7472*, 1998.

[13] G. Gentili, A. Micarelli, and F. Sciarrone, "Infoweb: An adaptive information filtering system for the cultural heritage domain," *Applied Artificial Intelligence*, vol. 17, no. 8-9, pp. 715–744, 2003.

[14] F. Gasparetti and A. Micarellli, "User profile generation based on a memory retrieval theory," in *Proceedings of the 1st International Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces, WPRSIUI 2005*, Reading, UK, October 2005.

[15] H. Haav and T. Lubi, "A survey of concept-based information retrieval tools on the web," in *5th East-European Conference, ADBIS 2001*, Vilnius, Lithuania, September 2001, pp. 29–41.

[16] J. Trajkova and S. Gauch, "Improving ontology-based user profiles," in *Proceedings of the Recherche d'Information Assiste par Ordinateur, RIAO 2004*, University of Avignon (Vaucluse), France, April 2004, pp. 380–389.

[17] A. Sieg, B. Mobasher, and R. Burke, "Web search personalization with ontological user profiles," in *ACM Sixteenth Conference on Information and Knowledge Management, CIKM 2007*, Lisbon, Portugal, November 2007.

[18] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing," *Web Intelligence and Agent Systems*, vol. 1, no. 3-4, 2003.

[19] D. Ravindran and S. Gauch, "Exploting hierarchical relationships in conceptual search," in *Proceedings of the 13th International Conference on Information and Knowledge Management, ACM CIKM 2004*, Washington DC, November 2004.

[20] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using odp metadata to personalize search," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005*, Salvador, Brazil, August 2005, pp. 178–185.

[21] C. Ziegler, K. Simon, and G. Lausen, "Automatic computation of semantic proximity using taxonomic knowledge," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006*, Arlington, VA, November 2006, pp. 465–474.

[22] F. Liu, C. Yu, and W. Meng, "Personalized web search for improving retrieval effectiveness," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 28–40, 2004.

[23] P. Chirita, C. Firan, and W. Nejdl, "Summarizing local context to personalize global web search," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006*, Arlington, VA, November 2006, pp. 287–296.

[24] F. Tanudjaja and L. Mui, "Persona: A contextualized and personalized web search," in *Proceedings of the 35th Annual Hawaii International Conference on System Sciences, HICSS 2002*, Big Island, Hawaii, January 2002, p. 67.

[25] H. Chang, D. Cohn, and A. McCallum, "Learning to create customized authority lists," in *Proceedings of the 7th International Conference on Machine Learning, ICML 2000*, San Francisco, CA, July 2000, pp. 127–134.

[26] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the 12th international conference on World Wide Web, WWW 2003*, Budapest, Hungary, May 2003, pp. 271–279.

[27] P. A. Chirita, D. Olmedilla, and W. Nejdl, "Pros: A personalized ranking platform for web search," in *Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2004*, Eindhoven, The Netherlands, August 2004.

[28] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the 11th International World Wide Web Conference, WWW 2002*, Honolulu, Hawaii, May 2002.

[29] F. Qiu and J. Cho, "Automatic identification of user interest for personalized search," in *Proceedings of the 15th International World Wide Web Conference, WWW 2006*, Edinburgh, Scotland, May 2006, pp. 727–736.

[30] B. Schilit and M. Theimer, "Disseminating active map information to mobile hosts," *IEEE Network*, vol. 8, no. 5, pp. 22–32, 1994.

[31] A. Sieg, B. Mobasher, and R. Burke, "Representing context in web search with ontological user profiles," in *Proceedings of the Sixth International and Interdisciplinary Conference on Modeling and Using Context*, Roskilde, Denmark, August 2007.

[32] T. R. Gruber, "Towards principles for the design of ontologies used for knowledge sharing," in *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993.

[33] S. Middleton, N. Shadbolt, and D. D. Roure, "Capturing interest through inference and visualization: Ontological user profiling in recommender systems," in *Proceedings of the International Conference on Knowledge Capture, K-CAP 2003*, Sanibel Island, Florida, October 2003, pp. 62–69.

[34] A. Sieg, B. Mobasher, S. Lytinen, and R. Burke, "Using concept hierarchies to enhance user queries in web-based information retrieval," in *Proceedings of the International Conference on Artificial Intelligence and Applications, IASTED 2004*, Innsbruck, Austria, February 2004.

[35] S. Dumais, T. Joachims, K. Bharat, and A. Weigend, "Implicit measures of user interests and preferences," *ACM SIGIR Forum*, vol. 37, no. 2, 2003.

[36] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[37] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill, 1983.

[38] G. Salton and C. Buckley, "On the use of spreading activation methods in automatic information," in *Proceedings of the 11th annual international ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 1988*, Grenoble, France, 1988, pp. 147–160.

[39] H. Alani, K. O'Hara, and N. Shadbolt, "Ontocopi: Methods and tools for identifying communities of practice," in *Proceedings of the IFIP 17th World Computer Congress - TC12 Stream on Intelligent Information Processing*, Deventer, The Netherlands, The Netherlands, 2002, pp. 225–236.

[40] C. Rocha, D. Schwabe, and M. P. de Aragao, "A hybrid approach for searching in the semantic web," in *Proceedings of the 13th international conference on World Wide Web, WWW 2004*, New York, NY, USA, 2004, pp. 374–383.

[41] A. Spink, H. Ozmutlu, S. Ozmutlu, and B. Jansen, "U.s. versus european web searching trends," *ACM SIGIR Forum*, vol. 15, no. 2, 2002.

Ahu Sieg is currently working on her Ph.D. in Artificial Intelligence at DePaul University. She has extensive experience as a senior developer and analyst in web and client/server technologies covering all phases of project development including business requirements analysis, architecture and design, user interface design, testing and quality assurance, implementation and deployment. She is a Microsoft Certified Solution Developer (MCSD) for .NET. She graduated from Illinois Wesleyan University in 1997 with a B.S. degree. She completed her undergraduate study with a double major in Computer Science and Business Administration. She received her M.S. degree in Management Information Systems in 2000 from Benedictine University. Her research interests include artificial intelligence, information retrieval, and cognitive science.

Dr. Bamshad Mobasher is an Associate professor of Computer Science and the director of the Center for Web Intelligence at DePaul CTI. He received his Ph.D. in Computer Science at Iowa State University in 1994. His research areas include data mining, Web mining, intelligent agents, machine learning, and computational logic. He has published over 100 scientific articles and book chapters in these areas. Dr. Mobasher is one of the leading authorities in the areas of Web mining, Web personalization, and recommender systems, and has served as an organizer and on the program committees of numerous related conferences. As the director of the Center for Web Intelligence, he directs research in these areas and regularly works with the industry on various joint projects. He has been an organizer of series of highly regarded workshops on Web mining and personalization, including the WebKDD series of workshops on Knowledge Discovery on the Web at ACM SIGKDD and a series of workshops on Intelligent Techniques for Web Personalization at AAAI. Dr. Mobasher serves on the editorial boards of several prominent journals, including User Modeling and User-Adapted Interaction and the Journal of Web Semantics.

Dr. Robin Burke is an Associate Professor at CTI. His research interests are in artificial intelligence (especially case-based reasoning) as applied to electronic commerce and digital libraries. His current work concentrates on the area of recommender systems, including comparative evaluation of hybrid designs and the investigation of the security properties of recommender system algorithms. Professor Burke earned his PhD in 1993 from Northwestern University, working with Professor Roger Schank, one of the founders of the field of cognitive science. He worked in post-doctoral positions at the University of Chicago, and then in 1998, left academic employment to help found a ”dot-com” startup. He returned to academic work in 2000 first at the University of California, Irvine and then at California State University, Fullerton. In the Fall of 2002, he began his current position at DePaul University.

Conversational Informatics and Human-Centered Web Intelligence

Toyoaki Nishida, *Member, IEEE*

***Abstract*— Conversation is the most natural communication means for people to communicate with each other. I believe that conversation plays a critical role in realizing a paradigm of human-centered web intelligence in which web intelligence engines are grounded on the human society. We are currently building a computational framework for circulating information in a conversational fashion, using information packages called conversation quanta that encapsulate conversational scenes. Technologies are being developed for acquiring conversation quanta on the spot, accumulating them in a visually recognizable form, and reusing them in a situated fashion. Conversational Informatics, based on measurement, analysis, and modeling of conversation, constitutes the theoretical foundation for these applications. I will overview recent results in Conversational Informatics that will help achieve our vision. I will also discuss our approach in the context of Social Intelligence Design aimed at the understanding and augmentation of social intelligence for collective problem solving and learning.**

***Index Terms*— Conversational Informatics, Social Intelligence Design, Human-centered computing, Human computer interaction**

I. Introduction

The goal of Web Intelligence is to create a world wide wisdom web (w4) by integrating individual intelligences available on the global network using technologies such as web agents, web mining and farming, web information retrieval, web knowledge management, web intelligence infrastructure, and social network intelligence [1].

In order for Web Intelligence to be able to maximally benefit the human society, it should be well-interfaced to the human society so that each member of the human society can benefit from it without much difficulty and Web Intelligence can gain enough sources of knowledge from the human society. Web Intelligence will synergistically co-evolve with the human society if it is intimately embedded in the human society.

One of the key issues in embedding Web Intelligence in the human society is information grounding, which roughly means that the information user is aware of the association between information and the real world. Web Intelligence need to provide information in such a way that people can readily ground it on their daily life. Even though potentially useful information is provided with Web Intelligence, it might be useless if the information user fails to recover the reference to real world or reconfigure the image implied by the given statement. Unfortunately, information grounding is not easy to establish once information is isolated from the original situation it is created unless special care is taken at the moment information is created. We need to invent a technology that permits people to preserve cues for information grounding on the spot so that they can help ground the information later at the situations different from the original one.

In this article, I focus on the role of conversation in information grounding and present a suite of technologies aimed at realizing a paradigm of human-centered web intelligence. Apparently, conversation is the most natural communication means for people to communicate with each other. A closer look reveals that various kinds of processes related to creation or recovering information grounding are in action in conversation. For example, pointing and gaze are basic forms of creating association with the real world and co-occurring propositions. Gestures and postures may suggest the scope and modality of the utterances. In addition, dialectic aspects of conversation help participants interpret the meaning of information through discussions.

We are currently building a computational framework for circulating information in a conversational fashion, using information packages called conversation quanta that encapsulate conversational scenes consisting of participants' behavior, references to the environment, and meta-descriptions. Technologies are being developed for acquiring conversation quanta on the spot, accumulating them in a visually recognizable form, and reusing them in a situated fashion.

Conversational Informatics is a field of research aimed at establishing the theoretical foundation for these applications, based on measurement, analysis, and modeling of conversation. The field exploits a foundation provided by Artificial Intelligence, Natural Language Processing, Speech and Image Processing, Cognitive Science, and Conversation Analysis. It is aimed at shedding light on meaning creation and interpretation resulting from the sophisticated mechanisms in verbal/nonverbal interactions during conversation, in search of better methods of computer-mediated communication, human/computer interaction, and support for knowledge

Toyoaki Nishida is with Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan (phone: +81-75-753-5371; fax: +81-75-753-4961; e-mail: nishida@i.kyoto-u.ac.jp).

creation. I will overview recent results in Conversational Informatics that will help achieve our vision.

I will also discuss our approach in the context of Social Intelligence Design aimed at the understanding and augmentation of social intelligence for collective problem solving and learning.

II. Conversational Informatics

Conversational Informatics [2] is a field of research that serves as a theoretical ground for understanding conversational phenomena and developing conversational systems. By integrating the methods in Artificial Intelligence, Pattern Recognition, and Cognitive Science, Conversational Informatics addresses understanding and augmenting conversation. Currently, we focus on shallow social implications that manifest in the nonverbal communications. The engineering aspects are emphasized to investigate conversations using sensors ranging from audio-visual and motion sensors to those for biological and brain measurement

A. The Lack of Situated Information

The advance of the information network infrastructure has connected people with each other and brought about the role of computers as a mediator in the human society. In spite of the huge amount of information made available on the net, we are still suffering from the lack of relevant information. Even for pursuing daily activities such as setting up a presentation for a lecture by connecting the PC to a projector, a certain amount of situation-specific information is needed (Figure 1). For example, the switches and controllers of presentation facilities are located in different places depending on the room, and there are subtle differences in operation sequences and semantics. Since such information is often shared by a handful of the local users and is too expensive to carefully maintain, it is often left implicit without much attention. As a result, new comers and casual users are left behind the latest updates, and disastrous failures take place from time to time. Certainly, we do not have enough situated information for daily life and need more.

B. Conversational Knowledge Process

Conversation is a handy means for people to communicate situated information. Conversation is dynamic information medium. In contrast with describing the situation in a static fashion, say by using a picture image and a written text, one can directly describe the situation by combining utterances with nonverbal communication actions, such as pointing or gaze, which are quite natural to people. For example, one might be able to communicate rather situation-dependent information as shown in Figure 2. In a more complex conversational setting with multiple participants, each participant may make structural interactions to manage shared information, as shown in Figure 3.

In conversation quantization [3], we introduce a conversation quantum that encapsulates interaction-oriented and content-oriented views of a conversation scene. A conversation quantum represents both content and interaction

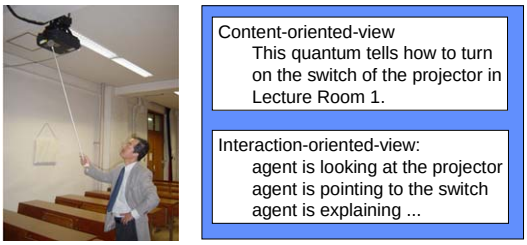
Figure 1: A certain amount of information is necessary even for a daily activity like setting up a presentation for a lecture.

Figure 2: Use of conversational description style to communicate situation-dependent information.

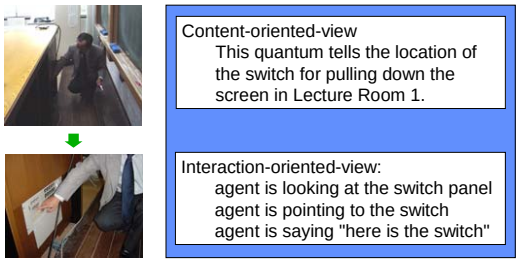
Figure 3: Multiparty conversation.

of the conversation scene. For example, a couple of conversation scenes shown in Figure 2 might be encoded in a conversational quantum as illustrated in Figure 4.

Conversational Knowledge Process is a framework for



(a) Conversation quantum for the scene in Figure 2a



(b) Conversation quantum for the scene in Figure 2b

Figure 4: Representing conversational scene by conversation quantum.

circulating conversation quanta in a community. It mainly consists of conversation quanta acquisition, accumulation and presentation, as shown in Figure 5.

Conversation quanta acquisition is a process of generating conversation quanta for a given conversation scene. So far, we have been manually encoding conversational interactions. We are now building a (semi) automated method by measuring and analyzing the participants behaviors in conversation.

Conversation quanta accumulation is a stage for accumulating conversation quanta on a server so that they can be reused in other conversation scenes. In order to allow the user to edit existing conversation quanta or to create new ones from the archive, we have developed a tool for visually manipulating the collection of conversation quanta.

Conversation quanta presentation is a stage for reproducing conversational interactions in conversation scenes. Embodied conversational agents or conversational robots are used to play a role of a participant in a conversation scene.

C. Conversation Measurement in the IMADE Room

We place much emphasis on the measuring nonverbal behaviors using the state-of-the-art sensing devices such as motion capture devices or eye trackers, rather than employing deep reasoning or planning algorithms, for we would like to gain the quality of conversation by preserving the subtle details, and also implement light-weight and robust algorithms.

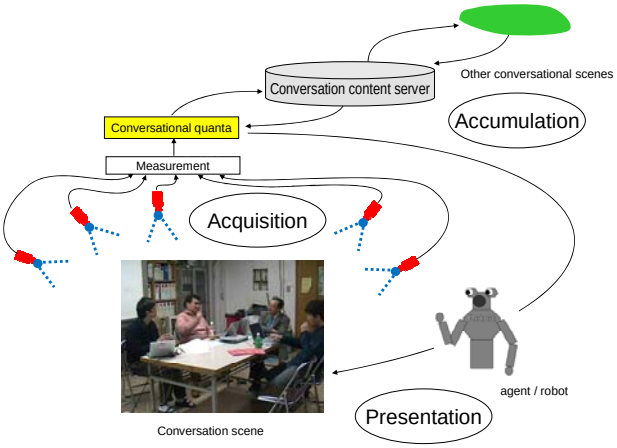
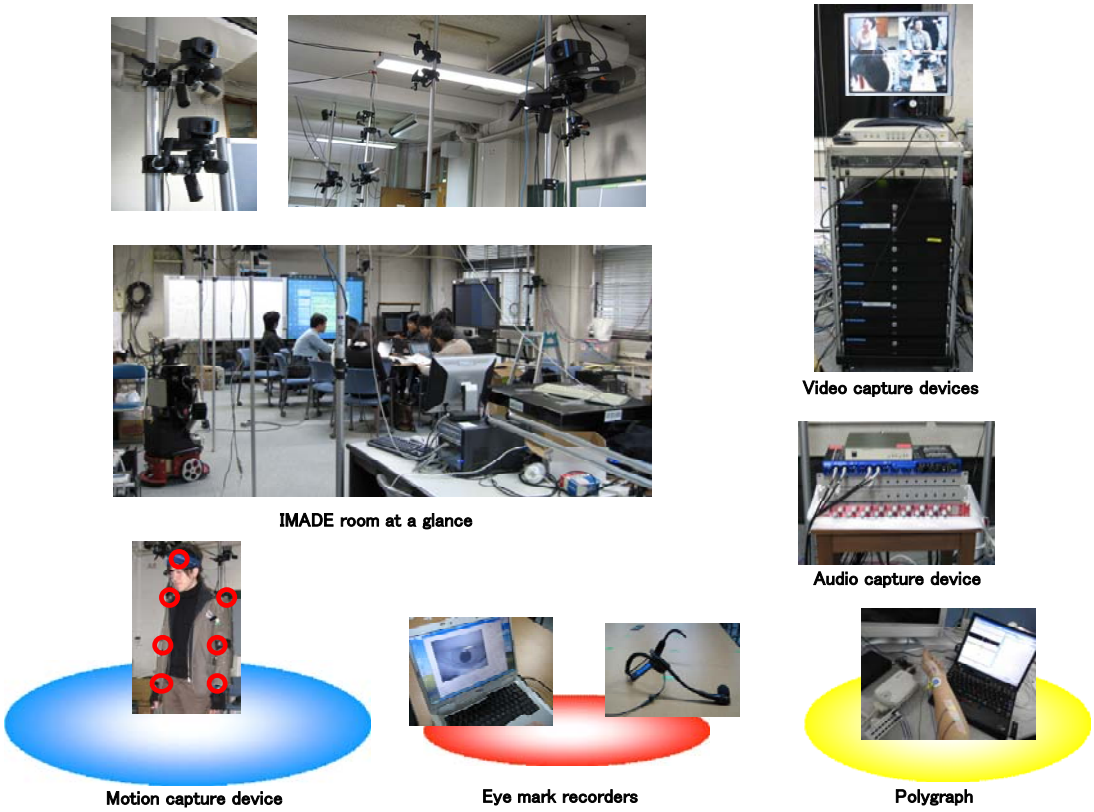


Figure 5: Conversational Knowledge Process based on conversation quantization.

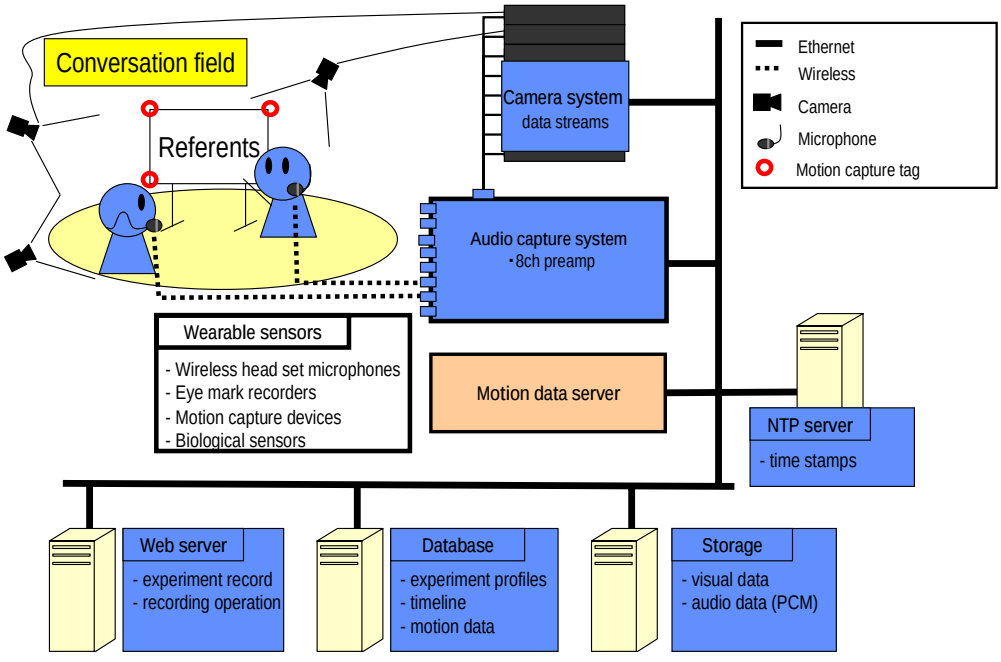
In order to study conversations by measurement and corpus building, we are developing an environment called IMADE (the real world Interaction Measurement, Analysis and Design Environment (Figure 6) [4,5]. In addition to multi-modal sensing devices, such as the wearable motion capture devices or eye mark recorders, we plan to introduce biological and brain measurement devices so that we can observe the internal activities and their interdependencies of each participant in a given conversational situation.

We have made preliminary experiments on conversation measurement and analysis. A tool called iCorpusStudio was developed for browsing, analyzing, annotating interaction corpus accumulating data obtained from experiment session [5]. In the first experiment, the behavior of group of people engaging in a collaborative design using a common display was recorded. The obtained data is being analyzed from the viewpoint of social discourse based on verbal and nonverbal interactions.

In the second experiment [5], a more complex setting was introduced to observe the dynamics of the participatory structure during the collaborative design session. In this experiment, two referents were placed in the field to see how the subject group would change the formation as discussion proceeded. Some interesting group behavior was observed that suggested the relationship between nonverbal behaviors of the participants and the group dynamics. Figure 6 demonstrates the analysis with iCorpusStudio. The analyzer is able to compare the video, audio-visual data, and annotations to study the interaction patterns observed in the session. In this example, although the subject S1 might appear to lead the migration from the left panel to the right at a glance, it turns out more likely that S1 simply dropped from the conversation and followed by the migration initiated by S2, according to the detailed analysis. Such detailed analysis is made available only by closely recording the gaze, gesture, posture, and speech of the subjects in detail, and showing an integrated view so that co-occurrences and temporal patterns of events across different modalities can be observed at a glance.



(a) Overview



(b) System configuration

Figure 5: IMADE (the real world Interaction, Measurement, Analysis and Design Environment) room [4].

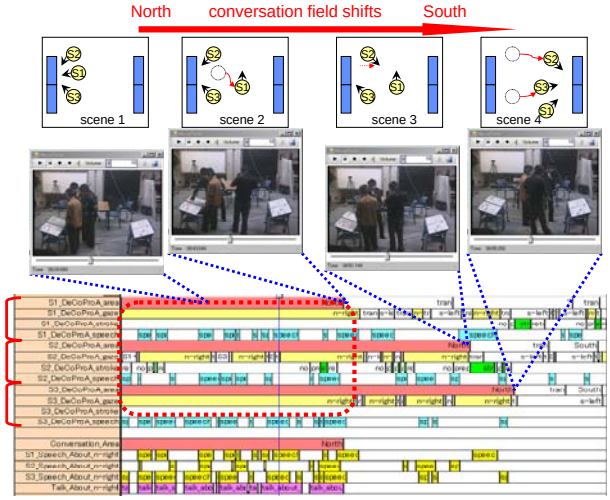


Figure 6: Analysis of multiparty interaction using iCorpusStudio [4].

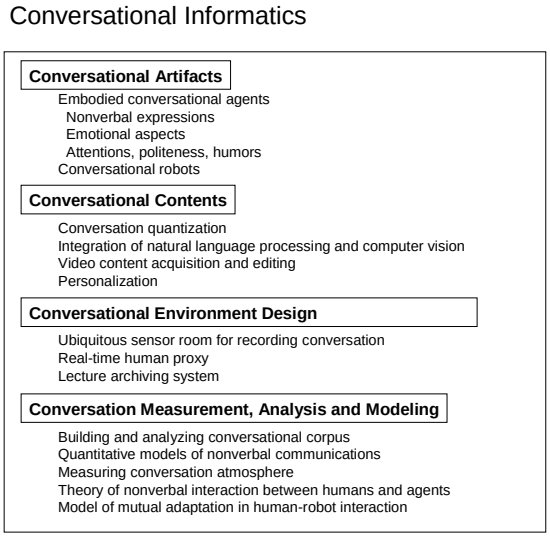


Figure 7: Framework of Conversational Informatics [5].

D. Conversational Informatics -- State of the Art

The current development of Conversational Informatics consists of four subjects (Figure 7).

The first subject is conversational artifacts (embodied conversational agents or conversational robots) that can participate in human conversations. Our study involves algorithms for interpreting and presenting nonverbal expressions to permit the user to interact with them in a conversational fashion, not only with natural language but also with eye gaze, facial expressions, gestures, or other nonverbal communication means. Socio-emotional implications of conversation such as attentions, politeness, friendliness, or personality are investigated with great interest.

The second subject is about manipulating conversational contents that encapsulate information arising in conversation scenes. Techniques are being developed for accumulating, editing, and converting conversational contents, using natural language processing, computer vision, and human computer interaction.

The third subject is conversation environment design. The primary goal is designing an intelligent environment that can sense and augment the conversational interactions. Work is in progress to provide situated information supports by combining wearable or environment sensors and displays in conversation scenes ranging from poster sessions to large classrooms.

The last subject is conversation measurement, analysis and modeling, driven by scientific interest. Introduction of powerful sensing technologies significantly accelerates the study. It will not only permit a data-driven quantitative understanding of conversational behaviors but also enable a corpus-based development of conversational systems that are more robust and sophisticated than those by pure programming.

In the next three sections, I would like to survey recent developments in Conversational Informatics.

III. Conversational Artifacts

The role of conversational artifacts is to provide a conversational interface with the user. Conversational artifacts consist of two major categories depending on whether they have a physical body or not.

A. Embodied Conversational Agents

Embodied conversational agents are interactive synthetic characters that have CG-based embodiment. Embodied conversational agents have a relatively long history of development. They originate from synthetic characters and natural language dialogue systems. More sophisticated nonverbal interaction functions have been incorporated as the technologies advance [6, 7]. The generic, component-based platform is being employed, rather than hard-wired application-specific architecture. More emphasis is placed on simulating subtle features of nonverbal expressions based on corpus and allowing large-scale rich content to be referred to in conversation.

The GECA (Generic ECA) is a generic framework for building an ECA system on multiple servers connected with each other by a computer network [8]. GECA allows for mediating and transporting data stream and command messages among software modules. It provides with a high-level protocol for exchanging XML messages among components such as input sensors, inference engines, the emotion model, the personality model, the dialogue manager, the face and body animation engines, etc. An application programming interface is made available on main-stream operating systems so that the programmer can easily adapt ECA software modules to incorporate into the GECA platform. The blackboard model is employed as the backbone.

GECA has been implemented and applied for various applications involving a navigation agent, a quiz agent, and a pedagogical agent for teaching cross cultural communication.

A navigation agent was designed to make a spatial navigation for the user. The user can talk to the navigation agent about objects in the background, by combining natural language, hand pointing and head movements. In response, the navigation agent combines speech with eye gaze, facial expressions, hand gestures, and postures to guide the user in a simulated place, possibly by moving around there.

The quiz agent was designed to entertain visitors at open house events of a research institute [9]. An interactive synthetic character is displayed on a large screen. When a visitor arrives, the quiz agent will give the user a number of puzzles. Each puzzle consists of a question followed by several alternatives. When the visitor chooses one, the quiz agent will tell whether the choice is correct or not, and explain the correct answer when the visitor's answer is wrong. Touch panel was chosen as the input device for complex sensors were considered to be unstable and might worry the visitor. Emotional feedback was implemented, using the PAD space model. Positive stimulus will be given to the emotion and mood when the visitor tries to answer the quiz. Even higher values will be given if the answer is correct. In contrast, negative stimulus will be given when the answer is wrong. The value on boredom axis will be increased when no input is given from the visitor in a certain amount of time.

This quiz agent was demonstrated to the public in a one-day public open lab event of NFRI (National Food Research Institute) on April 20th, 2007. In the demonstration session lasting for six hours, 307 visitors in small groups played the kiosk and 87 game sessions were run. The analysis of questionnaire revealed that most of the visitors enjoyed the game and felt that the knowledge explained by the agent was trustable.

B. Conversational Robots

Robots' physical embodiment normally yields a high presence and strong social implication in communication. Efforts have been made to build conversational robots that can participate in conversations. In early days of development, communication was made only with speech interface. Recent implementations, in contrast, place much emphasis on the nonverbal communication abilities. Nishida et al [10] has proposed the notion of robot as an embodied knowledge medium, where robots bear a role of mediating knowledge among people. The listener and presenter robots were prototyped to investigate the feasibility of the idea. The listener robot was designed to videotape critical scenes while interacting with an instructor. In the meanwhile, the presenter robot was designed to assist a novice user by showing appropriate video clip on a small display attached on the arm.

Both the listener and presenter robots were designed to detect critical behaviors of the user such as gaze or pointing to coordinate behaviors by intentionally making communicative acts such joint attention.

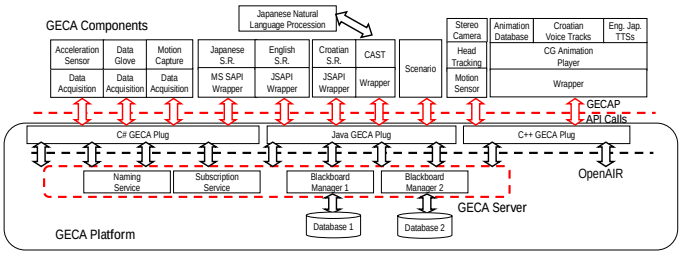
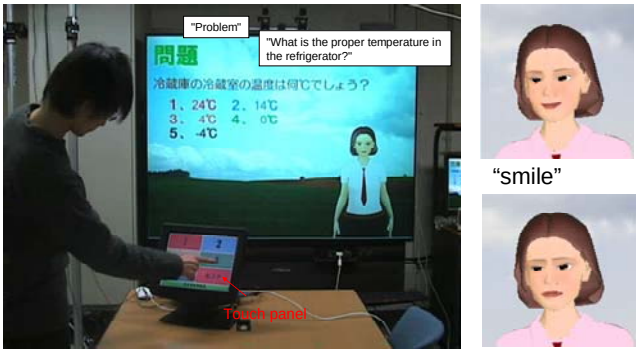


Figure 8: The Architecture of GECA [8].



the quiz agent in operation

Figure 9: The quiz agent [9].

In case of the listener robot, for example, ten 3D position sensors were attached to the instructor's body, and several 3D position sensors were used to identify the location of the salient objects in the environment. The user's status is sensed by a motion capture device and interpreted using Bayesian networks.

As a result, the listener robot can distinguish transitions of critical conversation modes, such as the talking-to, or talking-about modes.

Figure 10 shows how the listener robot interacts with the human user. In Figure 10a, the listener robot makes a joint attention according to the instructor's pointing gesture. Figure 10b shows the image of the object captured by the listener robot's eyes at that moment. Figure 10c and d shows how the listener robot interacts with two instructors. In Figure 10c, the two instructors are talking to the robot, and the robot replies to the person in the left, while in Figure 10d, both the instructor in the left and the robot are looking at the work of the instructor in the right.

Figure 11 illustrates the way the presenter robot behaves. The presenter robot coordinates eye gaze, posture, and motion


(a) Joint attention


(b) Image from the robot's eye


(c) *A* is talking


(d) Robot is looking at *B*

Figure 10: The listener robot.

according to the user's behavior. As a number of experimental evaluations, it turned out that the user was able to complete the task about 63% of the baseline setting with the fixed display position, with less (about 50%) interaction time and less frequency (54%) of interaction.

IV. Conversational Content

Technologies are being developed to help people create and manage a large amount of contents collected from conversations.

A. Visual Accumulation of Conversational Contents

The Sustainable Knowledge Globe (SKG) [11] is a system that allows the user to visually accumulate a large amount of conversational contents on the sphere surface so that s/he can a long-term relationship with them to complement the limitation of her/his biological memory. Conversational contents may be grouped into a tree structure so that the user can manipulate them as a group. A graphical user interface is employed to continuously zoom in/out the any region of the sphere surface, as shown in Figure 12. A linear zooming method is employed to avoid distortion of the landscape on the sphere surface. In order to help the user visually recognize the tree structure, we have introduced nesting contours. Embodied conversational agent was installed on the SKG system to navigate the user by presenting the content interactively.

B. Media Conversion

Media conversion is a powerful method to obtain conversational contents from a huge amount of legacy contents, such as natural language documents or archived videos. Kurohashi et al [12] developed a method for automatically converting a collection of series of short documents called knowledge cards into conversational contents consisting of spoken language scenario and summarization slides that can be

Figure 11: The presenter robot [9].

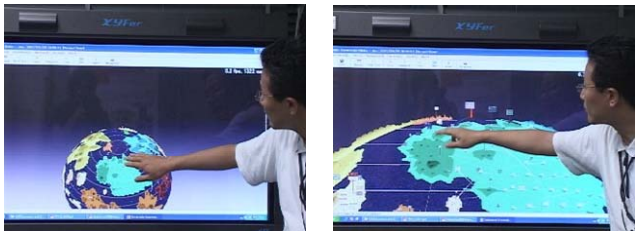


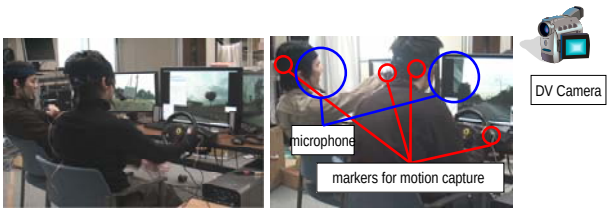
Figure 12: Sustainable Knowledge Globe [11].

automatically presented in a conversational fashion using embodied conversational agents. The method is based on corpus-driven natural language processing techniques for automatic construction of large-scale case frame, analysis of predicate-argument structure, and discourse structure analysis.

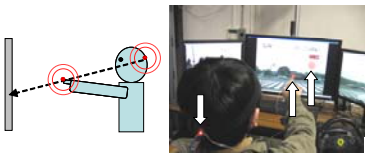
Nakamura [13] proposed an automated video content acquisition and editing for small meetings. The system follows two stages to generate a video summary for a dialogue. The first stage is content capture. The environment cameras and content production cameras are used for video capture. The system controls cameras to keep typical picture compositions such as close-up/bust shot, over-the-shoulder shot, or long shots. Contents capturing camera modules detect and track the face of the participants. The second stage is editing conversational scenes. At this stage, video streams from contents capturing cameras are edited into one stream, by maximally satisfying constraints extracted from various camera switching techniques. Speeches, motion, facial expressions, object movements, etc are taken into account for the processing.

V. Conversational Environment Design

The goal of conversational environment design is to provide a smart environment that allows people to pursue effective knowledge creation through conversations. Approaches vary depending on the size of the conversation environment, whether the environment is distributed or not, how much auxiliary devices can be introduced, how much quality is required, how much cooperation is expected from the participants, how much cost can be spent on the environment,

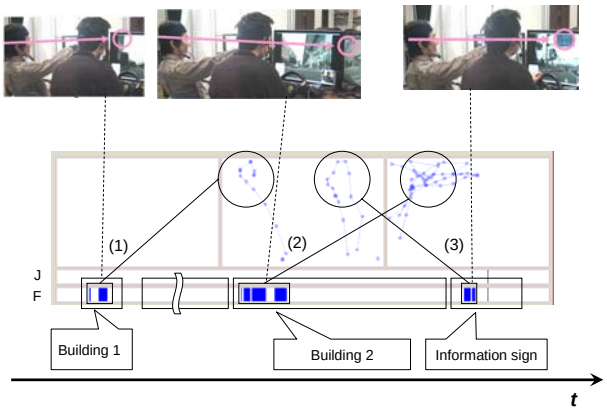


(a) setting of the recording devices



(b) detecting pointing gestures

Figure 13: An augmented conversational environment for a driving simulator [14].



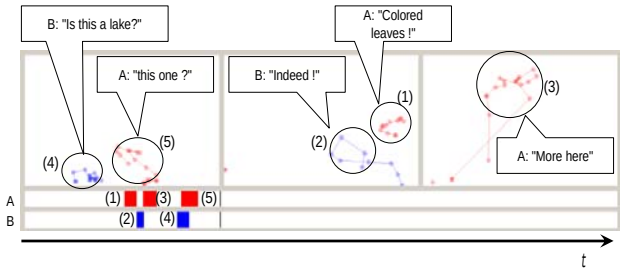


Figure 14: Pointing gestures and conversation discourse observed in the setting shown in Figure 13 [14].

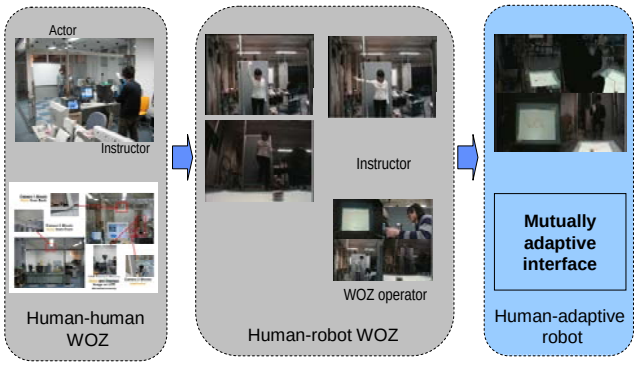


Figure 15: Three stage approach to mutual adaptation [28].

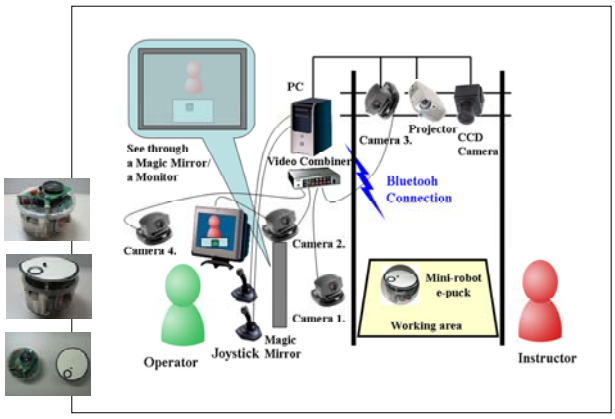


Figure 16: Facilities for measuring mutual adaptation in a human-robot WOZ [28].

speaker to gaze at the current speaker toward the end of his turn, but also a tendency for the other listener, who keeps silent during the next turn, to gaze at the next speaker around the end of the current turn and before the next speaker starts speaking." Ueda and Ohmoto prototyped a real-time system that can discriminate lies by measuring gaze directions and facial feature points [22]. The system can measure gaze directions and facial feature points, while allowing the user to move head position and orientation during the measurement and without requesting the user to place one or more markers on the face or preparing a face model in advance. Nagaoka et al [23] made a comprehensive survey of embodied synchrony (phenomenon of synchronization or similarity of nonverbal behaviors among participants) reported in diverse literature. The embodied synchrony manifests as body movement and gestures, facial behavior, vocal behavior, or physiological reactions. They surveyed the measurement and quantification techniques that have been employed in previous studies. They also attempted to attribute the embodied synchrony to interpersonal relations. Rutkowski and Mandic [24] addressed characterization of what may be called a communication atmosphere. They proposed the communication atmosphere space consisting of three dimensions: environmental, communicative, and emotional. They used audio-visual signal tracking to show the measurement for a handful of example data.

Mohammad and Nishida study human-robot communication of intentions using nonverbal behaviors. Early results include the use of interactive perception to establish and maintain joint intention [25] and a social robot that can express its internal state and intention to humans in a natural way using nonverbal feedback [26].

Mutual adaptation is a phenomenon we believe to exist between multiple learning agents being adapting with each other. Xu et al [27, 28] study mutual adaptation by taking a three stage approach consisting of a human-human WOZ experiment, a human-robot WOZ experiment, and a human-adaptive robot experiment (Figure 15). Instead of directly diving into the third stage, we observe in detail how

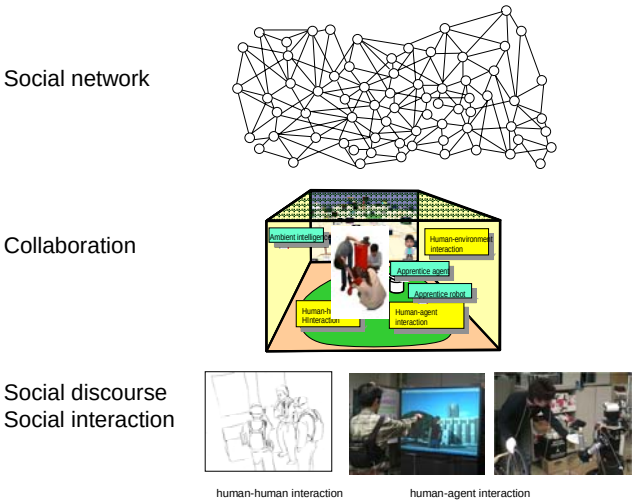


Figure 17: Three layer mode of social intelligence.

people adapt with each other and how people improve the protocols for interacting with robots. Figure 16 shows the experimental environment we developed for measuring mutual adaptation in a human-robot WOZ. We use a small mobile robot controlled by a hidden operator as if the robot was autonomous. By observing how the instructor interacts with the operator, we try to solicit the detailed observation of mutual adaptation.

VII. Social Intelligence Design

Social Intelligence Design [29] aimed at the understanding and augmentation of social intelligence for collective problem solving and learning. Social intelligence may manifest at the three levels (Figure 17). The base level comprises quick interactions at the milliseconds order where social intelligence is used to establish basic communications. The medium level encompasses a collaboration or negotiation in a small group to coordinate joint actions. The top level manifests at the community level to integrate individual intelligences into a

collective one. Conversational Informatics discussed in this article is most relevant to the social discourse level. The upper levels may have closer relationship with Web Intelligence. In order to realize Human-Centered Web Intelligence, we need to study how the layers interact with each other.

REFERENCES

[1] N. Zhong, J. Liu, Y.Y. Yao, "Envisioning Intelligent Information Technologies (iIT) from the Stand-Point of Web Intelligence (WI)," *Communications of the ACM*, 50(3), 2007, 89-94.

[2] T. Nishida, *Conversational Informatics: an Engineering Approach*. London: John Wiley & Sons Ltd, in press.

[3] T. Nishida, "Conversation Quantization for Conversational Knowledge Process," S. Bhalla (Ed.): *DNIS 2005*, LNCS 3433, Springer, 2005, pp. 15 – 33.

[4] S. Sumi, M. Bono, H. Kijima, and T. Nishida, "The IMADE (real world Interaction Measurement, Analysis and Design Environment)," unpublished.

[5] H. Kijima, M. Bono, Y. Sumi, and T. Nishida, "Development of the Environment for Multimodal Interaction Analysis," to be presented at SIG-HCI, IPSJ, Japan, 2007.

[6] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), *Embodied Conversational Agents*, MIT Press, 2000.

[7] H. Prendinger and M. Ishizuka M. (eds.), *Life-Like Characters: Tools, Affective Functions, and Applications*, Springer, 2004.

[8] H. Huang, A. Cerekovic, I. Pandzic, Y. Nakano, and T. Nishida, "Scripting Human-Agent Interactions in a Generic ECA Framework," to be presented at AI-2007, 2007, Cambridge, UK.

[9] H. Huang, T. Inoue, A. Cerekovic, I. Pandzic, Y. Nakano, and T. Nishida, "A Quiz Game Console Based on a Generic Embodied Conversational Agent Framework," *Proceedings of IVA 2007*, pp. 383-384.

[10] T. Nishida, K. Terada, T. Tajima, M. Hatakeyama, Y. Ogasawara, Y. Sumi, Y. Xu, Y. Mohammad, K. Tarasenko, T. Ohya, and T. Hiramatsu: Towards Robots as an Embodied Knowledge Medium, Invited Paper, Special Section on Human Communication II, *IEICE TRANSACTIONS on Information and Systems*, Vol. E89-D, No. 6, 2006, pp. 1768-1780.

[11] K. Kubota, Y. Sumi, and T. Nishida, "Conversation Quantization and Sustainable Knowledge Globe," Chapter 10 of [2].

[12] S. Kurohashi, D. Kawahara, N. Kaji, and T. Shibata, "Automatic Text Presentation for the Conversational Knowledge Process," Chapter 11 of [2].

[13] Y. Nakamura, "Video Content Acquisition and Editing for Conversation Scenes," Chapter 12 of [2].

[14] G. Okamura, H. Kubota, Y. Sumi, T. Nishida, H. Tsukahara, and H. Iwasaki, "Quantization and Reuse of Driving Conversations," Journal of IPSJ (in Japanese, to appear).

[15] L. Merckel, and T. Nishida, "Solution of the Perspective-Three-Point Problem," in *Proceedings IEA/AIE 2007*, pp. 324-333.

[16] H. Kubota, M. Takahashi, K. Satoh, Y. Kawaguchi, S. Nomura, Y. Sumi, and T. Nishida, "Conversation Quantization for Informal Information Circulation in a Community, The Fourth International Workshop on Social Intelligence Design (SID 2005), Stanford, USA.

[17] K. Saito, H. Kubota, Y. Sumi, and T. Nishida, "Analysis of Conversation Quanta for Conversational Knowledge Circulations," *Journal of Universal Computer Science*, vol. 13, no. 2, 2007, pp. 177-185.

[18] Y. Sumi, K. Mase, and T. Nishida, "Conversational Content Acquisition by Ubiquitous Sensors," Chapter 14 of [2].

[19] R. Taniguchi and D. Arita, "Real-time Human Proxy," Chapter 15 of [2].

[20] S. Nishiguchi, K. Kakusho, and M. Minoh, "Lecture Archiving System" Chapter 16 of [2].

[21] Y. Den and M. Enomoto, "A Scientific Approach to Conversational Informatics: Description, Analysis, and Modeling of Human Conversation," Chapter 17 of [2].

[22] Y. Ohmoto, K. Ueda, and T. Ohno, "Real-time system for measuring gaze direction and facial features: Towards automatic discrimination of lies using diverse nonverbal information," *AI & Society*, (online first).

[23] C. Nagaoka, M. Komori, and S. Yoshikawa, "Embodied Synchrony in Conversation," Chapter 18 of [2].

[24] T. Rutkowski and D. Mandic, "Modeling Communication Atmosphere," Chapter 19 of [2].

[25] Y. Mohammad and T. Nishida, "NaturalDraw: interactive perception based drawing for everyone," *Intelligent User Interfaces* 2007, pp. 251-260

[26] Y. Mohammad and T. Nishida, "TalkBack: Feedback From a Miniature Robot," to be presented at Australian AI Conference.

[27] Y. Xu, K. Ueda, T. Komatsu, T. Okadome, T. Hattori, Y. Sumi, and T. Nishida, "WOZ Experiments for Understanding Mutual Adaptation," AI&Society, online first.

[28] Y. Xu, M. Guillemot and T.iNishida, "An experiment study of gesture-based human-robot interface," IEEE/ICME International Conference on Complex Medical Engineering-CME2007, Beijing, China, 2007, pp. 458-464.

[29] T. Nishida, "Social Intelligence Design and Human Computing," *Artificial Intelligence for Human Computing 2007*, pp. 190-214.

Toyoaki Nishida (M'01) was born in 1954 in Kyoto, Japan. He is a professor of Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He received the Doctor of Engineering degree from Kyoto University in 1984. He received the B.E., the M.E., and the Doctor of Engineering degrees from Kyoto University in 1977, 1979, and 1984, respectively.

His research centers on artificial intelligence and human computer interaction. His current research focuses on social intelligence design and communicative intelligence. In 2001, he founded a series of international workshops on social intelligence design (see http://www.ii.ist.i.kyoto-u.ac.jp/sid/ for more details). Then, he broadened the scope of research to include understanding and augmenting conversational communication, and opened up a new field of research called Conversational Informatics. Currently, he leads several projects on social intelligence design and conversational informatics.

Prof. Nishida is a member of the board of directors of IPS (Information Processing Society) of Japan and JSAI (Japanese Society for Artificial Intelligence). He serves as an editorial board member of several academic journals, including Web Intelligence and Agent Systems, AI & Society, and Journal of JSAI (editor-in-chief).

Fuzzy Domain Ontology Discovery for Business Knowledge Management

Raymond Y.K. Lau
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon
Hong Kong
E-mail: raylau@cityu.edu.hk

Abstract

Ontology plays an essential role in the formalization of business information (e.g., products, services, relationships of businesses) for effective human-computer interactions. However, engineering of domain ontologies turns out to be very labor intensive and time consuming. Recently, some machine learning methods have been proposed for automatic discovery of domain ontologies. Nevertheless, the accuracy and computational ef ciency of the existing methods need to be improved to support large scale ontology construction for real-world business applications. This paper illustrates a novel fuzzy domain ontology discovery algorithm for supporting real-world business ontology engineering. By combining lexico-syntactic and statistical learning methods, the accuracy and the computational ef ciency of the ontology discovery process is improved. Empirical studies have con rme d that the proposed method can discover high quality fuzzy domain ontology which leads to signi cant improvement in information retrieval performance.

Keywords: *Domain Ontology, Fuzzy Sets, Text Mining, Information Retrieval, Knowledge Management.*

1 Introduction

Knowledge has been recognized as the most important corporate asset and it is the key for organizations to achieve sustainable competitive advantage. Knowledge management is a collection of processes that govern the creation, dissemination, and utilization of knowledge [25, 26]. To be able to effectively manage the intellectual capital, businesses need an effective approach to identify and capture information and knowledge about business processes, products, services, markets, customers, suppliers, and competitors, and to share this knowledge to improve the organizations' goal achievement. Ontologies allow domain knowledge such as products, services, markets, etc. to be captured in an explicit and formal way such that it can be shared among human and computer systems.

The notion of ontology is becoming very useful in various elds such as intelligent information extraction and retrieval, cooperative information systems, electronic commerce, and knowledge management [38]. Since Tim Berners-Lee, the inventor of the World Wide Web (Web), coined the vision of a Semantic Web [3], the proliferation of ontologies has been under tremendous growth. The success of Semantic Web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding [19]. Although there is not a universal consensus on the de niti on of ontology, it is generally accepted that ontology is a speci cat ion of conceptualization [9]. Ontology can take the simple form of a taxonomy (i.e., knowledge encoded in a minimal hierarchical structure) or as a vocabulary with standardized machine interpretable terminology supplemented with natural language de nitions. Ontology provides a number of potential bene ts in representing and processing knowledge, including the separation of domain knowledge from application knowledge, sharing of common knowledge of subjects among human and computers, and the reuse of domain knowledge for a variety of applications. Ontology is often speci ed in a declarative form by using semantic markup languages such as RDF and OWL [6]. igure 1 shows an example of the domain ontology extracted from the Reuters RCV1 corpus [16] and Figure 2 depicts the corresponding OWL statements.

Domain ontologies specify the knowledge for a particular type of domain [7]. This kind of ontologies generalize over application tasks in such domains such as medical, tourism, banking, nance, etc. A well-known example is the Uni ed Medical Language System (UMLS) and its component parts such as the Medical Subject Heading (MeSH). Although domain ontologies are useful in many

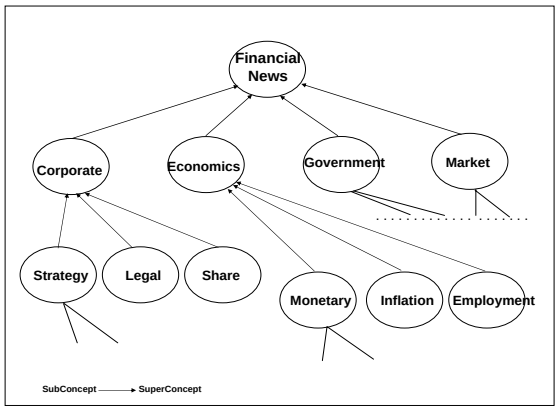


Figure 1. A Crisp Domain Ontology from the RCV-1 Corpus

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns\#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema\#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema\#"
xmlns:owl="http://www.w3.org/2002/07/owl\#"
xmlns="http://www.owl-ontologies.com/unnamed.owl\#"
xml:base="http://www.owl-ontologies.com/unnamed.owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="FinancialNews"/>
  <owl:Class rdf:ID="Corporate">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Economics">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Government">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Market">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Strategy">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#Corporate"/>
    </rdfs:subClassOf>
  </owl:Class>
  .....
</rdf:RDF>
```

Figure 2. OWL for the Financial News Ontology

areas, engineering of these ontologies turns out to be very labor intensive and time consuming. Therefore, many automatic or semi-automatic ontology engineering techniques have been proposed. Although fully automatic construction of perfect domain ontology is beyond the current state-of-the-art, we believe that the automatic ontology mining method illustrated in this paper can assist ontology engineers to build domain ontology quicker and more accurately.

Although some learning techniques have been applied to the extraction of domain ontology [4, 7, 31], these methods are still subject to further enhancement in terms of computational ef cienc y and accuracy. One of the ways to improve automated domain ontology discovery is to exploit contextual information from the knowledge sources. As domain ontology captures domain (context) dependent information, an effective discovery method should exploit contextual information in order to build relevant ontologies. On the other hand, since the taxonomy relations discovered from a text mining method often involve uncertainty, an uncertainty management mechanism is required to address such an issue. The notions of Fuzzy set and Fuzzy Relation are effective to represent knowledge with uncertainty [42]. Therefore, a fuzzy ontology rather than a crisp ontology is discovered by the proposed text mining method.

De nition 1 (Fuzzy Set) *A fuzzy set $\mathcal{F}$ consists of a set of objects drawn from a domain X and the membership of each object x_i in $\mathcal{F}$ is de ned by a membership function $\mu_{\mathcal{F}} : X \mapsto [0, 1]$. If Y is a crisp set, $\varphi(Y)$ denotes a fuzzy set generated from the traditional set of items Y.*

De nition 2 (Fuzzy Relation) *A fuzzy relation is de ned as the fuzzy set $\mathcal{G}$ on a domain $X \times Y$ where X and Y are two crisp sets.*

Figure 3 highlights the fuzzy domain ontology corresponding to the one depicted in Figure 2. The current OWL syntax can easily be extended to represent fuzzy domain ontology using approach similar to [8]. However, we will only focus on the mining of fuzzy concepts and fuzzy taxonomy relations in this paper. The term $\mu_{C\times C}(c_2, c_1)$ in Figure 3 denotes the membership value of the taxonomy relation from subclass c_2 to superclass c_1. From the text mining perspective, a keyword is an object and it belongs to different concepts (a linguistic class) with various memberships. The subsumption relations among linguistic concepts are often uncertain and are characterized by the appropriate fuzzy relations.

De nition 3 (Fuzzy Ontology) *A fuzzy ontology is a quadruple $Ont =< X, C, R_{XC}, R_{CC} >$, where X is a set of objects and C is a set of concepts. The fuzzy relation $R_{XC} : X \times C \mapsto [0, 1]$ maps the set of objects to the set of concepts by assigning the respective membership values, and the fuzzy relation $R_{CC} : C \times C \mapsto [0, 1]$ denotes the fuzzy taxonomy relations among the set of concepts C.*

The main contribution of our research work presented in this paper is the development of a novel fuzzy domain ontology discovery method which exploits contextual information embedded in textual databases (e.g., product description databases). By combining lexico-syntactic and statis-

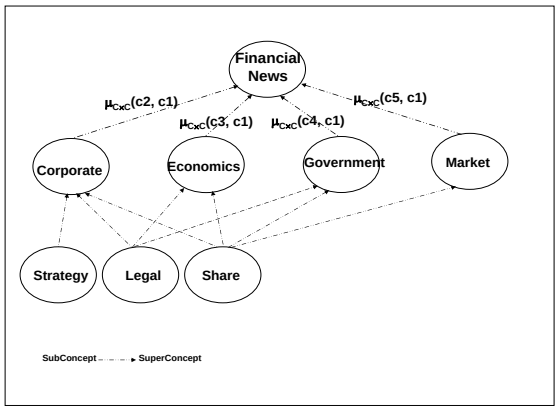


Figure 3. A Fuzzy Domain Ontology from the RCV-1 Corpus

tical learning approaches, the accuracy and the computational ef cienc y of the ontology discovery process is improved [20]. The remainder of the paper is organized as follows. Section 2 highlights previous research in the related area and compare these research work with ours. Section 3 gives an overview of our text mining methodology. The cognitive and linguistic foundations of the proposed context-sensitive ontology discovery method is described in Section 4. The computational details of the proposed ontology mining method are then illustrated in Section 5. Section 6 reports the empirical testing of our fuzzy domain ontology mining method. Finally, we offer concluding remarks and describe future direction of our research work.

2 Related Research

With the increasing importance of product information management in eCommerce environment, it is vital that precise de nition of product and services readily available in sharable, manageable, e xible, and scalable form, that is in the form of an ontology. Although the idea of utilizing ontology for e-Catalogs has been proposed long ago, an operational product ontology system for a speci c domain is not yet available. Lee et. al. [15] developed an operational product ontology system called KOCIS for the government procurement service. It consists of the ontology construction and management sub-system to build the ontology database from the product databases and to manage the the real-time processing for update operations while maintaining a consistency of the ontology data. In addition, the ontology search sub-system can retrieves and navigates the product ontology information. The search subsystem addresses the problem of ranking keyword search results by modeling the product ontology as a Bayesian belief network. Although, the KOCIS system addresses the operational aspects of an ontology management and search system, it does not support automated or semi-automated discovery of product ontology from information sources. This paper focuses on the development of a fuzzy ontology discovery algorithm for automatic business knowledge management; the proposed method can be readily applied to discover product ontology for eCommerce.

Cimiano et al. have presented an automatic taxonomy learning algorithm to extract concept hierarchies from a text corpus [5]. In particular, their taxonomy learning method is based on formal concept analysis [40]. Formal concept analysis is a systematic method for deriving implicit relationships among objects described by a set of attributes. Formal concept analysis can be seen as a conceptual clustering techniques at it provides intensional descriptions for the abstract concepts. Central to formal concept analysis is the notion of a context which is essentially the prominent attributes or features common to a set of objects of the same class. A formal context is a triple $K = (G, M, I)$ where G and M represent a set of objects and attributes respectively and I is a binary relation between G and M. Thereby, a formal concept (A, B) is de ned by $A = \{g \in G | \forall_{m \in M}(g, m) \in I\}$ and $B = \{m \in M | \forall_{g \in G}(g, m) \in I\}$. In order to derive attributes from a certain corpus, part-of-speech tagging and linguistic analysis are performed to extract verb/prepositional phrase complement, verb/object and verb/subject dependencies. For each noun appearing as head of the extracted syntactic structures, the corresponding verbs are taken as the attributes for building the formal context. Their approach is evaluated by comparing the automatically generated concept hierarchies with hand-crafted taxonomies in a tourism and a nance domain. The fuzzy ontology discovery method illustrated in this paper employs a novel subsumption based mechanism rather than the formal concept analysis approach to generate concept lattice. Semantically richer context vectors are used to represent concepts in our approach as opposed to the simple verb-based features employed by formal concept analysis. In addition, our concept hierarchy represents a fuzzy taxonomy of relations rather than a crisp taxonomy as proposed in [5].

The FOGA framework for fuzzy ontology generation has been proposed [37]. The FOGA framework consists of fuzzy formal concept analysis, fuzzy conceptual clustering, fuzzy ontology generation, and semantic representation conversion. Essentially, the FOGA method extends the formal concept analysis approach, which has also been applied to ontology extraction, with the notions of fuzzy sets. The notions of formal context and formal concept have been fuzzi ed by introducing the respective membership functions. In addition, an approximate reasoning method is developed so that the automatically generated fuzzy ontology can be incrementally furnished with the arrival of new in-

stances. The FOGA framework is evaluated in a small citation database. Our method discussed in this paper differs from the FOGA framework in that a more compact representation of fuzzy ontology is developed. The proposed method is based on previous work in computational linguistic and with the computational mechanism built on the concept of fuzzy relations. We believe that the proposed method is computationally more ef cient and be able to scale up for huge textual databases which typically consists of millions of records and thousands of terms. Finally, our proposed method is validated in a standard benchmark textual database which is considerably larger than the citation database used in [37].

A fuzzy ontology which is an extension of the domain ontology with crisp concepts is utilized for news summarization purpose [14]. In this semi-automatic ontology discovery approach, the domain ontology with various events of news is pre-de ned by domain experts. A document preprocessing mechanism will generate the meaningful terms based on the news corpus and a Chinese news dictionary pre-de ned by the domain experts. The meaningful terms are classi ed according to the events of the news by a term classi er . Basically, every fuzzy concept has a set of membership degrees associated with the various events of the domain ontology. The main function of the fuzzy inference mechanism is to generate the membership degrees (classi- cation) for each event with respect to the fuzzy concepts de ned in the fuzzy ontology. The standard triangular membership function is used for the classi cation purpose. The method discussed in this paper is a fully automatic fuzzy domain ontology discovery approach. There is no pre-de ned fuzzy concepts and taxonomy of concepts, instead our text mining method will automatically discover such concepts and generate the taxonomy relations. In addition, there is no need to set the arti c ial threshold values for the triangular membership function, instead our membership function can automatically derive the membership values based on the lexico-syntactic and statistical features of the terms observed in a textual database.

An ontology mining technique is proposed to extract patterns representing users' information needs [17]. The ontology mining method consists of two parts: the top backbone and the base backbone. The former represents the relations between compound classes of the ontology. The latter indicates the linkage between primitive classes and compound classes. The Dempster-Shafer theory of evidence model is adopted to model the relations among classes. The presented method can effectively synthesizing taxonomic relation and non-taxonomic relation in a single ontology model. In addition, a novel method is proposed to capture the evolving patterns in order to re ne the discovered ontology. Finally, a formal model is developed to assess the relevance of the discovered ontology with respect to the user's information needs. The ontology mining method is validated based on the Reuters RCV-1 benchmark collection. The research work presented in this paper focuses on fuzzy domain ontology discovery rather than the discovery of crisp ontology representing users' information needs.

Personalized Abstract Search Services (PASS) is a domain speci c search engine providing abstracts of papers from IEEE Transactions sponsored by the IEEE Neural Network Council [39]. The system uses a fuzzy ontology of term associations to support semantic based information retrieval. The fuzzy ontology is automatically built using information obtained from the system's document collection. The system extracts a set of two or three consecutive words exhibiting some linguistic patterns such as "noun noun", "adj noun", etc. from a corpus. The system then eliminates the phrases that contain at least one stop word from a pre-de ned control le. The notions of narrower and broader term relations are introduced and a fuzzy conjunction operator is applied to compute the membership values of the term relations. By evaluating the users' searching activities, it was found that the fuzzy ontology of term relations significantly contributes to the information retrieval process. Our work presented in this paper differs from the PASS system in the fuzzy concepts (instead of terms) are rst identi ed and the taxonomy relations of concepts are then developed. In addition, our fuzzy ontology mining approach has been evaluated based on a bench-mark collection in the eld of information retrieval.

An ontology based text mining system that extracts fuzzy relations from biological texts is present [1]. This approach preserves the basic structured knowledge format for storing domain knowledge, but allows for update of information at the same time. The document processor parses the text documents and removes the tags pertaining to the biological domain. The strength of association between a tag pair E_i and E_j representing two biological entities is computed according to a fuzzy conjunction operator. Basically, the membership values of the relations are functions of frequency of co-occurrence of concepts. The fuzzy relations between the biological terms are used to guide information retrieval from a medical document collection called GENIA. The ontology discovery method presented in this paper deals with general textual databases rather than speci cally tagged biological documents. Concept extraction in our approach is based on the lexico-syntactic characteristic of tokens appearing in a corpus rather than the pre-de ned semantic of speci c biological tags.

A semiautomatic ontology engineering environment called OntoEdit has been developed [19, 20]. The workbench supports ontology import, extraction, pruning, re nement, and evaluation. Merging existing semantic structures or de ning mapping rules between these structures allows importing and reusing available ontologies. Ontology ex-

traction is one of the main tasks of ontology engineering, which deals with learning the appropriate ontologies from the domain sources. The initial ontology which results from import, reuse, and extraction, is then pruned to better t the purpose of the particular application. Traditional text processing techniques such as n-gram [30] is used to extend the set of lexical entries L based on source documents. Hierarchical clustering is applied to learn the taxonomy relations H_C. In addition, morphological analysis and generalized association rule mining are applied to learn the relations R among some concepts C. Our work presented in this paper focuses on the ontology extraction stage of the ontology engineering cycle. Moreover, a subsumption-based computational method rather than the traditional clustering method is used for the extraction of concept lattice.

3 An Overview of the Text Mining Methodology

Figure 4 depicts the proposed text mining methodology for the automatic discovery of fuzzy domain ontology from a textual database (corpus). A text corpus is parsed to analyze the lexico-syntactic elements. For instance, stop words such as "a, an, the" are removed from the source documents since these words appear in any contexts and they cannot provide useful information to describe a domain concept. For our implementation, a stop word le is constructed based on the standard stop word le used in the SMART retrieval system [29]. Lexical pattern is identi ed by applying Part-of-Speech (POS) tagging to the source documents and then followed by token stemming based on the Porter stemming algorithm [28]. We refer to the WordNet lexicon [21] to tag each word during this process. During the linguistic pattern ltering stage, certain linguistic patterns are extracted based on the speci c requirements speci ed by the ontology engineers. For example, the ontology engineers may only focus on the "Noun Noun" and "Adjective Noun" patterns instead of all the linguistic patterns. This is in fact a good way to gain computational ef cienc y by reducing the number of patterns for further statistical analysis. In addition, to extract relevant domain speci c concepts, the appearances of concepts across different domains should be taken into account. The basic intuition is that a concept frequently appears in a speci c domain (corpus) rather than many different domains is more likely to be a relevant domain concept. The statistical Token Analysis step employs the information theoretic measure to compute the co-occurrence statistics of the targeting linguistic patterns. Finally, taxonomy of domain concepts is developed according to the fuzzy conjunction operator. The details of the proposed ontology mining method will be discussed in Section 5.

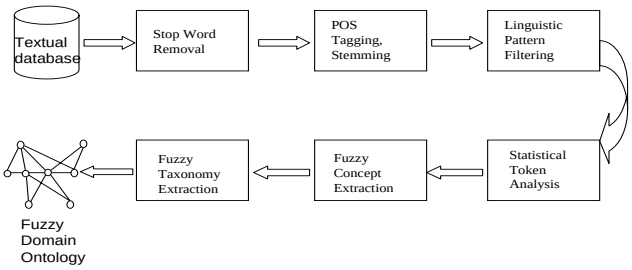


Figure 4. Context-Sensitive Fuzzy Ontology Discovery Process

4 The Linguistic Foundations

The proposed context-sensitive fuzzy ontology discovery method is based on the *distributional hypothesis* which assumes that terms (concepts) are similar according to the extent that they share similar linguistic contexts [10]. In particular, we borrow the notion of *collocational expressions* from computational linguistic to identify the semantics of some lexical elements such as concepts from text corpora. For computational linguistic, a term refers to one or more tokens (words) and a term is also a concept if it carries recognizable meaning speci c to a domain [23]. Collocational expressions are groups of words related in meaning, and the constituent words of an expression are frequently found in a near loci of a few adjacent words in a textual unit [33, 35]. The collocational expressions are indeed providing the underlying context of a given concept embedded in natural language text such as Web documents.

Contextual information has long been recognized as one of the major contributors to concept learning in the eld of computer science [43]. Nevertheless, to automatically detect the semantics (meanings) of a concept is not a trivial task since the meanings of a concept is context (domain) dependent. For example, the concept "bank" can refer to a nancial institute such as a "commercial bank, or refer to the raised shelf of ground such as the "river bank". Therefore, to accurately extract domain ontologies from text, contextual information must be exploited to disambiguate different senses. In this regard, *static* lexicons (i.e., generic linguistic ontologies) such as WordNet [21] with meanings (senses) computed *a priori* may not be able to capture the speci c semantics of concepts pertaining to a particular application domain. However, WordNet can be used to boot-

strap the performance of information extraction when domain ontologies are built [22, 24]. Our general approach is that the collocational expressions are rst extracted from the source documents; these collocational expressions which carry context-sensitive semantics are then used to de ne the meanings of the concepts.

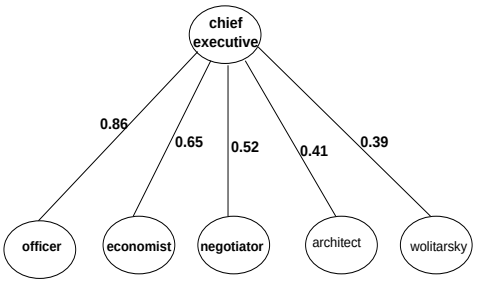


Figure 5. Domain Specific Semantics of the Concept "Chief Executive"

In the eld of information retrieval (IR), the notion of *context vectors* [11, 32] has been proposed to give computer-based representations of concepts. In this approach, a concept is represented by a vector of words and their numerical weights. The weight of a word indicates the extent to which the particular word is *associated* with the underlying concept. For example, the concept "chief executive" is represented by the words such as of cer , negotiator, economist, etc. as depicted in Figure 5, which is an interesting example by parsing the Reuters-21578 corpus (http://www.daviddlewis.com/resources/testcollections/). The context vector of "chief executive" is shown as follows:

Concept: chief executive
Context Vector:
$\{(officer, 0.72), (economist, 0.65), (negotiator, 0.63), (architect, 0.61), (wolitarsky, 0.44)\}$

The context vector can be seen as a point in a multi-dimensional geometric information space with each dimension representing a property term. It should be noted that the meanings (senses) of "chief executive" is "head of state" or "presidency" as de ned in WordNet [21], which is quite different from that discovered by our context-sensitive text mining method. The last term in the example context vector is "wolitarsky" which is the name of the chief executive of a nancial institution often mentioned in the Reuters nancial news in that period. So, our method can really discover domain speci c relation such as "wolitarsky" is a chief executive. Static lexicons such as WordNet can only capture the lexical knowledge of a concept, but fails to represent domain speci c non-lexical knowledge. A linguistic concept such as "chief executive" can be taken as a class (set) with respect to the fuzzy set framework. A term such as "wolitarsky" will then be treated as an object which belongs to the set with certain degree.

5 Text Mining for Fuzzy Ontology Discovery

It is believed that the main challenge in mining taxonomy relations from textual databases is to lter out the noisy relations[18, 20]. Accordingly, our text mining method is speci cally designed to deal with such an issue. After standard document pre-processing such as stop word removal, POS tagging, and word stemming [30], a *windowing process* is conducted over the collection of documents. The windowing process can help reduce the number of noisy term relationships. For each document (e.g., Net news, Web page, email, etc.), a *virtual window* of δ words is moved from left to right one word at a time until the end of a textual unit (e.g., a sentence) is reached. Within each window, the statistical information among tokens is collected to develop collocational expressions. Such a windowing process has successfully been applied to text mining before [13]. The windowing process is repeated for each document until the entire collection has been processed. According to previous studies, a text window of 5 to 10 terms is effective [11, 27], and so we adopt this range as the basis to perform our windowing process. To improve computational ef cienc y and lte r noisy relations, only the speci c linguistic pattern (e.g., Noun Noun, and Adjective Noun) de ned by an ontology engineer will be analyzed. The following is an example segment of a news article in the Reuters-21578 collection:

```
<REUTERS OLDID="5545" NEWID="2"><TEXT>
<TITLE>STANDARD OIL TO FORM FINANCIAL
UNIT</TITLE>
<BODY>Standard Oil Co and BP North
America Inc said they plan to form
a venture to manage the money market
borrowing and investment activities
of both companies.
</BODY></TEXT> </REUTERS>
```

After parsing the main body of the news article, our ontology extraction program will remove the stop words, apply POS tagging and stem the words. So, the result will look like:

```
standard (Adj) oil (N) co (N)
```

```
bp (N) north (Adj) america (N)
inc (N) said (V) plan (V) form (V)
venture (N) manage (V) money (N)
market (N) borrow (V) investment (N)
activit (N) compan (N) .
```

Assuming that the window size of 5 is used and the ontology engineer speci es the "Noun Noun" linguistic pattern as the only focus, the potential concepts "Oil Co" and "Co BP" will be extracted from the rst virtual text window. The concept "Oil Co" might be represented by the features such as "standard", "bp", and "north". After parsing the whole corpus, the statistical data (by statistical token analysis) about the potential concepts can be collected. If a word has an association weight lower than a pre-de ned threshold value, it will be discarded from the context vector of the concept. This is equivalent to the α-cut operation for fuzzy sets.

For statistical token analysis, several information theoretic methods are employed. Mutual Information has been applied to collocational analysis [27, 36] in previous research. Mutual Information is an information theoretic method to compute the dependency between two entities and is de ned by [34]:

$$MI(t_i, t_j) = \log_2 \frac{Pr(t_i, t_j)}{Pr(t_i)Pr(t_j)} \quad (1)$$

where $MI(t_i, t_j)$ is the mutual information between term t_i and term t_j. $Pr(t_i, t_j)$ is the joint probability that both terms appear in a text window, and $Pr(t_i)$ is the probability that a term t_i appears in a text window. The probability $Pr(t_i)$ is estimated based on $\frac{|w_t|}{|w|}$ where $|w_t|$ is the number of windows containing the term t and $|w|$ is the total number of windows constructed from a textual database (i.e., a collection). Similarly, $Pr(t_i, t_j)$ is the fraction of the number of windows containing both terms out of the total number of windows.

We develop *Balanced Mutual Information* (BMI) to compute the degree of association among tokens. This method considers both term presence and term absence as the evidence of the implicit term relationships.

$$\begin{aligned}\mu_{c_i}(t_j) &\approx BMI(t_i, t_j) \\ &= \beta(Pr(t_i, t_j)\log_2(\tfrac{Pr(t_i,t_j)}{Pr(t_i)Pr(t_j)})+ \\ &\quad Pr(\neg t_i, \neg t_j)\log_2(\tfrac{Pr(\neg t_i,\neg t_j)}{Pr(\neg t_i)Pr(\neg t_j)})) - \\ &\quad (1-\beta)(Pr(t_i, \neg t_j)\log_2(\tfrac{Pr(t_i,\neg t_j)}{Pr(t_i)Pr(\neg t_j)})+ \\ &\quad Pr(\neg t_i, t_j)\log_2(\tfrac{Pr(\neg t_i,t_j)}{Pr(\neg t_i)Pr(t_j)}))\end{aligned} \quad (2)$$

where $\mu_{c_i}(t_j)$ is the membership function to estimate the degree of a term $t_j \in X$ belonging to a concept $c_i \in C$. $\mu_{c_i}(t_j)$ is the computational mechanism for the relation R_{XC} de ned in the fuzzy ontology $Ont =< X, C, R_{XC}, R_{CC} >$. The membership function $\mu_{c_i}(t_j)$ is

Algorithm FuzzyOntoMine(*D, Para, Ont*)
Input: corpus D and vector of threshold values $Para$
Output: a fuzzy domain ontology Ont
Main Procedure:

1. $Ont = \{\}$
2. Foreach document $d \in D$ Do
   (a) Construct text windows $w \in d$
   (b) Remove stop words sw from w
   (c) Perform POS tagging for each term $t_i \in w$
   (d) Apply Porter stemming to each term t_i
   (e) Accumulate the frequency for $t_i \in w$ and the joint frequency for any pair $t_i, t_j \in w$
   (f) IF $lower \leq Feq(t_i) \leq upper$, $X = X \cup t_i$
3. End for
4. Foreach term $t_i \in X$ Do
   (a) compute its context vector c_i using BMI, MI, JA, CP, KL, or ECH
   (b) $C = C \cup c_i$
5. End for
6. Foreach $c_i \in C$ Do /* Concept Pruning - α-cut */
   (a) IF $\forall t_i \in c_i : \mu_{c_i}(t_i) < \alpha$
   (b) THEN $C = C - c_i$
7. End for
8. Foreach pair of concepts $c_i, c_j \in C$ Do
   (a) Compute the taxonomy relation $R(c_i, c_j)$ using $Spec(c_i, c_j)$
   (b) IF $\mu_{C\times C}(c_i, c_j) > \lambda$, $R = R \cup R(c_i, c_j)$
9. End For
10. Foreach $R(c_i, c_j) \in R$ Do /* Taxonomy Pruning */
    (a) IF $\mu_{C\times C}(c_i, c_j) < \mu_{C\times C}(c_j, c_i)$
    (b) THEN $R = R - R(c_i, c_j)$
    (c) IF $\exists P(c_i \rightarrow c_x, \ldots, c_y \rightarrow c_j)$
    (d) AND $\mu_{C\times C}(c_i, c_j) \leq \min(\{\mu_{C\times C}(c_i, c_x), \mu_{C\times C}(c_x, c_y), \ldots, \mu_{C\times C}(c_y, c_j)\})$
    (e) THEN $R = R - R(c_i, c_j)$
11. End For
12. Output Ont

Figure 6. The Fuzzy Domain Ontology Discovery Algorithm

indeed approximated by the BMI score. $Pr(t_i, t_j)$ is the joint probability that both terms appear in a text window, and $Pr(\neg t_i, \neg t_j)$ is the joint probability that both terms are absent in a text window. The weight factor $\beta > 0.5$ is used to control the relative importance of two kinds of evidence (positive and negative). In Eq.(2), each MI value is then normalized by the corresponding joint probabilities. For the special case where $Pr(t_i, t_j) = 1$ is true, the joint probability value is replaced by a large positive integer because terms t_i, t_j have the strongest association. An α-cut is applied to discard terms from the potential concept if their membership values are below the threshold α. After computing all the BMI values in a collection, these values are subject to linear scaling such that each membership value is within the unit interval $\forall_{c_i \in C, t_j \in X} \mu_{c_i}(t_j) \in [0, 1]$. It should be noted that the constituent terms of a concept are always belonging to the concept with the maximal membership 1. Other measures that can be used to estimate the membership values of $t_j \in c_i$ include Jaccard (JA), conditional probability (CP), Kullback-Leibler divergence (KL), and Expected Cross Entropy (ECH) [12]:

$$\begin{aligned} \mu_{c_i}(t_j) &\approx Jacc(c_i, t_j) \\ &= \frac{Pr(c_i \wedge t_j)}{Pr(c_i \vee t_j)} \end{aligned} \tag{3}$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx Pr(c_i|t_j) \\ &= \frac{Pr(c_i, t_j)}{Pr(t_j)} \end{aligned} \tag{4}$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx KL(c_i||t_j) \\ &= \sum_{c_i \in C} Pr(c_i|t_j) \log_2 \frac{Pr(c_i|t_j)}{Pr(c_i)} \end{aligned} \tag{5}$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx ECH(t_j, c_i) \\ &= Pr(t_j) \sum_{c_i \in C} Pr(c_i|t_j) \log_2 \frac{Pr(c_i|t_j)}{Pr(c_i)} \end{aligned} \tag{6}$$

To further lter the noisy concept relations, only the relatively prominent concepts for a domain will be further explored. We adopt the TFIDF [30] like heuristic to lter non-relevant domain concepts. Similar approach has also been used in ontology learning [24]. For example, if a concept is signi cant for a particular domain, it will appear more frequently in that domain when compared with its appearance in other domains. The following measure is used to compute the relevance score of a concept:

$$Rel(c_i, D_j) = \frac{Dom(c_i, D_j)}{\sum_{k=1}^{n} Dom(c, D_k)} \tag{7}$$

where $Rel(c_i, D_j)$ is the relevance score of a concept c_i in the domain D_j. The term $Dom(c_i, D_j)$ is the domain frequency of the concept c_i (i.e., number of documents containing the concept divided by the total number of documents in the corpus). The higher the value of $Rel(c_i, D_j)$, the more relevant the concept is for domain D_j. Based on empirical testing, we can estimate a threshold rel for a particular domain. Only the concepts with relevance score greater than the threshold will be selected. For each selected concept, its context vector will be expanded based on the synonymy relation de ned in WordNet [21]. This is in fact a *smoothing* procedure [5]. The intuition is that some words that belong to a particular concept may not co-occur with the concept in a corpus. To make our ontology discovery method more robust, we need to consider these missing associations. For instance, our example context vector for "chief executive" will be expanded with the feature "presidency" based on the synonymy relation of WordNet, and a default membership value will be applied to such a term.

The na l stage towards our ontology discovery method is fuzzy taxonomy generation based on subsumption relations among extracted concepts. Let $Spec(c_x, c_y)$ denotes that concept c_x is a specialization (sub-class) of another concept c_y. The degree of such a specialization is derived by:

$$\begin{aligned} \mu_{C \times C}(c_x, c_y) &\approx Spec(c_x, c_y) \\ &= \frac{\sum_{t_x \in c_x, t_y \in c_y, t_x = t_y} \mu_{c_x}(t_x) \otimes \mu_{c_y}(t_y)}{\sum_{t_x \in c_x} \mu_{c_x}(t_x)} \end{aligned} \tag{8}$$

where $\otimes$ is a fuzzy conjunction operator which is equivalent to the min function. The above formula states that the degree of subsumption (speci city) of c_x to c_y is based on the ratio of the sum of the minimal membership values of the common terms belonging to the two concepts to the sum of the membership values of terms in the concept c_x. For instance, if every object of c_x is also an object of c_y, a high speci city value will be derived. The $Spec(c_x, c_y)$ function takes its values from the unit interval $[0, 1]$ and the subsumption relation is asymmetric. When the taxonomy is built, we only select the subsumption relations such that $Spec(c_x, c_y) > Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ where λ is a threshold to distinguish signi can t subsumption relations. The parameter λ is estimated based on empirical tests. If $Spec(c_x, c_y) = Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ is established, the *equivalent* relation between c_x and c_y will be extracted. In addition, a pruning step is introduced such that the redundant taxonomy relations are removed. If the membership of a relation $\mu_{C \times C}(c_1, c_2) \leq \min(\{\mu_{C \times C}(c_1, c_i), \ldots, \mu_{C \times C}(c_i, c_2)\})$, where $c_1, c_i, \ldots, c_2$ form a path P from c_1 to c_2, the relation $R_(c_1, c_2)$ is removed because it can be derived from other stronger taxonomy relations in the ontology. The fuzzy domain ontology mining algorithm is summarized and shown in Figure 6.

6 Evaluation

Since one of the most important applications of domain ontology is for intelligent information retrieval, our context-

sensitive fuzzy ontology mining method is evaluated within the context of information retrieval. Our rst experiment is similar to the routing tasks used in the Text REtrieval Conference (TREC) (http://trec.nist.gov/) which is a well-known international benchmark forum for information retrieval systems. The Reuters-21578 standard corpus with the Lewis-Split subset which contains 19,813 documents is used in our experiments. The training set consists of 13,625 documents and the test set consists of 6,188 documents. Our fuzzy domain ontology is automatically constructed based on the training set only. It takes 19 minutes only to complete the ontology mining process on a Pentium-4 2.2GHz PC. In this experiment, a window size of 5, a term size of 1, a single Noun pattern, and the (BMI) computational method with $\beta = 0.7$ are used.

For our ontology extraction method, a concept's relevance score de ned in Eq. 7 is computed with respect to a variety of domains. Therefore, several other corpora are constructed based on the Web documents retrieved under different Yahoo categories such as "computer", "entertainment", "education" etc. For the Reuters-21578 corpus, a set of queries are composed based on the pre-de ned Reuters topics and the top ve (weighted by TFIDF) terms from one relevant document of the training set. For each Reuters subject code such as "acq", the corresponding subject description such as "acquisitions or mergers" is retrieved from the Reuters-21578 category description le. Each query is then applied to the testing set and the documents are ranked with respect to their relevance to the query. The vector-space model [29] is employed in this routing task. For instance, the standard TFIDF term weighting scheme is used to compute the term weights of a document and a query respectively, and the cosine similarity measure is used to rank each document:

$$sim(\overrightarrow{q}, \overrightarrow{d}) = \frac{\sum_{i=1}^{n} w_q(k_i) \times w_d(k_i)}{\sqrt{\sum_{i=1}^{n}(w_q(k_i))^2} \times \sqrt{\sum_{i=1}^{n}(w_d(k_i))^2}} \quad (9)$$

where $\overrightarrow{q}$ and $\overrightarrow{d}$ are the query vector and the document vector respectively. The term $w_q(k_i)$ represents the weight of the ith keyword k_i in the query vector $\overrightarrow{q}$, and the term $w_d(k_i)$ represents the weight of the ith keyword k_i in the document vector $\overrightarrow{d}$.

The routing tasks are performed with (the experimental group) and without (the control group) the help of our automatically constructed fuzzy domain ontology. Basically, the domain ontology is used for query expansion [41] for the routing task. For instance, each term in the original query is expanded with respect to the domain ontology to obtain a equivalent, a broader, or a more speci c term. In this experiment, the type of relations is selected manually from the fuzzy domain ontology to optimize the retrieval effectiveness. Standard performance measures [30] such as precision, recall, and F-measure are then computed based on the top 100 documents retrieved in both groups:

$$Precision = \frac{a}{a+b} \quad (10)$$

$$Recall = \frac{a}{a+c} \quad (11)$$

$$F_\eta = \frac{(1+\eta^2) Precision \times Recall}{\eta^2 Precision + Recall} \quad (12)$$

where a, b, c represent the number of retrieved relevant documents, the number of retrieved non-relevant documents, and the number of not retrieved relevant documents respectively. The $F_{\eta=1}$ measure and the recall results of 15 randomly selected Reuters topics are depicted in Table 1. The rst column in Table 1 shows the topic names of the Reuters-21578 collection; the second column shows the number of true relevant documents for each topic. The remaining two columns are the $F_{\eta=1}$ and the recall results achieved when domain ontology is applied to expand initial query. The last two columns show the $F_{\eta=1}$ and the recall gures when domain ontology is not used for query expansion. Except for the topic of "coffee", the IR performance is improved with the help of the fuzzy domain ontology for query expansion. The reason why there is no improvement for the "coffee" topic is that the automatically generated domain ontology does not provide additional knowledge to expand the initial query. The difference of IR performance (both F-measure and Recall) between these two groups is statistically signi can t ($p < 0.01$) according to a paired one tail t-test. The average improvement of the $F_{\eta=1}$ measure is 58.3%. Therefore, we can conclude that the automatically discovered fuzzy domain ontology is with good quality and it is useful for enhancing information retrieval performance.

In our second experiment, various information theoretic measures are tested for the purpose of extracting domain concepts from a corpus. The same routing task is conducted except the use of different computational methods such as BMI, MI, JA, CP, and KL to estimate the membership of a term for a concept. The topic "carcass" is used to illustrate the typical performance of these methods. The precision-recall graph of these runs is plotted in Figure 7. The x axis indicates the various recall levels and the y axis shows the precision values obtained at the corresponding recall level. For example, the recall level 0.1 indicates the Nth position where 7 relevant documents (there are 68 relevant records for this topic) are found from the ranked list, and the corresponding precision values indicate the retrieval effectiveness of various methods (e.g., the best precision 0.36 is achieved by BMI). In general, the higher the precision curve, the better performance the information retrieval system is. As can be seen, the BMI method leads to the best performance because it can take into account both positive

indeed approximated by the BMI score. $Pr(t_i, t_j)$ is the joint probability that both terms appear in a text window, and $Pr(\neg t_i, \neg t_j)$ is the joint probability that both terms are absent in a text window. The weight factor $\beta > 0.5$ is used to control the relative importance of two kinds of evidence (positive and negative). In Eq.(2), each MI value is then normalized by the corresponding joint probabilities. For the special case where $Pr(t_i, t_j) = 1$ is true, the joint probability value is replaced by a large positive integer because terms t_i, t_j have the strongest association. An α-cut is applied to discard terms from the potential concept if their membership values are below the threshold α. After computing all the BMI values in a collection, these values are subject to linear scaling such that each membership value is within the unit interval $\forall_{c_i \in C, t_j \in X} \mu_{c_i}(t_j) \in [0, 1]$. It should be noted that the constituent terms of a concept are always belonging to the concept with the maximal membership 1. Other measures that can be used to estimate the membership values of $t_j \in c_i$ include Jaccard (JA), conditional probability (CP), Kullback-Leibler divergence (KL), and Expected Cross Entropy (ECH) [12]:

$$\begin{aligned} \mu_{c_i}(t_j) &\approx Jacc(c_i, t_j) \\ &= \frac{Pr(c_i \wedge t_j)}{Pr(c_i \vee t_j)} \end{aligned} \tag{3}$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx Pr(c_i|t_j) \\ &= \frac{Pr(c_i, t_j)}{Pr(t_j)} \end{aligned} \tag{4}$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx KL(c_i||t_j) \\ &= \sum_{c_i \in C} Pr(c_i|t_j) \log_2 \frac{Pr(c_i|t_j)}{Pr(c_i)} \end{aligned} \tag{5}$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx ECH(t_j, c_i) \\ &= Pr(t_j) \sum_{c_i \in C} Pr(c_i|t_j) \log_2 \frac{Pr(c_i|t_j)}{Pr(c_i)} \end{aligned} \tag{6}$$

To further lter the noisy concept relations, only the relatively prominent concepts for a domain will be further explored. We adopt the TFIDF [30] like heuristic to lter non-relevant domain concepts. Similar approach has also been used in ontology learning [24]. For example, if a concept is signi cant for a particular domain, it will appear more frequently in that domain when compared with its appearance in other domains. The following measure is used to compute the relevance score of a concept:

$$Rel(c_i, D_j) = \frac{Dom(c_i, D_j)}{\sum_{k=1}^{n} Dom(c, D_k)} \tag{7}$$

where $Rel(c_i, D_j)$ is the relevance score of a concept c_i in the domain D_j. The term $Dom(c_i, D_j)$ is the domain frequency of the concept c_i (i.e., number of documents containing the concept divided by the total number of documents in the corpus). The higher the value of $Rel(c_i, D_j)$, the more relevant the concept is for domain D_j. Based on empirical testing, we can estimate a threshold rel for a particular domain. Only the concepts with relevance score greater than the threshold will be selected. For each selected concept, its context vector will be expanded based on the synonymy relation de ned in WordNet [21]. This is in fact a *smoothing* procedure [5]. The intuition is that some words that belong to a particular concept may not co-occur with the concept in a corpus. To make our ontology discovery method more robust, we need to consider these missing associations. For instance, our example context vector for "chief executive" will be expanded with the feature "presidency" based on the synonymy relation of WordNet, and a default membership value will be applied to such a term.

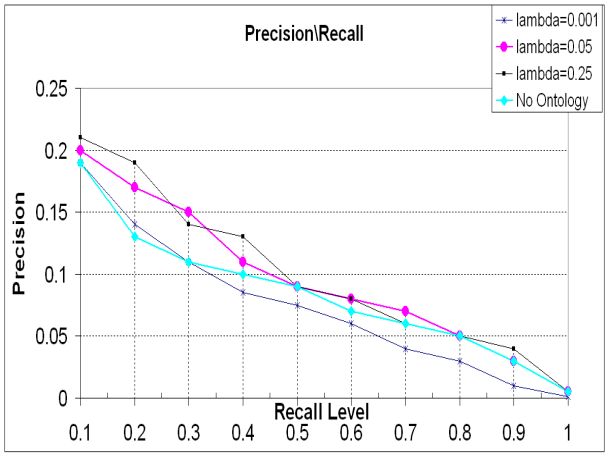
The na l stage towards our ontology discovery method is fuzzy taxonomy generation based on subsumption relations among extracted concepts. Let $Spec(c_x, c_y)$ denotes that concept c_x is a specialization (sub-class) of another concept c_y. The degree of such a specialization is derived by:

$$\begin{aligned} \mu_{C \times C}(c_x, c_y) &\approx Spec(c_x, c_y) \\ &= \frac{\sum_{t_x \in c_x, t_y \in c_y, t_x = t_y} \mu_{c_x}(t_x) \otimes \mu_{c_y}(t_y)}{\sum_{t_x \in c_x} \mu_{c_x}(t_x)} \end{aligned} \tag{8}$$

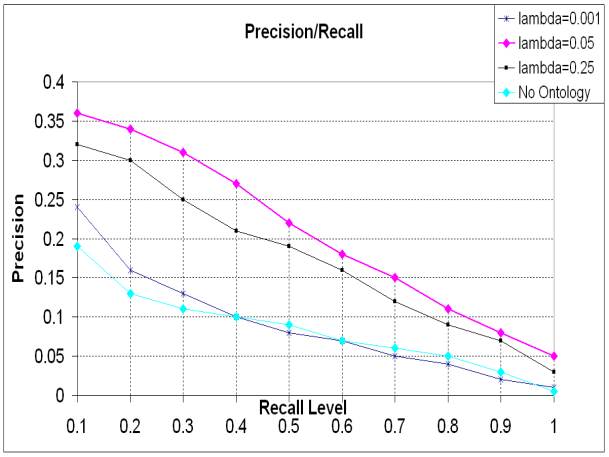
where $\otimes$ is a fuzzy conjunction operator which is equivalent to the min function. The above formula states that the degree of subsumption (speci city) of c_x to c_y is based on the ratio of the sum of the minimal membership values of the common terms belonging to the two concepts to the sum of the membership values of terms in the concept c_x. For instance, if every object of c_x is also an object of c_y, a high speci city value will be derived. The $Spec(c_x, c_y)$ function takes its values from the unit interval $[0, 1]$ and the subsumption relation is asymmetric. When the taxonomy is built, we only select the subsumption relations such that $Spec(c_x, c_y) > Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ where λ is a threshold to distinguish signi can t subsumption relations. The parameter λ is estimated based on empirical tests. If $Spec(c_x, c_y) = Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ is established, the *equivalent* relation between c_x and c_y will be extracted. In addition, a pruning step is introduced such that the redundant taxonomy relations are removed. If the membership of a relation $\mu_{C \times C}(c_1, c_2) \leq \min(\{\mu_{C \times C}(c_1, c_i), \ldots, \mu_{C \times C}(c_i, c_2)\})$, where $c_1, c_i, \ldots, c_2$ form a path P from c_1 to c_2, the relation $R_(c_1, c_2)$ is removed because it can be derived from other stronger taxonomy relations in the ontology. The fuzzy domain ontology mining algorithm is summarized and shown in Figure 6.

6 Evaluation

Since one of the most important applications of domain ontology is for intelligent information retrieval, our context-



a. Window Size 20 and windowing across sentences



b. Window Size 5 and windowing within sentences

Figure 8. The Impact of Windowing and Taxonomy Pruning

too many noisy taxonomy relations exist in the ontology which leads to poor query expansion. On the other hand, if the threshold λ is too high, many useful taxonomy relations are ltere d out such that the ontology is not useful for query re nement. When the threshold $\lambda = 0.001$ is used, a noisy ontology will be generated which leads to retrieval performance worse than the baseline where no ontology is used for query expansion. If an appropriate window size is employed and the windowing process is carried out within sentence boundary, a fuzzy domain ontology with higher quality is generated (as depicted in Figure 8.b).

7 Conclusions

The manipulation and exchange of semantically enriched business intelligence (e.g., products, services, markets, etc.) can enhance the quality of an eCommerce system and offer a high level of inter-operability among different enterprise systems. Ontology certainly plays an important role in the formalization of business knowledge. However, the biggest challenge for the wide spread applications of ontologies is on the construction of these ontologies because it is a very labor intensive and time consuming process. As uncertainty often presents in real-world applications, it is less likely that domain ontologies with crisp concepts and relations can satisfy these applications. This paper illustrates a novel fuzzy domain ontology discovery algorithm to facilitate the ontology engineering process. In particular, contextual information of a domain is exploited so that higher quality fuzzy domain ontologies can be automatically constructed. The proposed discovery method combines lexico-syntactic and statistical learning approaches so as to reduce the chance of generating noisy concepts and relations. Empirical studies have been performed to evaluate the quality of the fuzzy domain ontology discovered by the proposed ontology mining algorithm. Our preliminary results show that the automatically generated fuzzy domain ontology can signi cantly improve the effectiveness in information retrieval. Future work involves comparing the accuracy and the computational ef cienc y of our fuzzy ontology mining method with that of the other approaches. In addition, larger scale of quantitative evaluation of our fuzzy ontology mining algorithm in the context of business information management will be conducted.

References

[1] Muhammad Abulaish and Lipika Dey. Biological ontology enhancement with fuzzy relations: A text-mining framework. In Andrzej Skowron, Rakesh Agrawal, Michael Luck, Takahira Yamaguchi, Pierre Morizet-Mahoudeaux, Jiming Liu, and Ning Zhong, editors, *Proceedings of the 2005 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2005)*, pages 379–385, Compiegne, France, September 19–22 2005. IEEE Computer Society.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago de Chile, Chile, September 12–15 1994. Morgan Kaufmann Publishers.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scienti c American*, 284(5):34–43, 2001.

[4] Shan Chen, Damminda Alahakoon, and Maria Indrawan. Background knowledge driven ontology discovery. In *Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service*, pages 202–207, 2005.

[5] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Arti cial Intelligence Research*, 24:305–339, 2005.

[6] The World Wide Web Consortium. Web Ontology Language, 2004. Available from `http://www.w3.org/2004/OWL/`.

[7] Michael Dittenbach, Helmut Berger, and Dieter Merkl. Improving domain ontologies by mining semantics from text. In *Proceedings of the First Asia-Paci c Conference on Conceptual Modelling (APCCM2004)*, pages 91–100, 2004.

[8] Mingxia Gao and Chunnian Liu. Extending OWL by fuzzy description logic. In *Proceedings of the 17th IEEE International Conference on Tools with Arti cial Intelligence (ICTAI'05)*, pages 562–567. IEEE Computer Society, 2005.

[9] T. R. Gruber. A translation approach to portable ontology speci cations. *Knowledge Acquisition*, 5(2):199–220, 1993.

[10] Z. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.

[11] Hongyan Jing and Evelyne Tzoukermann. Information retrieval based on context distance and morphology. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Language Analysis, pages 90–96, 1999.

[12] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 170–178, Nashville, Tennessee, 1997. Morgan Kaufmann Publishers, San Francisco, California.

[13] R.Y.K. Lau. Context-Sensitive Text Mining and Belief Revision for Intelligent Information Retrieval on the Web. *Web Intelligence and Agent Systems An International Journal*, 1(3-4):1–22, 2003.

[14] Chang-Shing Lee, Zhi-Wei Jian, and Lin-Kai Huang. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(5):859–880, 2005.

[15] Taehee Lee, Ig hoon Lee, Suekyung Lee, Sang goo Lee, Dongkyu Kim, Jonghoon Chun, Hyunja Lee, and Junho Shim. Building an operational product ontology system. *Electronic Commerce Research and Applications*, 5(1):16–28, 2006.

[16] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[17] Yuefeng Li and Ning Zhong. Mining ontology for automatically acquiring web user information needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.

[18] A. Maedche, V. Pekar, and S. Staab. Ontology learning part one on discovering taxonomic relations from the web. In N. Zhong, J. Liu, and Y. Yao, editors, *Web Intelligence*, pages 3–24. Springer, 2003.

[19] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[20] Alexander Maedche and Steffen Staab. Ontology learning. In *Handbook on Ontologies*, pages 173–190. 2004.

[21] G. A. Miller, Beckwith R., C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244, 1990.

[22] Michele Missikoff, Roberto Navigli, and Paola Velardi. Integrated approach to web ontology learning and engineering. *IEEE Computer*, 35(11):60–63, 2002.

[23] Christine A. Montgomery. Concept extraction. *American Journal of Computational Linguistics*, 8(2):70–73, 1982.

[24] Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31, 2003.

[25] I. Nonaka. A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1):14–37, 1994.

[26] I. Nonaka and H. Takeuchi. *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, New York, 1995.

[27] Patrick Perrin and Frederick Petry. Extraction and representation of contextual information for knowledge discovery in texts. *Information Sciences*, 151:125–152, 2003.

[28] M. Porter. An algorithm for suf x stripping. *Program*, 14(3):130–137, 1980.

[29] G. Salton. Full text information processing using the smart system. *Database Engineering Bulletin*, 13(1):2–9, March 1990.

[30] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, New York, 1983.

[31] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213. ACM, 1999.

[32] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124, 1998.

[33] Satoshi Sekine, Jeremy J. Carroll, So a Ananiadou, and Jun'ichi Tsujii. Automatic learning for semantic collocation. In *Proceedings of the third Conference on Applied Natural Language Processing*, pages 104–110, Trento, Italy, March 31–April 3 1992. Association for Computational Linguistics.

[34] C. Shannon. A mathematical theory of communication. *Bell System Technology Journal*, 27:379–423, 1948.

[35] G. Smith. *Computers and Human Language*. Oxford University Press, New York, New York, 1991.

[36] Mark A. Stairmand. Textual context analysis for information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–147, 1997.

[37] Quan Thanh Tho, Siu Cheung Hui, Alvis Cheuk M. Fong, and Tru Hoang Cao. Automatic fuzzy ontology generation for semantic web. *IEEE Transactions on Knowledge and Data Engineering*, 18(6):842–856, 2006.

[38] Christopher A. Welty. Ontology research. *AI Magazine*, 24(3):11–12, 2003.

[39] Dwi H. Widyantoro and John Yen. A fuzzy ontology-based abstract search engine and its user studies. In *The 10th IEEE International Conference on Fuzzy Systems*, pages 1291–1294. IEEE Press, 2001.

[40] Rudolf Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors, *Formal Concept Analysis, Foundations and Applications*, volume 3626, pages 1–33. Springer, 2005.

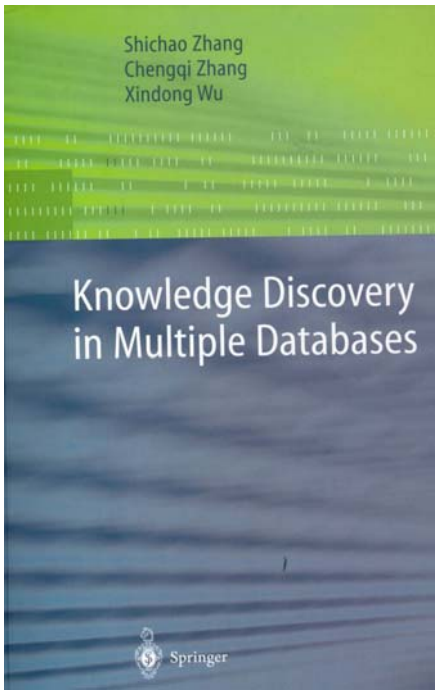
[41] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland, 1996.

[42] L. A. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8:338–353, 1965.

[43] Wlodek Zadrozny. Context and ontology in understanding of dialogs. In *Proceedings of the IJCAI'95 Workshop on Context in NLP*, May 15 1995.

Knowledge Discovery in Multiple Databases

BY SHICHAO ZHANG, CHENGQI ZHANG, XINDONG WU, 2004. ISBN: 978-1-85233-703-2



REVIEW BY RAMESH K. RAYUDU*

In this digital world, we are inundated with terrabytes of data every day. The databases are being stored in distributed environment as companies become global and have offices all over the world. This lead to an interesting research problem in the field of data mining termed multi-database mining. Most solutions in this regard specify merging the databases into a single dataset. But this kind of merging can lead to many problematic issues such as data explosion, destruction of database distribution information, loss of data, and unnecessary inclusion of some data. In simple words, a single dataset may not reflect the real nature of multi-datasets.

*Massey University, New Zealand.
E-mail: r.k.rayudu@massey.ac.nz

"Knowledge Discovery in Multiple Databases" written by S. Zhang, C. Zhang and X. Wu is a book that addresses the issue of data-mining multiple databases. The book is a description of authors' research work and development of a new strategy termed local pattern analysis.

Local pattern analysis is claimed to discover useful patterns that cannot be mined in traditional multi-database mining techniques. The book discusses knowledge discovery principles at different levels of detail. Novice dataminers, researchers, academics and students will find the book helpful.

Functionally "Knowledge Discovery in Multiple Databases" is organised in two parts. The first half introduces knowledge discovery, multi-databases and some related research. The second half discusses the authors' techniques for pre-processing the data and identifying patterns from multi-databases. The authors provide a smooth transition from introducing a reader with multi-database mining (MDM) to their application and research. The authors also point out some very important shortcomings of earlier research and provide a good insight into the practical aspects of MDM.

Chapter 1 provides an introduction to data mining and discusses association rule mining in a format that keeps the reader interested in the topic. The chapter provided a good insight into mining aspect of both single and multi-databases. Authors' statement that dual-level applications present many challenges could be clearly identified from the reading the chapter.

The authors clearly identified the practical issues of a MDM process with emphasis on design of application independent database clustering. Identifying quality knowledge and resolving conflicts are important aspects of any MDM and challenge the boundaries of MDM research. Authors address these topics and also provide solutions. The chapter ends with authors' discussion on identifying the features of MDM and contrasting their research against each feature.

Knowledge discovery in databases (KDD) has been an active research area since mid-nineties and, since then, several books have been published in the area. The authors' also discuss this topic in their second chapter of the book. They start with the traditional approach of discussing the processing steps involved in KDD and continue on to discuss the latest research in each process. From this broad overview, the authors narrow their discussion on to association rule mining to discuss its effects on mining mono-databases. The final part of the chapter discusses the relevant research into MDM. Several algorithms including meta-learning and parallel datamining were discussed in appropriate detail so that a reader can understand the concepts.

Chapter 3 introduces authors' Local Pattern Analysis (LPA) strategy that is based on a competing model in sports. As each sport has a set of rules to choose its winners, LPA was developed to recognize patterns in multi-databases based on a dual-level multi-rule strategy. Through the strategy, they identify the three useful patterns: high-vote patterns, exceptional and global patterns.

To recognize patterns in multi-databases, the authors demonstrate the structure of a pattern and represent it in a multi-dimension space where each dimension is a selection factor. The details of the algorithm were avoided in the chapter and were dealt in the

subsequent chapters. In the final section of the chapter, the authors demonstrate and discuss the effectiveness of their algorithm LPA.

Chapter 4 is dedicated towards detection of quality knowledge in high-veridical data sources. The authors demonstrate the effectiveness of identifying quality data by considering one internal and six external databases as an example and then apply several existing techniques on the databases. They demonstrate that the technique discussed in the book can successfully detect the frequent itemsets while still preserving the distributive nature of the databases. The basic techniques that are necessary for identifying quality data has been discussed extensively in the chapter.

They developed several semantics to state that a veridical data-source combines the collected knowledge with a set of possibilities to obtain a higher level knowledge. The authors readily acknowledge that there may not be veridical data sources in the real-world but veridical properties can originate from other sources. Their developed framework was then applied to real-world databases. The application demonstrated that the stated algorithm can successfully identify data-sources with high success ratio based on their veridicality. The authors' claims that their algorithm works by distinguishing internal and external knowledge and the elimination of untrustworthy and fraudulent knowledge by veridicality analysis can be established through this chapter.

Chapter 5 is dedicated to identification of relevant databases for a datamining application. The authors use classification techniques to identify relevant databases. Classification is a challenging process and depends extensively on selected features related to an object being classified. The authors develop a new clustering algorithm that is application independent and utilizes MDM.

To search for a good classification from given multiple databases a two-step process is stated in this chapter. The first step is to design a procedure that generates a classification for a given threshold. The second step is to develop an algorithm that can search for a good classification based on a distance measure that measures the goodness of a database class. The developed algorithm was then applied to a set of databases and their results analyzed. The performance improvement is impressive and certainly advanced from other algorithms.

Decision making systems that use negative association rules to identify the mutually exclusive correlations among data items can create a problem when dealing with multi-databases. Identifying these conflicts and resolving them is an important aspect and is the topic of Chapter 6 of the book. The authors address the issue by introducing a local pattern synthesizing operator that can identify the local pattern set and resolves inconsistency using the weighted majority principle. The authors review some basic concepts of modal logic and construct a proof theory of the proposed logic. The last section of the chapter discusses on how to use the proposed logic framework to identify quality knowledge from multiple databases.

Chapters 7 and 8 discuss the detection of high-vote and exceptional patterns. To identify a high-vote pattern, the authors introduce a voting rate that is calculates voting of each branch. The measure of interest of a high voting pattern is stemmed from the relationship between the voting rate and the average voting rate (random pattern area). A fuzzy logic controller is then used to further identify the high-vote patterns.

While high-vote patterns show the commonality between different branches of a company, exceptional patterns depict the patterns that are unique to each branch. The authors' algorithm for identifying exceptional patterns was discussed in Chapter 8. The exceptional patterns were identified by a 'measure of interestingness' developed by the authors. After a good discussion of the algorithm, they demonstrate the algorithm using an example.

The major highlight of this book is the discussion of algorithms in relation to many practical problems and the functional hierarchy of a company. The intricacies and problems related to a multi-branch organization were discussed and related with the algorithms specified in the book. One such intricacy is the importance given to different branches in a company. The authors consider this importance and incorporate it into their model for synthesizing global patterns from local patterns.

As specified in Chapter 6 they use a weighting measure to achieve the synthesis. The resulting algorithm is then discussed in Chapter 9 of the book. The stated model synthesizes the association rules from multiple databases using a weighting process. The weighting process is elaborate and well discussed in the book. The algorithm developed is stated to synthesize rules from different databases and different databases can be mined concurrently.

The final chapter highlights the book's contributions and addresses the future issues related to the research.

The authors, through this book, provide a practical and logical approach to solving problems related to knowledge discovery in multi-database systems. Data mining multiple databases is a complex task and needs a proper direction. This book provides that direction. The authors addressed the problem well and discussed their solutions systematically.

One highlight of the book is the discussion on authors' creation of several MDM techniques and methodologies towards solving some practical problems. The primary focus of the book is to demonstrate some new techniques in mining multi-databases and the authors have certainly succeeded.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

AWIC'08
The Sixth Atlantic Web Intelligence Conference
Cape Town, South Africa
June 30-July 3, 2008
http://www.fullcycles.org/AWIC2008/

The Atlantic Web Intelligence Conference (Spain – 2003, Mexico – 2004, Poland – 2005, Israel – 2006, France – 2007) brings together scientists, engineers, computer users, and students to exchange and share their experiences, new ideas, and research results about all aspects (theory, applications and tools) of intelligent methods applied to Web based systems, and to discuss the practical challenges encountered and the solutions adopted.

The conference will cover a broad set of intelligent methods, with particular emphasis on soft computing. Methods such as (but not restricted to): Neural Networks, Fuzzy Logic, Multivalued Logic, Rough Sets, Ontologies, Evolutionary Programming, Intelligent CBR, Genetic Algorithms, Semantic Networks, Intelligent Agents, Reinforcement Learning, Knowledge Management, etc. should be related to applications on the Web like: Web Design, Information Retrieval, Electronic Commerce, Conversational Systems, Recommender Systems, Browsing and Exploration, Adaptive Web, User Profiling/Clustering, E-mail/SMS filtering, Negotiation Systems, Security, Privacy, and Trust, Web-log Mining, etc.

.

WI 2008
The 2008 IEEE/WIC/ACM International Conference on Web Intelligence
Sydney, Australia
December 9-12, 2008

Web Intelligence (WI) has been recognized as a new direction for scientific research and development to explore the fundamental roles as well as practical impacts of Artificial Intelligence (AI) (e.g., knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence) and advanced Information Technology (IT) (e.g., wireless networks, ubiquitous devices, social networks, semantic Web, wisdom Web, and data/knowledge grids) on the next generation of Web-empowered products, systems, services, and activities. It is one of the most important as well as promising IT research fields in the era of Web and agent intelligence.

The 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI'08) will be jointly held with the 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'08). The IEEE/WIC/ACM 2008 joint conferences are organized by University of Technology, Sydney, Australia, and sponsored by IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), Web Intelligence Consortium (WIC), and ACM-SIGART.

Following the great successes of WI'01 held in Maebashi City, Japan , WI'03 held in Halifax, Canada, WI'04 held in Beijing, China, WI'05 in Compiegne University of Technology, France, and WI'06 held in Hong Kong, WI'07 held in Silicon Valley USA, WI'08 provides a leading international forum for researchers and practitioners (1) to present the state-of-the-art of WI technologies; (2) to examine performance characteristics of various approaches in Web-based intelligent information technology; and (3) to cross-fertilize ideas on the development of Web-based intelligent information systems among different domains. By idea-sharing and discussions on the underlying foundations and the enabling technologies of Web intelligence, WI'08 will capture current important developments of new models, new methodologies and new tools for building a variety of embodiments of Web-based intelligent information systems.

IAT 2008
The 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology
Sydney, Australia
December 9-12, 2008

Following the great successes of IAT'01, IAT'03, IAT'04, IAT'05, IAT'06 and IAT'07, we are excited to propose Sydney as the site for the 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'08), to be jointly held with the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI'08), will be held in Sydney, Australia. IAT 2008 is sponsored by the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), the Web Intelligence Consortium (WIC), and ACM-SIGART.

IAT'08 will provide a leading international forum to bring together researchers and practitioners from diverse fields, such as computer science, information technology, business, education, human factors, systems engineering, and robotics, to (1) examine the design principles and performance characteristics of various approaches in intelligent agent technology, and (2) increase the cross-fertilization of ideas on the development of autonomous agents and multi-agent systems among different domains. By encouraging idea-sharing and discussions on the underlying logical, cognitive, physical, and sociological foundations as well as the enabling technologies of intelligent agents, IAT 2008 will foster the development of novel paradigms and advanced solutions in agent-based computing.

ICDM'08
The Eighth IEEE International Conference on Data Mining
Pisa, Italy
December 15-19, 2008
http://icdm08.isti.cnr.it

The IEEE International Conference on Data Mining series (ICDM) has established itself as the world's premier research conference in data mining, providing a leading forum for presentation of original research results, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of data mining, including algorithms, software and systems, and applications. In addition, ICDM

draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems, and high performance computing. By promoting novel, high quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to continuously advance the state-of-the-art in data mining. Besides the technical program, the conference will feature workshops, tutorials, panels, and the ICDM data mining contest. High quality papers in all data mining areas are solicited. Original papers exploring new directions will receive especially careful consideration. Papers that have already been accepted or are currently under review for other conferences or journals will not be considered for ICDM '08.

A selected number of IEEE ICDM '08 accepted papers will be invited for possible inclusion, in expanded and revised form, in the Knowledge and Information Systems journal (http://www.cs.uvm.edu/~kais/) published by Springer-Verlag. IEEE ICDM Best Paper Awards will be conferred at the conference on the authors of (1) the best research paper and (2) the best application paper. Strong, foundational, results will be considered for the best research paper award and application-oriented submissions will be considered for the best application paper award. Other than technical paper presentation sessions, ICDM'08 will host short and long tutorials as well as workshops that focus on new research directions and initiatives. All accepted workshop papers will be included in a separate workshop proceedings published by the IEEE Computer Society Press. Also, a call for organizing a data mining contest will be issued to challenge researchers and practitioners with a real practical data mining problem.

Related Conferences

AAMAS'08
The Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems
Estoril, Portugal
May 12-16, 2008
http://gaips.inesc-id.pt/aamas2008/

AAMAS is the leading scientific conference for research in autonomous agents and multi-agent systems. The AAMAS conference series was initiated in 2002 as a merger of three highly respected individual conferences: the International Conference in Autonomous Agents, the International Workshop on Agent Theories, Architectures, and Languages, and the International Conference on Multi-Agent Systems. The aim of the joint conference is to provide a single, high-profile, internationally respected archival forum for research in all aspects of the theory and practice of autonomous agents and multi-agent systems.

Tutorials and workshops will be held on Monday May 12th and Tuesday May 13th, 2008. Accepted technical papers and invited talks will be presented from Wednesday May 14th through Friday May 16th, 2008. This year AAMAS will feature two special tracks: one on Multi-Robots and the other on Virtual Agents. The goal is to provide an opportunity for interaction and cross-fertilization between the AAMAS community and researchers working in these fields and to strengthen links between the two communities. For more information please contact Pedro Lima at robotics@aamas2008.org and Elisabeth Andre at synthetic@aamas2008.org. AAMAS will also continue to feature an 'Industry and Applications' track and a Demonstrations session.

ISWC'08
The Seventh International Semantic Web Conference
Karlsruhe
http://iswc2008.semanticweb.org/

ISWC is a major international forum where visionary and state-of-the-art research of all aspects of the Semantic Web are presented. ISWC'06 follows the 1st International Semantic Web Conference (ISWC'02 which was held in Sardinia, Italy, 9-12 June 2002), the 2nd International Semantic Web Conference (ISWC'03 which was held in Florida, USA, 20 - 23 October 2003), 3rd International Semantic Web Conference (ISWC'04 which was held in Hiroshima, Japan, 7 - 11 November 2004), 4th International Semantic Web Conference 2005 (ISWC'05 which was held in Galway, Ireland, 6 - 10 November, 2005), 5th (ISWC'06 which was held in Athens, GA, USA 5 - 9 November, 2006), and 6th (ISWC'07 which was held in Busan, Korea 11 - 15 November, 2007).

SDM'08
2008 SIAM International Conference on Data Mining
Atlantic, Georgia, USA
April 24-26, 2008
http://www.siam.org/meetings/sdm08/

Data mining and knowledge discovery is rapidly becoming an important tool in all walks of human endeavor including science, engineering, industrial processes, healthcare, business, medicine and society. The datasets in these fields are large, complex, and often noisy. Extracting knowledge requires the use of sophisticated, high-performance and principled analysis techniques and algorithms, based on sound statistical foundations. These techniques in turn require powerful visualization technologies; implementations that must be carefully tuned for performance; software systems that are usable by scientists, engineers, and physicians as well as researchers; and infrastructures that support them. For the main conference the program committee seeks outstanding papers in all areas pertaining to data mining and knowledge discovery.

This conference provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. It also provides an ideal setting for graduate students and others new to the field to learn about cutting-edge research by hearing outstanding invited speakers and attending tutorials (included with conference registration). A set of focused workshops are also held on the last day of the conference. The proceedings of the conference are published in archival form, and are also made available on the SIAM web site.

AAAI'08
The Twenty-Third Conference on Artificial Intelligence
Chicago, Illinois, USA
July 13–17, 2008
http://www.aaai.org/Conferences/AAAI/

AAAI'08 is the Twenty-Third AAAI Conference on Artificial Intelligence (AI). Sponsored by the Association for the Advancement of Artificial Intelligence, the purpose of this conference is to promote research in AI and scientific exchange among AI researchers, practitioners, scientists, and engineers in related disciplines. AAAI'08 will have multiple technical tracks, student abstracts, poster sessions, invited speakers, and exhibit programs, all selected according to the highest reviewing standards.

AAAI'08 welcomes submissions on mainstream AI topics as well as novel cross-cutting work in related areas. Topics include but are not limited to the following: Agents; Cognitive modeling and human

interaction; Commonsense reasoning; Constraint satisfaction; Evolutionary computation; Game playing and interactive entertainment; Information integration and extraction; Knowledge acquisition and ontologies; Machine learning and data mining; Model-based systems; Natural language processing; Planning and scheduling; Robotics; Search; Semantic web; Vision and perception.

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398