ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 5, July 2012

A Customized Ontological-Based Web Information Collection Model

Krishna Madhav Majety, Uma Devi Deva, Sailaja Sanaka

Abstract —It is known that Ontologies can be used to describe and formalize knowledge to represent user profiles in personalized web information collection. Though, there are many models adopted to represent user profiles either from a global knowledge base or a user local information repository. In this paper, a customized ontological-based model is proposed for representing reasoning among user profiles. The proposed ontology model learns ontological user profiles which include world knowledge base and the instance of local user repositories. As the ontological model is compared it against existing benchmark models in web information gathering, the results shows the significant improvement.

Index Terms – Customization, local user repository, knowledge description, Ontology, User Profiles, Web Information.

I. INTRODUCTION

In recent trends, the exponential growth of Information available in World Wide Web needs more efficient techniques and proficient mechanisms to locate and capture relevant web information that we required. Current web information gathering systems attempt to satisfy user requirements by capturing their information needs. For this purpose, user profiles are created for user background knowledge description [12], [22], [23].

User profiles represent the concept models possessed by users when gathering web information. A concept model is implicitly possessed by users and is generated from their background knowledge. While this concept model cannot be proven in laboratories, many web ontologists have observed it in user behavior [23]. When users read through a document, they can easily determine whether or not it is of their interest or relevance to them, a judgment that arises from their implicit concept models. If a user's concept model can be simulated, then a superior representation of user profiles can be built.

To simulate user concept models, ontologies—a knowledge description and formalization model are utilized in personalized web information gathering. Such Ontologies are called ontological user profiles [12], [35] or personalized Ontologies [39]. To represent user profiles, many researchers have attempted to discover user background knowledge through global or local analysis.

Global analysis uses existing global knowledge bases for user background knowledge representation. Commonly used knowledge bases include generic ontologies (e.g., WordNet [26]), thesauruses (e.g., digital libraries), and online knowledge bases (e.g., online categorizations and Wikipedia). The global analysis techniques produce effective performance for user background knowledge extraction. However, global analysis is limited by the quality of the used knowledge base. For example, WordNet was reported as helpful in capturing user interest in some areas but useless for others [44].

Local analysis investigates user local information or observes user behavior in user profiles. For example, Li and Zhong [23] discovered taxonomical patterns from the user's local text documents to learn ontologies for user profiles. Some groups [12], [35] learned personalized Ontologies adaptively from user's browsing history. Alternatively, Sekine and Suzuki [33] analyzed query logs to discover user background knowledge. In some works, such as [32], users were provided with a set of documents and asked for relevance feedback. User background knowledge was then discovered from this feedback for user profiles. However, because local analysis techniques rely on data mining or classification techniques for knowledge discovery, occasionally the discovered results contain noisy and uncertain information. As a result, local analysis suffers from ineffectiveness at capturing formal user knowledge.

From this, we can hypothesize that user background knowledge can be better discovered and represented if we can integrate global and local analysis within a hybrid model. The knowledge formalized in a global knowledge base will constrain the background knowledge discovery from the user local information. Such a personalized ontology model should produce a superior representation of user profiles for web information gathering.

In this paper, an ontology model to evaluate this hypothesis is proposed. This model simulates users' concept models by using personalized ontologies, and attempts to improve web information gathering performance by using ontological user profiles. The world knowledge and a user's local instance repository (LIR) are used in the proposed model. World knowledge is commonsense knowledge acquired by people from experience and education [46]; an

Manuscript received Oct 15, 2011.

Krishna Madhav Majety (M.Tech-CSE), Department of Computer Science and Engineering, Sri Mittapalli College of Engineering, affiliated to Jawaharlal Nehru Technological University Kakinada, India, (e-mail: krishnamadhav.mtech@gamil.com).

Uma Devi Deva (P.hD), M.Tech, Head of the Department, Department of Computer Science and Engineering, Sri Mittapalli College of Engineering, Jawaharlal Nehru Technological University Kakinada, India, (e-mail: p_umapremchand@yahoo.co.in).

Sailaja Sanaka, Asst. Professor, Department of Computer Science and Engineering, SRK Institute of Technology, Jawaharlal Nehru Technological University Kakinada, India (e-mail: sailaja2012research@gmail.com).

LIR is a user's personal collection of information items. From a world knowledge base, we construct personalized ontologies by adopting user feedback on interesting knowledge. A multidimensional ontology mining method, Specificity and Exhaustivity, is also introduced in the proposed model for analyzing concepts specified in ontologies. The users' LIRs are then used to discover background knowledge and to populate the personalized ontologies. The proposed ontology model is evaluated by comparison against some benchmark models through experiments using a large standard data set. The evaluation results show that the proposed ontology model is successful.

The research contributes to knowledge engineering, and has the potential to improve the design of personalized web information gathering systems. The contributions are original and increasingly significant, considering the rapid explosion of web information and the growing accessibility of online documents.

II. Related Work

A. Ontology Learning

Global knowledge bases were used by many existing models to learn ontologies for web information gathering. For example, Gauch et al. [12] and Sieg et al. [35] learned personalized ontologies from the Open Directory Project to specify users' preferences and interests in web search. On the basis of the Dewey Decimal Classification, King et al. [18] developed IntelliOnto to improve performance in distributed web information retrieval. Wikipedia was used by Downey et al. [10] to help understand underlying user interests in queries. These works effectively discovered user background knowledge; however, their performance was limited by the quality of the global knowledge bases.

Aiming at learning personalized ontologies, many works mined user background knowledge from user local information. Li and Zhong [23] used pattern recognition and association rule mining techniques to discover knowledge from user local documents for ontology construction. Tran et al. [42] translated keyword queries to Description Logics' conjunctive queries and used ontologies to represent user background knowledge. Zhong [47] proposed a domain ontology learning approach that employed various data mining and natural-language understanding techniques. Navigli et al. [28] developed OntoLearn to discover semantic concepts and relations from web documents. Web content mining techniques were used by Jiang and Tan [16] to discover semantic knowledge from domain-specific text documents for ontology learning. Finally, Shehata et al. [34] captured user information needs at the sentence level rather than the document level, and represented user profiles by the Conceptual Ontological Graph. The use of data mining techniques in these models leads to more user background knowledge being discovered. However, the knowledge discovered in these works contained noise and uncertainties.

Additionally, ontologies were used in many works to improve the performance of knowledge discovery. Using a fuzzy domain ontology extraction algorithm, a mechanism was developed by Lau et al. [19] in 2009 to construct concept maps based on the posts on online discussion forums. Quest and Ali [31] used ontologies to help data mining in biological databases. Jin et al. [17] integrated data mining and information retrieval techniques to further enhance knowledge discovery. Doan et al. [8] proposed a model called GLUE and used machine learning techniques to find similar concepts in different ontologies. Dou et al. [9] proposed a framework for learning domain ontologies using pattern decomposition, clustering/classification, and association rules mining techniques. These works attempted to explore a route to model world knowledge more efficiently.

B. User Profiles

User profiles were used in web information gathering to interpret the semantic meanings of queries and capture user information needs [12], [14], [23], [41], [48]. User profiles were defined by Li and Zhong [23] as the interesting topics of a user's information need. They also categorized user profiles into two diagrams: the data diagram user profiles acquired by analyzing a database or a set of transactions [12], [23], [25], [35], [37]; the information diagram user profiles acquired by using manual techniques, such as questionnaires and interviews [25], [41] or automatic techniques, such as information retrieval and machine learning [30]. Van der Sluijs and Huben [43] proposed a method called the Generic User Model Component to improve the quality and utilization of user modeling. Wikipedia was also used by [10], [27] to help discover user interests. In order to acquire a user profile, Chirita et al. [6] and Teevan et al. [40] used a collection of user desktop text documents and emails, and cached web pages to explore user interests. Makris et al. [24] acquired user profiles by a ranked local set of categories, and then utilized web pages to personalize search results for a user. These works attempted to acquire user profiles in order to discover user background knowledge.

User profiles can be categorized into three groups: interviewing, semi-interviewing, and non-interviewing. Interviewing user profiles can be deemed perfect user profiles. They are acquired by using manual techniques, such as questionnaires, interviewing users, and analyzing user classified training sets. One typical example is the TREC Filtering Track training sets, which were generated manually [32]. The users read each document and gave a positive or negative judgment to the document against a given topic. Because, only users perfectly know their interests and preferences, these training documents accurately reflect user background knowledge. Semi-interviewing user profiles are acquired by semi-automated techniques with limited user involvement. These techniques usually provide users with a list of categories and ask users for interesting or non-interesting categories. One typical example is the web training set acquisition model introduced by Tao et al. [38], which extracts training sets from the web based on user fed back categories. Non-interviewing techniques do not involve users at all, but ascertain user interests instead. They acquire user profiles by observing user activity and behavior and discovering user background knowledge [41]. A typical model is OBIWAN, proposed by Gauch et al. [12], which acquires user profiles based on users' online browsing history. The interviewing, semi-interviewing, and

non-interviewing user profiles can also be viewed as manual, semiautomatic, and automatic profiles, respectively.

III. Personalized Ontology Construction

Personalized ontologies are a conceptualization model that formally describes and specifies user background knowledge. From observations in daily life, we found that web users might have different expectations for the same search query. For example, for the topic "New York," business travelers may demand different information from leisure travelers. Sometimes even the same user may have different expectations for the same search query if applied in a different situation. A user may become a business traveler when planning for a business trip, or a leisure traveler when planning for a family holiday. Based on this observation, an assumption is formed that web users have a personal concept model for their information needs. A user's concept model may change according to different information needs. In this section, a model constructing personalized Ontologies for web users's concept models is introduced.

A. World Knowledge Representation

World knowledge is important for information gathering. According to the definition provided by [46], world knowledge is commonsense knowledge possessed by people and acquired through their experience and education. Also, as pointed out by Nirenburg and Raskin [29], "world knowledge is necessary for lexical and referential disambiguation, including establishing co-reference relations and resolving ellipsis as well as for establishing and maintaining connectivity of the discourse and adherence of the text to the text producer's goal and plans." In this proposed model, user background knowledge is extracted from a world knowledge base encoded from the Library of Congress Subject Headings (LCSH).

We first need to construct the world knowledge base. The world knowledge base must cover an exhaustive range of topics, since users may come from different backgrounds. For this reason, the LCSH system is an ideal world knowledge base. The LCSH was developed for organizing and retrieving information from a large volume of library collections. For over a hundred years, the knowledge contained in the LCSH has undergone continuous revision and enrichment. The LCSH represents the natural growth and distribution of human intellectual work, and covers comprehensive and exhaustive topics of world knowledge [5].

	LCSH	LCC	DDC	RC
# of Topics	394,070	4,214	18,462	100,000
Structure	Directed Acyclic Graph	Tree	Tree	Directed Acyclic Graph
Depth	37	7	23	10
Semantic Relations	Broader, Used-for, Related-to	Super- and Sub-class	Super- and Sub-class	Super- and Sub-class

table 1. Comparison of Different World Taxonomies

In addition, the LCSH is the most comprehensive non-specialized controlled vocabulary in English. In many respects, the system has become a de facto standard for subject cataloging and indexing, and is used as a means for enhancing subject access to knowledge management systems [5].

The LCSH system is superior compared with other world knowledge taxonomies used in previous works. Table 1 presents a comparison of the LCSH with the Library of Congress Classification (LCC) used by Frank and Paynter [11], the Dewey Decimal Classification (DDC) used by Wang and Lee [45] and King et al. [18], and the reference categorization (RC) developed by Gauch et al. [12] using online categorizations. As shown in Table 1, the LCSH covers more topics, has a more specific structure, and specifies more semantic relations. The LCSH descriptors are classified by professionals, and the classification quality is guaranteed by well-defined and continuously refined cataloging rules [5]. These features make the LCSH an ideal world knowledge base for knowledge engineering and management.

The structure of the world knowledge base used in this research is encoded from the LCSH references. The LCSH system contains three types of references: Broader term (BT), Used-for (UF), and Related term (RT) [5]. The BT references are for two subjects describing the same topic, but at different levels of abstraction (or specificity). In our model, they are encoded as the is-a relations in the world knowledge base. The UF references in the LCSH are used for many semantic situations, including broadening the semantic extent of a subject and describing compound subjects and subjects subdivided by other topics. The complex usage of UF references makes them difficult to encode. During the investigation, we found that these references are often used to describe an action or an object. When object A is used for an action, A becomes a part of that action (e.g., "a fork is used for dining"); when A is used for another object, B, A becomes a part of B (e.g., "a wheel is used for a car"). These cases can be encoded as the part-of relations. Thus, we simplify the complex usage of UF references in the LCSH and encode them only as the part-of relations in the world knowledge base. The RT references are for two subjects related in some manner other than by hierarchy. They are encoded as the related-to relations in our world knowledge base.

The primitive knowledge unit in our world knowledge base is subjects. They are encoded from the subject headings in the LCSH. These subjects are formalized as follows:

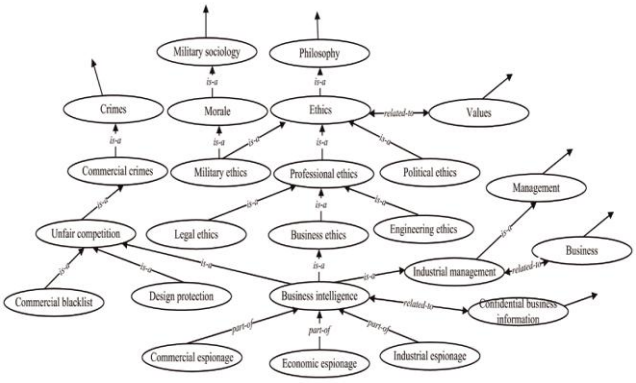


Fig 1. A sample part of the world knowledge base

Definition 1. Let S be a set of subjects, an element s ∈ S is formalized as a 4-tuple s := <label; neighbor; ancestor; Descendant>, where

- label is the heading of s in the LCSH thesaurus;
- neighbor is a function returning the subjects that have direct links to s in the world knowledge base;
- ancestor is a function returning the subjects that have a higher level of abstraction than s and link to s directly or indirectly in the world knowledge base;
- descendant is a function returning the subjects that are more specific than s and link to s directly or indirectly in the world knowledge base.

The subjects in the world knowledge base are linked to each other by the semantic relations of is-a, part-of, and related-to. The relations are formalized as follows:

Definition 2. Let R be a set of relations, an element r ∈ R is a 2-tuple r:= <edge, type>, where

- an edge connects two subject that hold a type of relation;
- a type of relations is an element of {is-a, part-of, related-to}.

With Definitions 1 and 2, the world knowledge base can then be formalized as follows:

Definition 3. Let WKB be a world knowledge base, which is a taxonomy constructed as a directed acyclic graph. The WKB consists of a set of subjects linked by their semantic relations, and can be formally defined as a 2-tuple WKB := <S, R>, where

Fig.1 illustrates a sample of the WKB dealing with the topic "Economic espionage." (This topic will also be used as an example throughout this paper to help explanation.)

B. Ontology Contruction

The subjects of user interest are extracted from the WKB via user interaction. A tool called Ontology Learning Environment (OLE) is developed to assist users with such interaction. Regarding a topic, the interesting subjects consist of two sets: positive subjects are the concepts relevant to the information need, and negative subjects are the concepts resolving paradoxical or ambiguous interpretation of the information need. Thus, for a given topic, the OLE provides users with a set of candidates to identify positive and negative subjects. These candidate subjects are extracted from the WKB.

Fig. 2 is a screen-shot of the OLE for the sample topic "Economic espionage." The subjects listed on the top-left panel of the OLE are the candidate subjects presented in hierarchical form. For each s 2 SS, the s and its ancestors are retrieved if the label of s contains any one of the query terms in the given topic (e.g., "economic" and "espionage"). From these candidates, the user selects positive subjects for the topic. The user-selected positive subjects are presented on the top-right panel in hierarchical form.

The candidate negative subjects are the descendants of the user-selected positive subjects. They are shown on the bottom-left panel. From these negative candidates, the user selects the negative subjects. These user-selected negative subjects are listed on the bottom-right panel (e.g., "Political ethics" and "Student ethics"). Note that for the completion of the structure, some positive subjects (e.g., "Ethics," "Crime," "Commercial crimes," and "Competition Unfair") are also included on the bottom-right panel with the negative subjects. These positive subjects will not be included in the negative set.

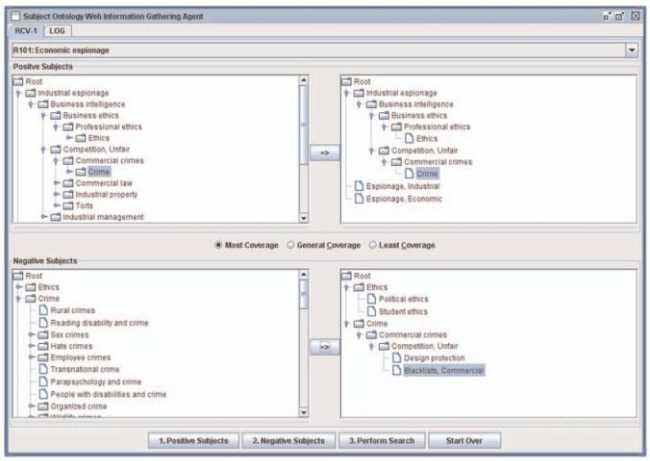


Fig. Ontology learning environment

The remaining candidates, which are not fed back as either positive or negative from the user, become the neutral subjects to the given topic.

An ontology is then constructed for the given topic using these user fed back subjects. The structure of the ontology is based on the semantic relations linking these subjects in the WKB. The ontology contains three types of knowledge: positive subjects, negative subjects, and neutral subjects. Fig. 3 illustrates the ontology (partially) constructed for the sample topic "Economic espionage," where the white nodes are positive, the dark nodes are negative, and the gray nodes are neutral subjects.

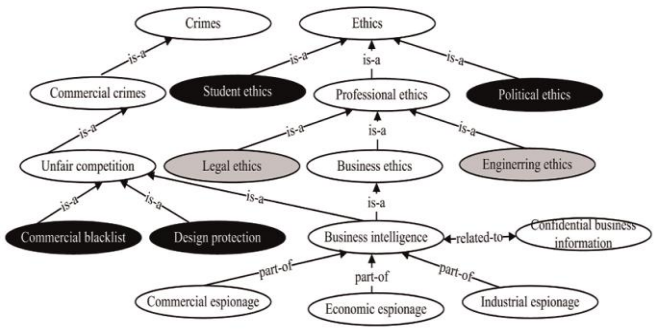


Fig 3. An ontology (partial) constructed for topic "Economic Espionage."

IV. MULTI-DIMENSIONAL ONTOLOGY MINING

Ontology mining discovers interesting and on-topic knowledge from the concepts, semantic relations, and instances in an ontology. In this section, a 2D ontology mining method is introduced: Specificity and Exhaustivity. Specificity (denoted spe) describes a subject's focus on a given topic. Exhaustivity (denoted exh) restricts a subject's semantic space dealing with the topic. This method aims to investigate the subjects and the strength of their associations in an ontology.

A. Semantic Specificity

The semantic specificity is investigated based on the structure of *O(T)* inherited from the world knowledge base. The strength of such a focus is influenced by the subject's locality in the taxonomic structure taxS of *O(T)* (this is also argued by [42]). As stated in Definition 4, the tax^s of *O(T)* is a graph linked by semantic relations. The subjects located at upper bound levels toward the root are more abstract than those at lower bound levels toward the "leaves." The upper bound level subjects have more descendants, and thus refer to more concepts, compared with the lower bound level subjects. Thus, in terms of a concept being referred to by both an upper bound and lower bound subjects, the lower bound subject has a stronger focus because it has fewer concepts in its space. Hence, the semantic specificity of a lower bound subject is greater than that of an upper bound subject.

The semantic specificity is measured based on the hierarchical semantic relations (is-a and part-of) held by a subject and its neighbors in taxS.1 Because subjects have a fixed locality on the tax^s of *O(T)*, semantic specificity is also called absolute specificity and denoted by $spe_a(s)$.

input : a personalized ontology $\mathcal{O}(\mathcal{T}) := \langle tax^{\mathcal{S}}, rel\rangle$; a coefficient θ between (0,1).
output: $spe_a(s)$ applied to specificity.

1 set $k = 1$, get the set of leaves S_0 from $tax^{\mathcal{S}}$, for $(s_0 \in S_0)$ assign $spe_a(s_0) = k$;
2 get S' which is the set of leaves in case we remove the nodes S_0 and the related edges from $tax^{\mathcal{S}}$;
3 **if** $(S' == \emptyset)$ **then** return;//*the terminal condition*;
4 **foreach** $s' \in S'$ **do**
5 **if** $(isA(s') == \emptyset)$ **then** $spe_a^1(s') = k$;
6 **else** $spe_a^1(s') = \theta \times min\{spe_a(s)|s \in isA(s')\}$;
7 **if** $(partOf(s') == \emptyset)$ **then** $spe_a^2(s') = k$;
8 **else** $spe_a^2(s') = \frac{\sum_{s \in partOf(s')} spe_a(s)}{|partOf(s')|}$;
9 $spe_a(s') = min(spe_a^1(s'), spe_a^2(s'))$;
10 **end**
11 $k = k \times \theta, S_0 = S_0 \cup S'$, go to step 2.

Algorithm 1. Analyzing semantic relations for specificity

B. Topic Specificity

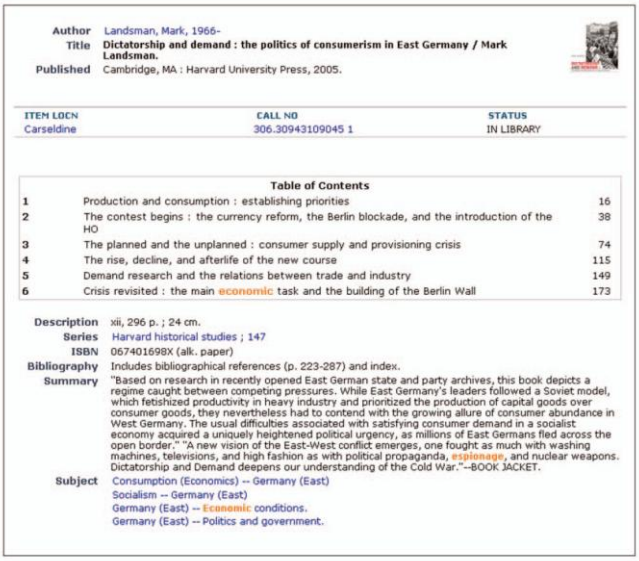
The topic specificity of a subject is investigated, based on the user background knowledge discovered from user local information.

User Local Instance Repository

User background knowledge can be discovered from user local information collections, such as a user's stored documents, browsed web pages, and composed/received emails [6]. The ontology O(T) constructed in Section 3 has only subject labels and semantic relations specified. In this section, we populate the ontology with the instances generated from user local information collections. We call such a collection the user's local instance repository (LIR).

Generating user local LIRs is a challenging issue. The documents in LIRs may be semi-structured (e.g., the browsed HTML and XML web documents) or unstructured (e.g., the stored local DOC and TXT documents). In some semi-structured web documents, content-related descriptors are specified in the metadata sections. These descriptors have direct reference to the concepts specified in a global knowledge base, for example, the infoset tags in some XML documents citing control vocabularies in global lexicons.

These documents are ideal to generate the instances for ontology population. When different global knowledge bases are used, ontology mapping techniques can be used to match the concepts in different representations. Approaches like the concept map generation mechanism developed by Lau et al. [19], the GLUE system developed by Doan et al. [8], and the approximate concept mappings introduced by Gligorov et al. [13] are useful for such mapping of different world knowledge bases.


Author Landsman, Mark, 1966-
Title **Dictatorship and demand : the politics of consumerism in East Germany / Mark Landsman.**
Published Cambridge, MA : Harvard University Press, 2005.

ITEM LOCN	CALL NO	STATUS
Carseldine	306.30943109045 1	IN LIBRARY

Table of Contents

1	Production and consumption : establishing priorities	16
2	The contest begins : the currency reform, the Berlin blockade, and the introduction of the HO	38
3	The planned and the unplanned : consumer supply and provisioning crisis	74
4	The rise, decline, and afterlife of the new course	115
5	Demand research and the relations between trade and industry	149
6	Crisis revisited : the main economic task and the building of the Berlin Wall	173

Description xii, 296 p. ; 24 cm.
Series Harvard historical studies ; 147
ISBN 067401698X (alk. paper)
Bibliography Includes bibliographical references (p. 223-287) and index.
Summary "Based on research in recently opened East German state and party archives, this book depicts a regime caught between competing pressures. While East Germany's leaders followed a Soviet model, which fetishized productivity in heavy industry and prioritized the production of capital goods over consumer goods, they nevertheless had to contend with the growing allure of consumer abundance in West Germany. The usual difficulties associated with satisfying consumer demand in a socialist economy acquired a uniquely heightened political urgency, as millions of East Germans fled across the open border." "A new vision of the East-West conflict emerges, one fought as much with washing machines, televisions, and high fashion as with political propaganda, espionage, and nuclear weapons. Dictatorship and Demand deepens our understanding of the Cold War."--BOOK JACKET.
Subject Consumption (Economics) -- Germany (East)
Socialism -- Germany (East)
Germany (East) -- Economic conditions.
Germany (East) -- Politics and government.

Fig. 4. An information item in QUT library catalogs.

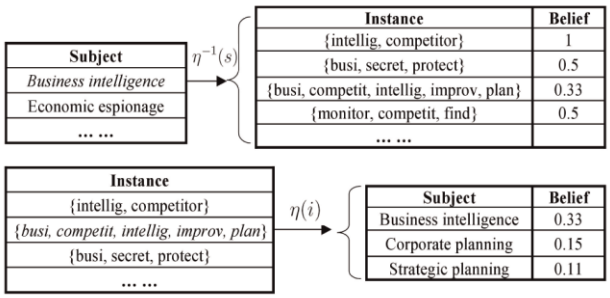


Fig. 5. Mappings of subjects and instances.

C. Multidimensional Analysis of Subjects

The exhaustivity of a subject refers to the extent of its concept space dealing with a given topic. This space extends if a subject has more positive descendants regarding the topic. In contrast, if a subject has more negative descendants,

its exhaustivity decreases. Based on this, let desc(s) be a function that returns the descendants of s (inclusive) in O(T).

V. Architecture Of The Ontology Model

The proposed ontology model aims to discover user background knowledge and learns personalized ontologies to represent user profiles. Fig. 6 illustrates the architecture of the ontology model. A personalized ontology is constructed, according to a given topic. Two knowledge resources, the global world knowledge base and the user's local instance repository, are utilized by the model. The world knowledge base provides the taxonomic structure for the personalized ontology. The user background knowledge is discovered from the user local instance repository. Against the given topic, the specificity and exhaustivity of subjects are investigated for user background knowledge discovery.

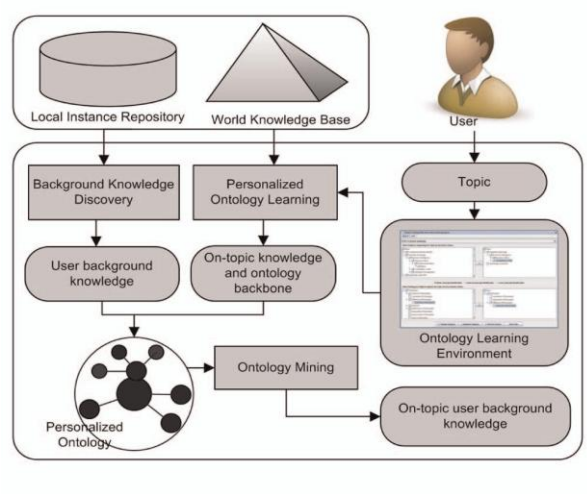


Fig 6. Architecture of the ontology model.

VI. Evaluation

A. Experiment Design

The proposed ontology model was evaluated by objective experiments. Because it is difficult to compare two sets of knowledge in different representations, the principal design of the evaluation was to compare the effectiveness of an information gathering system (IGS) that used different sets of user background knowledge for information gathering. The knowledge discovered by the ontology model was first used for a run of information gathering, and then the knowledge manually specified by users was used for another run. The latter run set up a benchmark for the evaluation because the knowledge was manually specified by users. Under the same experimental conditions, if the IGS could achieve the same (or similar) performance in two different runs, we could prove that the discovered knowledge has the same quality as the user specified knowledge. The proposed ontology model could then be proven promising to the domain of web information gathering.

In information gathering evaluations, a common batchstyle experiment is developed for the comparison of different models, using a test set and a set of topics associated with relevant judgments [36]. Our experiments followed this style and were performed under the experimental environment set up by the TREC-11 Filtering Track.3 This track aimed to evaluate the methods of persistent user

profiles for separating relevant and nonrelevant documents in an incoming stream [32].

User background knowledge in the experiments was represented by user profiles, such as those in the experiments of [23] and the TREC-11 Filtering Track. A user profile consisted of two document sets: a positive document set D+ containing the on-topic, interesting knowledge, and a negative document set D- containing the paradoxical, ambiguous concepts. Each document d held a support value support(d) to the given topic. Based on this representation, the baseline models in our experiments were carefully selected.

User profiles can be categorized into three groups: interviewing, semi-interviewing, and non-interviewing profiles, as previously discussed in Section 2. In an attempt to compare the proposed ontology model to the typical models representing these three group user profiles, four models were implemented in the experiments:

1. The Ontology model that implemented the proposed ontology model. User background knowledge was computationally discovered in this model.
2. The TREC model that represented the perfect interviewing user profiles. User background knowledge was manually specified by users in this model.
3. The Category model that represented the non-interviewing user profiles.
4. The Web model that represented the semi-interviewing user profiles.

The experiment dataflow is illustrated in Fig. 7. The topics were distributed among four models, and different user profiles were acquired. The user profiles were used by a common web information gathering system, the IGS, to gather information from the testing set. Because the user profiles were the only difference made by the experimental models to the IGS, the change of IGS performance reflected the effectiveness of user profiles, and thus, the performance of experimental models. The details of the experiment design are given as follows:

The TREC-11 Filtering Track testing set and topics were used in our experiments. The testing set was the Reuters Corpus Volume 1 (RCV1) corpus [21] that contains 806,791 documents and covers a great range of topics. This corpus consists of a training set and a testing set partitioned by the TREC. The documents in the corpus have been processed by substantial verification and validation of the content, attempting to remove spurious or duplicated documents, normalization of dateline and byline formats, addition of copyright statements, and so on. We have also further processed these documents by removing the stopwords, and stemming and grouping the terms.

In the experiments, we attempted to evaluate the proposed

model in an environment covering a great range of topics.

However, it is difficult to obtain an adequate number of users who have a great range of topics in their background knowledge. The TREC-11 Filtering Track provided a set of 50 topics specifically designed manually by linguists, covering various domains and topics [32]. For these topics, we assumed that each one came from an individual user. With this, we simulated 50 different users in our experiments. Buckley and Voorhees [3] stated that 50 topics are substantial to make a benchmark for stable evaluations in information gathering experiments. Therefore, the 50 topics used in our experiments also ensured high stability in the evaluation.

Each topic has a title, a description, and a narrative, provided by the topic author. In the experiments, only the titles of topics were used, based on the assumption that in the real world users often have only a small number of terms in their queries [15].

B. Web Information Gathering System

The information gathering system, IGS, was designed for common use by all experimental models. The IGS was an implementation of a model developed by Li and Zhong [23] that uses user profiles for web information gathering. The input support values associated with the documents in user profiles affected the IGS's performance acutely. Li and Zhong's model was chosen since not only is it better verified than the Rocchio and Dempster-Shafer models, but it is also extensible in using support values of training documents for web information gathering.

C. Proposed Model: Ontology Model

This model was the implementation of the proposed ontology model. As shown in Fig. 7, the input to this model was a topic and the output was a user profile consisting of positive documents (D+) and negative documents (D-). Each document d was associated with a support(d) value indicating its support level to the topic.

The WKB was constructed based on the LCSH system, as introduced in Section 3.1. The LCSH authority records distributed by the Library of Congress were a single file of 130 MB compiled in MAchine-Readable Cataloging (MARC) 21 format. After data preprocessing using expression techniques, these records were translated to human-readable form and organized in an SQL database, approximately 750 MB in size. Theoretically, the LCSH authority records consisted of subjects for personal names, corporate names, meeting names, uniform titles, bibliographic titles, topical terms, and geographic names. In order to make the Ontology model run more efficiently, only the topical, corporate, and geographic subjects were kept in the WKB, as they covered most topics in daily life. The BT, UF, and RT references (referred to by "450 | w | a", "450," and "550" in the records, respectively) linking the subjects in the LCSH thesaurus, were also extracted and encoded as the semantic relations of is-a, part-of, and related-to in the WKB, respectively. Eventually, the constructed WKB contained 394,070 subjects covering a wide range of topics linked by semantic relations.

The user personalized ontologies were constructed as described in Section 3.2 via user interaction. The authors played the user role to select positive and negative subjects for ontology construction, following the descriptions and narratives associated with the topics. On average, each personalized ontology contained about 16 positive and 23 negative subjects.

For each topic T, the ontology mining method was performed on the constructed O(T) and the user LIR to discover interesting concepts, as discussed in Section 4. The user LIRs were collected through searching the subject catalog of the QUT library by using the given topics. The catalog was distributed by QUT library as a 138 MB text file containing information for 448,590 items. The information was preprocessed by removing the stopwords, and stemming and grouping the terms. Librarians and authors have assigned title, table of content, summary, and a list of subjects to each information item in the catalog. These were used to represent the instances in LIRs. For each one of the 50 experimental topics, and thus, each one of the 50 corresponding users, the user's LIR was extracted from this catalog data set. As a result, there were about 1,111 instances existing in one LIR on average.

The semantic relations of is-a and part-of were also analyzed in the ontology mining phase for interesting knowledge discovery. For the coefficient _ in Algorithm 1, some preliminary tests had been conducted for various values (0.5, 0.7, 0.8, and 0.9). As a result, _ ¼ 0:9 gave the testing model the best performance and was chosen in the experiments.

Finally, a document d in the user profile was generated from an instance i in the LIR. The d held a support value support(d) to the T , which was measured by

$$support(d_i) = str(i, \mathcal{T}) \times \sum_{s \in \eta(i)} spe(s, \mathcal{T}),$$

where s E S of O(T), str(i,T) was defined by (4), and spe(s,T) by (6). When conducting the experiments, we tested various thresholds of support(d) to classify positive and negative documents. However, because the constructed ontologies were personalized and focused on various topics, we could not find a universal threshold that worked for all topics. Therefore, we set the threshold as support(d)=0, following the nature of positive and negative defined in this paper. The documents with support(d) > 0 formed D+, and those with negative support(d) ≤ 0 formed D- eventually.

D. Golden Model: TREC Model

The TREC model was used to demonstrate the interviewing user profiles, which reflected user concept models perfectly. As previously described, the RCV1 data set consisted of a training set and a testing set. The 50 topics were designed manually by linguists and associated with

positive and negative training documents in the RCV1 set [32]. These training documents formed the user profiles in the TREC model. For each topic, TREC users were given a set of documents to read and judged each as relevant or nonrelevant to the topic. If a document d was judged relevant, it became a positive document in the TREC user profile and support(d)= $\frac{1}{|D^+|}$ otherwise, it became a negative document and support(d)=0. The TREC user profiles perfectly reflected the users' personal interests, as the relevant judgments were provided by the same people who created the topics as well, following the fact that only users know their interests and preferences perfectly. Hence, the TREC model was the golden model for our proposed model to be measured against. The modeling of a user's concept model could be proven if our proposed model achieved the same or similar performance to the TREC model.

E. Baseline Model: Category Model

This model demonstrated the non-interviewing user profiles, in particular Gauch et al.'s OBIWAN [12] model. In the OBIWAN model, a user's interests and preferences are described by a set of weighted subjects learned from the user's browsing history. These subjects are specified with the semantic relations of superclass and subclass in an ontology. When an OBIWAN agent receives the search results for a given topic, it filters and re-ranks the results based on their semantic similarity with the subjects. The similar documents are awarded and reranked higher on the result list. In this Category model, the sets of positive subjects were manually fed back by the user via the OLE and from the WKB, using the same process as that in the Ontology model. The Category model differed from the Ontology model in that there were no is-a, part-of, and related-to knowledge considered and no ontology mining performed in the model. The positive subjects were equally weighted as one, because there was no evidence to show that a user might prefer some positive subjects more than others. The training sets in this model were extracted through searching the subject catalog of the QUT library, using the same process as in the Ontology model for user LIRs. However, in this model, a document's support(d) value was determined by the number of positive subjects cited by d. Thus, more positive subjects cited by d would give the document a stronger support(d) value. There was no negative training set generated by this model, as it was not required by the OBIWAN model.

F. Baseline Model: Web Model

The web model was the implementation of typical semi-interviewing user profiles. It acquired user profiles from the web by employing a web search engine. For a given topic, a set of feature terms {t|t ∈ T+} and a set of noisy terms {t|t ∈ T-} were first manually identified. The feature terms referred to the interesting concepts of the topic. The noisy terms referred to the paradoxical or ambiguous concepts. Also identified were the certainty factors CF(t) of the terms that determined their supporting rates ([-1, 1]) to the topic.

By using the feature and noisy terms, the Google4 API was employed to perform two searches for the given topic. The first search used a query generated by adding "+" symbols in front of the feature terms and "-" symbols in front of the noisy terms. By using this query, about 100 URLs were retrieved for the positive training set. The second search used a query generated by adding "-" symbols in front of feature terms and "+" symbols in front of noisy terms. Also, about 100 URLs were retrieved for the negative set.

These positive and negative documents were filtered by their certainty degrees CD. The CD(d) was determined by the docum ent's index ind(d) on the returned list from Google and Google's precision rate $\wp$. The $\wp$ was set as 0.9, based on the preliminary test results using experimental topics.

VII. RESULTS AND DISCUSSIONS

The experiments were designed to compare the information gathering performance achieved by using the proposed (Ontology) model, to that achieved by using the golden (TREC) and baseline (web and Category) models.

A. Experimental Results

The performance of the experimental models was measured by three methods: the precision averages at 11 standard recall levels (11SPR), the mean average precision (MAP), and the F1 Measure. These are modern methods based on precision and recall, the standard methods for information gathering evaluation [1], [3]. Precision is the ability of a system to retrieve only relevant documents. Recall is the ability to retrieve all relevant documents. An 11SPR value is computed by summing the interpolated precisions at the specified recall cutoff, and then dividing by the number of topics:

$$\frac{\sum_{i=1}^{N} precision_{\lambda}}{N};\ \ \lambda = \{0.0, 0.1, 0.2, \ldots, 1.0\},$$

where N denotes the number of topics, and λ= indicates the cutoff points where the precisions are interpolated. At each λ point, an average precision value over N topics is calculated. These average precisions then link to a curve describing the recall-precision performance. The experimental 11SPR results are plotted in Fig. 8, where the 11SPR curves show that the Ontology model was the best, followed by the TREC model, the web model, and finally, the Category model.

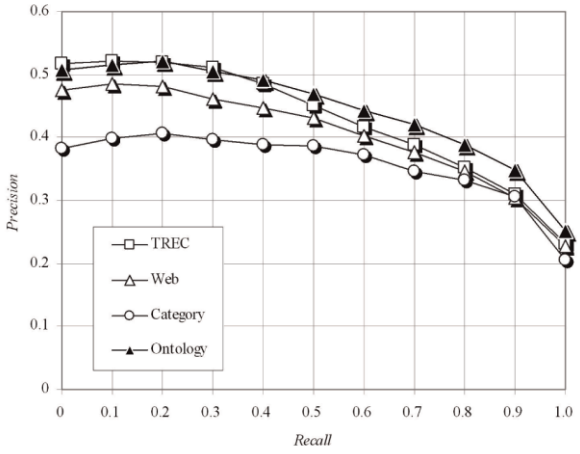


Fig 8. The 11SPR experimental results.

The MAP is a discriminating choice and recommended for general-purpose information gathering evaluation [3]. The

average precision for each topic is the mean of the precision obtained after each relevant document is retrieved. The MAP for the 50 experimental topics is then the mean of the average precision scores of each of the individual topics in the experiments. Different from the 11SPR measure, the MAP reflects the performance in a non-interpolated recall-precision curve. The experimental MAP results are presented in Table 2. As shown in this table, the TREC model was the best, followed by the Ontology model, and then the web and the Category models.

VIII. Conclusions And Future Work

In this paper, an ontology model is proposed for representing user background knowledge for personalized web information gathering. The model constructs user personalized ontologies by extracting world knowledge from the LCSH system and discovering user background knowledge from user local instance repositories. A multidimensional ontology mining method, exhaustivity and specificity, is also introduced for user background knowledge discovery. In evaluation, the standard topics and a large testbed were used for experiments. The model was compared against benchmark models by applying it to a common system for information gathering. The experiment results demonstrate that our proposed model is promising. A sensitivity analysis was also conducted for the ontology model. In this investigation, we found that the combination of global and local knowledge works better than using any one of them. In addition, the ontology model using knowledge with both is-a and part-of semantic relations works better than using only one of them. When using only global knowledge, these two kinds of relations have the same contributions to the performance of the ontology model. While using both global and local knowledge, the knowledge with part-of relations is more important than that with is-a. The proposed ontology model in this paper provides a solution to emphasizing global and local knowledge in a single computational model. The findings in this paper can be applied to the design of web information gathering systems. The model also has extensive contributions to the fields of Information Retrieval, web Intelligence, Recommendation Systems, and Information Systems.

In our future work, we will investigate the methods that generate user local instance repositories to match the representation of a global knowledge base. The present work assumes that all user local instance repositories have content-based descriptors referring to the subjects; however, a large volume of documents existing on the web may not have such content-based descriptors. For this problem, in Section 4.2, strategies like ontology mapping and text classification/clustering were suggested. These strategies will be investigated in future work to solve this problem.

The investigation will extend the applicability of the ontology model to the majority of the existing web documents and increase the contribution and significance of the present work.

Acknowledgment

This paper presents the extensive work of, but significantly beyond, an earlier research works. The authors also thank the anonymous reviewers for their valuable comments.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.
[2] G.E.P. Box, J.S. Hunter, and W.G. Hunter, Statistics For Experimenters. John Wiley & Sons, 2005.
[3] C. Buckley and E.M. Voorhees, "Evaluating Evaluation Measure Stability," Proc. ACM SIGIR '00, pp. 33-40, 2000.
[4] Z. Cai, D.S. McNamara, M. Louwerse, X. Hu, M. Rowe, and A.C. Graesser, "NLS: A Non-Latent Similarity Algorithm," Proc. 26th Ann. Meeting of the Cognitive Science Soc. (CogSci '04), pp. 180-185, 2004.
[5] L.M. Chan, Library of Congress Subject Headings: Principle and Application. Libraries Unlimited, 2005.
[6] P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," Proc. ACM SIGIR ('07), pp. 7-14, 2007.
[7] R.M. Colomb, Information Spaces: The Architecture of Cyberspace. Springer, 2002.
[8] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to Map between Ontologies on the Semantic Web," Proc. 11th Int'l Conf. World Wide Web (WWW '02), pp. 662-673, 2002.
[9] D. Dou, G. Frishkoff, J. Rong, R. Frank, A. Malony, and D. Tucker, "Development of Neuroelectromagnetic Ontologies(NEMO): A Framework for Mining Brainwave Ontologies," Proc. ACM SIGKDD ('07), pp. 270-279, 2007.
[10] D. Downey, S. Dumais, D. Liebling, and E. Horvitz, "Understanding the Relationship between Searchers' Queries and Information Goals," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 449-458, 2008.
[11] E. Frank and G.W. Paynter, "Predicting Library of Congress Classifications from Library of Congress Subject Headings," J. Am. Soc. Information Science and Technology, vol. 55, no. 3, pp. 214-227, 2004.
[12] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.
[13] R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen, "Using Google Distance to Weight Approximate Ontology Matches," Proc. 16th Int'l Conf. World Wide Web (WWW '07), pp. 767-776, 2007.
[14] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[15] B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web," ACM SIGIR Forum, vol. 32, no. 1, pp. 5-17, 1998.
[16] X. Jiang and A.-H. Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents," Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM '05), pp. 665-668, 2005.
[17] W. Jin, R.K. Srihari, H.H. Ho, and X. Wu, "Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 193-202, 2007.
[18] J.D. King, Y. Li, X. Tao, and R. Nayak, "Mining World Knowledge for Analysis of Search Engine Content," Web Intelligence and Agent Systems, vol. 5, no. 3, pp. 233-253, 2007.
[19] R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, and J.X. Hao, "Towards a Fuzzy Domain Ontology Extraction Method for Adaptive e- Learning," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 6, pp. 800-813, June 2009.
[20] K.S. Lee, W.B. Croft, and J. Allan, "A Cluster-Based Resampling Method for Pseudo-Relevance Feedback," Proc. ACM SIGIR '08, pp. 235-242, 2008.
[21] D.D. Lewis, Y. Yang, T.G. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," J. Machine Learning Research, vol. 5, pp. 361-397, 2004.
[22] Y. Li and N. Zhong, "Web Mining Model and Its Applications for Information Gathering," Knowledge-Based Systems, vol. 17, pp. 207-217, 2004.
[23] Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.
[24] C. Makris, Y. Panagis, E. Sakkopoulos, and A. Tsakalidis, "Category Ranking for Personalized Search," Data and Knowledge Eng., vol. 60, no. 1, pp. 109-125, 2007.
[25] S.E. Middleton, N.R. Shadbolt, and D.C. De Roure, "Ontological User Profiling in Recommender Systems," ACM Trans. Information Systems (TOIS), vol. 22, no. 1, pp. 54-88, 2004.

[26] G.A. Miller and F. Hristea, "WordNet Nouns: Classes and Instances," Computational Linguistics, vol. 32, no. 1, pp. 1-3, 2006.
[27] D.N. Milne, I.H. Witten, and D.M. Nichols, "A Knowledge-Based Search Engine Powered by Wikipedia," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 445-454, 2007.
[28] R. Navigli, P. Velardi, and A. Gangemi, "Ontology Learning and Its Application to Automated Terminology Translation," IEEE Intelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.
[29] S. Nirenburg and V. Rasin, Ontological Semantics. The MIT Press, 2004.
[30] A.-M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05), pp. 339-346, 2005.
[31] D. Quest and H. Ali, "Ontology Specific Data Mining Based on Dynamic Grammars," Proc. IEEE Computational Systems Bioinformatics Conf. (CSB '04), pp. 495-496, 2004.
[32] S.E. Robertson and I. Soboroff, "The TREC 2002 Filtering Track Report," Proc. Text REtrieval Conf., 2002.
[33] S. Sekine and H. Suzuki, "Acquiring Ontological Knowledge from Query Logs," Proc. 16th Int'l Conf. World Wide Web (WWW '07), pp. 1223-1224, 2007.
[34] S. Shehata, F. Karray, and M. Kamel, "Enhancing Search Engine Quality Using Concept-Based Text Retrieval," Proc. IEEE/WIC/ ACM Int'l Conf. Web Intelligence (WI '07), pp. 26-32, 2007.
[35] A. Sieg, B. Mobasher, and R. Burke, "Web Search Personalization with Ontological User Profiles," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 525-534, 2007.
[36] M.D. Smucker, J. Allan, and B. Carterette, "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 623-632, 2007.
[37] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW '04), pp. 675- 684, 2004.
[38] X. Tao, Y. Li, N. Zhong, and R. Nayak, "Automatic Acquiring Training Sets for Web Information Gathering," Proc. IEEE/WIC/ ACM Int'l Conf. Web Intelligence, pp. 532-535, 2006.
[39] X. Tao, Y. Li, N. Zhong, and R. Nayak, "Ontology Mining for Personalized Web Information Gathering," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 351-358, 2007.
[40] J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities," Proc. ACM SIGIR '05, pp. 449-456, 2005.
[41] J. Trajkova and S. Gauch, "Improving Ontology-Based User Profiles," Proc. Conf. Recherche d'Information Assistee par Ordinateur (RIAO '04), pp. 380-389, 2004.
[42] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-Based Interpretation of Keywords for Semantic Search," Proc. Sixth Int'l Semantic Web and Second Asian Semantic Web Conf. (ISWC '07/ ASWC '07), pp. 523-536, 2007.
[43] K. van der Sluijs and G.J. Huben, "Towards a Generic User Model Component," Proc. Workshop Personalization on the Semantic Web (PerSWeb '05), 10th Int'l Conf. User Modeling (UM '05), pp. 43-52, 2005.
[44] E.M. Voorhees and Y. Hou, "Vector Expansion in a Large Collection," Proc. First Text REtrieval Conf., pp. 343-351, 1993.
[45] J. Wang and M.C. Lee, "Reconstructing DDC for Interactive Classification," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 137-146, 2007.
[46] L.A. Zadeh, "Web Intelligence and World Knowledge—The Concept of Web IQ (WIQ)," Proc. IEEE Ann. Meeting of the North American Fuzzy Information Soc. (NAFIPS '04), vol. 1, pp. 1-3, 2004.
[47] N. Zhong, "Representation and Construction of Ontologies for Web Intelligence," Int'l J. Foundation of Computer Science, vol. 13, no. 4, pp. 555-570, 2002.
[48] N. Zhong, "Toward Web Intelligence," Proc. First Int'l Atlantic Web Intelligence Conf., pp. 1-14, 2003.
[49] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering Personally Meaningful Places: An Interactive Clustering

Krishna Madhav Majety pursuing M.Tech(CSE) II Year in Sri Mittapalli College of Engineering, Affiliated to Jawaharlal Nehru Technological University Kakinada. Interested in the areas of Data Mining, Ontology Systems, Web & Mobile Technologies.

Uma Devi Deva (P.hD), M.Tech(CSE), Head of the Department, Computer Science & Engineering Department, Sri Mittapalli College of Engineering, Affiliated to Jawaharlal Nehru Technological University Kakinada. Research areas include Data Mining, Safety & Critical System Analysis, Software Engineering,

Sailaja Sanaka M.Tech(CSE), Assistant Professor, SRK Institute of Technology, Affiliated to Jawaharlal Nehru Technological University Kakinada. Research areas include Data Mining, Database Optimizations.