

A Personalized Ontology Model for Web Information Gathering

Xiaohui Tao, Yuefeng Li, and Ning Zhong, *Senior Member, IEEE*

Abstract—As a model for knowledge description and formalization, ontologies are widely used to represent user profiles in personalized web information gathering. However, when representing user profiles, many models have utilized only knowledge from either a global knowledge base or a user local information. In this paper, a personalized ontology model is proposed for knowledge representation and reasoning over user profiles. This model learns ontological user profiles from both a world knowledge base and user local instance repositories. The ontology model is evaluated by comparing it against benchmark models in web information gathering. The results show that this ontology model is successful.

Index Terms—Ontology, personalization, semantic relations, world knowledge, local instance repository, user profiles, web information gathering.

1 INTRODUCTION

ON the last decades, the amount of web-based information available has increased dramatically. How to gather useful information from the web has become a challenging issue for users. 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 users' 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

• X. Tao and Y. Li are with the Computer Science Discipline, Faculty of Science and Technology, Queensland University of Technology (QUT), GPO Box 2434, Brisbane Qld 4001, Australia.
Email: {x.tao, y2.li}@qut.edu.au.

• N. Zhong is with the Knowledge Information Systems Laboratory, Department of Systems and Information Engineering, Maebashi Institute of Technology, 460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan.
Email: zhong@maebashi-it.ac.jp.

Manuscript received 21 Nov. 2008; revised 26 June 2009; accepted 30 Nov. 2009; published online 24 Aug. 2010.

Recommended for acceptance by C. Bettini.

For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-11-0613. Digital Object Identifier no. 10.1109/TKDE.2010.145.

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 *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.

The paper is organized as follows: Section 2 discusses the related work; in Section 3, we introduce how personalized ontologies are constructed for users; and in Section 4, we present the multidimensional ontology mining method. After that, Section 5 gives the architecture of the proposed model; Section 6 discusses the evaluation issues, and the results are analyzed in Section 7. Finally, Section 8 makes conclusions and addresses our future work.

2 RELATED WORK

2.1 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 lead 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.

2.2 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 *noninterviewing*. 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 semiautomated techniques with limited user involvement. These techniques usually provide users with a list of categories and ask users for interesting or

noninteresting 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 feedback categories. Noninterviewing 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 noninterviewing user profiles can also be viewed as manual, semiautomatic, and automatic profiles, respectively.

3 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.

3.1 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 coreference 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]. In addition, the LCSH is the most comprehensive nonspecialized 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].

TABLE 1
Comparison of Different World Taxonomies

	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

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:

Definition 1. Let \mathbb{S} be a set of subjects, an element $s \in \mathbb{S}$ is formalized as a 4-tuple $s := \langle \text{label}, \text{neighbor}, \text{ancestor}, \text{descendant} \rangle$, 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;

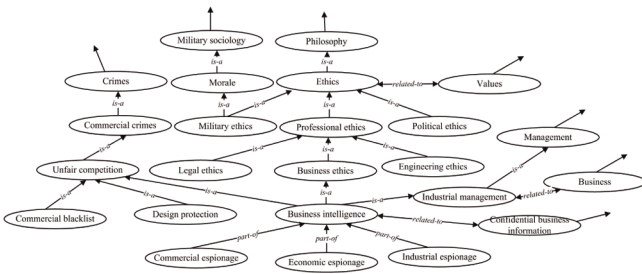


Fig. 1. A sample part of the world knowledge base.

- 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 \mathbb{R} be a set of relations, an element $r \in \mathbb{R}$ is a 2-tuple $r := \langle edge, type \rangle$, where

- an edge connects two subjects 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 $WKB := \langle \mathbb{S}, \mathbb{R} \rangle$, where

- \mathbb{S} is a set of subjects $\mathbb{S} := \{s_1, s_2, \dots, s_m\}$;
- \mathbb{R} is a set of semantic relations $\mathbb{R} := \{r_1, r_2, \dots, r_n\}$ linking the subjects in \mathbb{S} .

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.)

3.2 Ontology Construction

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 $s \in \mathbb{S}$, the s and its ancestors are

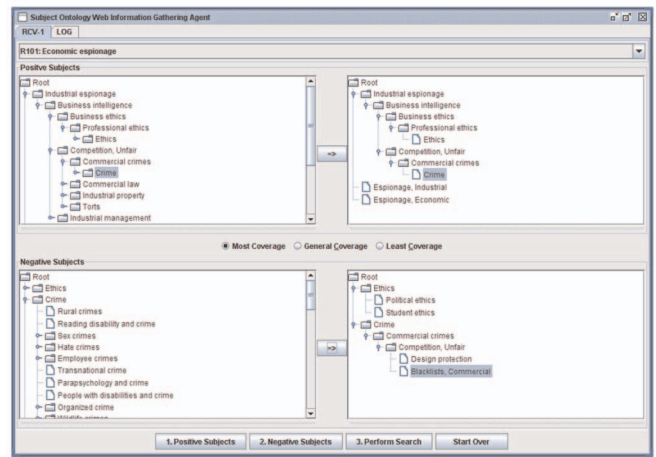


Fig. 2. Ontology learning environment.

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.

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

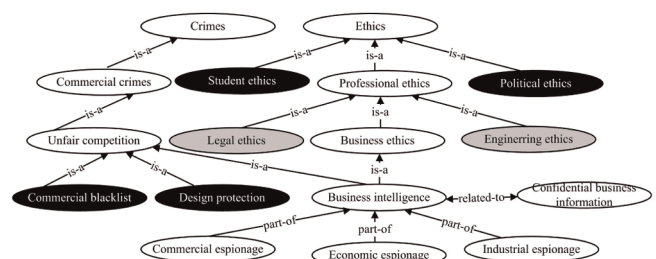


Fig. 3. An ontology (partial) constructed for topic “Economic Espionage.”

nodes are neutral subjects. Here, we formalize the ontology constructed for a given topic:

Definition 4. *The structure of an ontology that describes and specifies topic \mathcal{T} is a graph consisting of a set of subject nodes. The structure can be formalized as a 3-tuple $\mathcal{O}(\mathcal{T}) := \langle \mathcal{S}, tax^S, rel \rangle$, where*

- \mathcal{S} is a set of subjects consisting of three subsets \mathcal{S}^+ , \mathcal{S}^- , and \mathcal{S}° , where \mathcal{S}^+ is a set of positive subjects regarding \mathcal{T} , $\mathcal{S}^- \subseteq \mathcal{S}$ is negative, and $\mathcal{S}^\circ \subseteq \mathcal{S}$ is neutral;
- tax^S is the taxonomic structure of $\mathcal{O}(\mathcal{T})$, which is a noncyclic and directed graph $(\mathcal{S}, \mathcal{E})$. For each edge $e \in \mathcal{E}$ and $type(e) = is-a$ or $part-of$, iff $\langle s_1 \rightarrow s_2 \rangle \in \mathcal{E}$, $tax(s_1 \rightarrow s_2) = True$ means s_1 is-a or is a part-of s_2 ;
- rel is a boolean function defining the related-to relationship held by two subjects in \mathcal{S} .

The constructed ontology is personalized because the user selects positive and negative subjects for personal preferences and interests. Thus, if a user searches "New York" and plans for a business trip, the user would have different subjects selected and a different ontology constructed, compared to those selected and constructed by a leisure user planning for a holiday.

4 MULTIDIMENSIONAL 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.

We argue that a subject's specificity has two focuses: 1) on the referring-to concepts (called semantic specificity), and 2) on the given topic (called topic specificity). These need to be addressed separately.

4.1 Semantic Specificity

The semantic specificity is investigated based on the structure of $\mathcal{O}(\mathcal{T})$ inherited from the world knowledge base. The strength of such a focus is influenced by the subject's locality in the taxonomic structure tax^S of $\mathcal{O}(\mathcal{T})$ (this is also argued by [42]). As stated in Definition 4, the tax^S of $\mathcal{O}(\mathcal{T})$ 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 tax^S .¹ Because subjects have a fixed locality on the tax^S of $\mathcal{O}(\mathcal{T})$, semantic specificity is also called absolute specificity and denoted by $spe_a(s)$.

The determination of a subject's spe_a is described in Algorithm 1. The $isA(s')$ and $partOf(s')$ are two functions in the algorithm satisfying $isA(s') \cap partOf(s') = \emptyset$. The $isA(s')$ returns a set of subjects $s \in tax^S$ that satisfy $tax(s \rightarrow s') = True$ and $type(s \rightarrow s') = is-a$. The $partOf(s')$ returns a set of subjects $s \in tax^S$ that satisfy $tax(s \rightarrow s') = True$ and $type(s \rightarrow s') = part-of$. Algorithm 1 is efficient with the complexity of only $O(n)$, where $n = |\mathcal{S}|$. The algorithm terminates eventually because tax^S is a directed acyclic graph, as defined in Definition 4.

Algorithm 1. Analyzing semantic relations for specificity

```

input : a personalized ontology  $\mathcal{O}(\mathcal{T}) := \langle tax^S, rel \rangle$ ; a
        coefficient  $\theta$  between (0,1).
output:  $spe_a(s)$  applied to specificity.
1 set  $k = 1$ , get the set of leaves  $S_0$  from  $tax^S$ , 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  $tax^S$ ;
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.

```

As the tax^S of $\mathcal{O}(\mathcal{T})$ is a graphic taxonomy, the leaf subjects have no descendants. Thus, they have the strongest focus on their referring-to concepts and the highest $spe_a(s)$. By setting the spe_a range as (0, 1] (greater than 0, less than or equal to 1), the leaf subjects have the strongest $spe_a(s)$ of 1, and the root subject of tax^S has the weakest $spe_a(s)$ and the smallest value in (0, 1]. Toward the root of tax^S , the $spe_a(s)$ decreases for each level up. A coefficient θ is applied to the $spe_a(s)$ analysis, defining the decreasing rate of semantic specificity from lower bound toward upper bound levels. ($\theta = 0.9$ was used in the related experiments presented in this paper.)

From the leaf subjects toward upper bound levels in tax^S , if a subject has *is-a* child subjects, it has no greater semantic specificity compared with any one of its *is-a* child subjects. In *is-a* relationships, a parent subject is the abstract description of its child subjects. However, the abstraction sacrifices the focus and specificity of the referring-to concepts. Thus, we define the $spe_a(s)$ value of a parent subject as the smallest $spe_a(s)$ of its *is-a* child subjects, applying the decreasing rate θ .

If a subject has *part-of* child subjects, the $spe_a(s)$ of all *part-of* child subjects takes part of their parent subject's semantic specificity. As a *part-of* relation, the concepts referred to by a parent subject are the combination of its *part-of* child subjects. Therefore, we define the parent's spe_a

1. In this analysis, the *related-to* semantic relations are not considered because they are nontaxonomic. In this paper, we assume they have no influence on each other in terms of *specificity*. However, this is an interesting issue and will be pursued in our future work.

as the average spe_a value of its *part-of* child subjects applying θ .

If a subject has direct child subjects mixed with *is-a* and *part-of* relationships, a spe_a^1 and a spe_a^2 are addressed separately with respect to the *is-a* and *part-of* child subjects. The approaches to calculate spe_a^1 and spe_a^2 are the same as described previously. Following the principle that specificity decreases for the subjects located toward upper bound levels, the smaller value of spe_a^1 or spe_a^2 is chosen for the parent subject.

In summary, the semantic specificity of a subject is measured, based on the investigation of subject locality in the taxonomic structure tax^S of $\mathcal{O}(T)$. In particular, the influence of locality comes from the subject's taxonomic semantic (*is-a* and *part-of*) relationships with other subjects.

4.2 Topic Specificity

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

4.2.1 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 $\mathcal{O}(T)$ 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 semistructured (e.g., the browsed HTML and XML web documents) or unstructured (e.g., the stored local DOC and TXT documents). In some semistructured 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.

However, many documents do not have such direct, clear references. For such documents in *LIRs*, data mining techniques, clustering, and classification in particular, can help to establish the reference, as in the work conducted by [20], [49]. The clustering techniques group the documents into unsupervised (nonpredefined) clusters based on the document features. These features, usually represented by terms, can be extracted from the clusters. They represent the user background knowledge discovered from the user *LIR*. By measuring the semantic similarity between these features and the subjects in $\mathcal{O}(T)$, the references of these clustered documents to the subjects in $\mathcal{O}(T)$ can be established and the strength of each reference can be scaled by using methods like Nonlatent Similarity [4]. The


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 LDCN	CALL NO	STATUS
Carseldine	306.30943:109045 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.

documents with a strong reference to the subjects in $\mathcal{O}(T)$ can then be used to populate these subjects.

Classification is another strategy to map the unstructured/semistructured documents in user *LIRs* to the representation in the global knowledge base. By using the subject labels as the feature terms, we can measure the semantic similarity between a document in the *LIR* and the subjects in $\mathcal{O}(T)$. The documents can then be classified into the subjects based on their similarity, and become the instances populating the subjects they belong to. Ontology mapping techniques can also be used to map the features discovered by using clustering and classification to the subjects in $\mathcal{O}(T)$, if they are in different representations.

Because ontology mapping and text classification/clustering are beyond the scope of the work presented in this paper, we assume the existence of an ideal user *LIR*. The documents in the user *LIR* have content-related descriptors referring to the subjects in $\mathcal{O}(T)$. In particular, we use the information items in the catalogs of the QUT library² as user *LIR* to populate the $\mathcal{O}(T)$ constructed from the *WKB* in the experiments.

The *WKB* is encoded from the LCSH, as discussed in Section 3.1. The LCSH contains the content-related descriptors (subjects) in controlled vocabularies. Corresponding to these descriptors, the catalogs of library collections also contain descriptive information of library-stored books and documents. Fig. 4 displays a sample information item used as an instance in an *LIR*. The descriptive information, such as the title, table of contents, and summary, is provided by authors and librarians. This expert classified and trustworthy information can be recognized as the extensive knowledge from the LCSH. A list of content-based descriptors (subjects) is also cited on the bottom of Fig. 4, indexed by their focus on the item's content. These subjects provide a connection between the extensive knowledge and the concepts formalized in the *WKB*. User background knowledge is to be discovered from both the user's *LIR* and $\mathcal{O}(T)$.

2. The Queensland University of Technology Library, <http://library.qut.edu.au>.

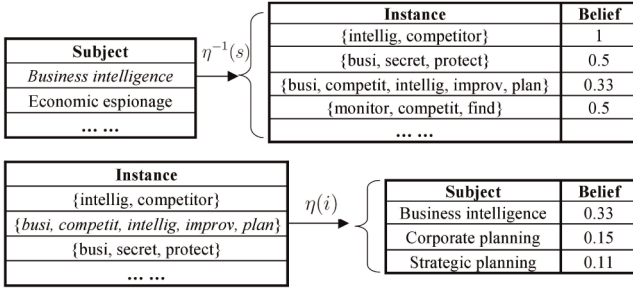


Fig. 5. Mappings of subjects and instances.

The reference strength between an instance and a subject needs to be evaluated. As mentioned previously, the subjects cited by an instance are indexed by their focus. Many subjects cited by an instance may mean loose specificity of subjects, because each subject deals with only a part of the instance. Hence, denoting an instance by i , the strength of i to a subject s is determined by

$$str(i, s) = \frac{1}{priority(s, i) \times n(i)}; \quad (1)$$

where $n(i)$ is the number of subjects on the citing list of i and $priority(s, i)$ is the index (starting with one) of s on the citing list. The $str(i, s)$ aims to select the right instances to populate $\mathcal{O}(T)$.

With the $str(i, s)$ determined, the relationship between an \mathcal{LIR} and $\mathcal{O}(T)$ can be defined. Let $\Omega = \{i_1, i_2, \dots, i_k\}$ be a finite and nonempty set of instances in an \mathcal{LIR} , and min_str be the minimal str value for filtering out the noisy pairs with weak strengths. Given an $i \in \Omega$, we can get a set of subjects using the following mapping:

$$\eta : \Omega \rightarrow 2^{\mathcal{S}}, \quad \eta(i) = \{s \in \mathcal{S} | str(i, s) \geq min_str\}. \quad (2)$$

The mapping function $\eta(i)$ describes the subjects cited by i . In order to classify instances, the reverse mapping η^{-1} of η can also be defined as

$$\eta^{-1} : \mathcal{S} \rightarrow 2^{\Omega}, \quad \eta^{-1}(s) = \{i \in \Omega | str(i, s) \geq min_str\}. \quad (3)$$

The mappings η and η^{-1} reveal the relationships between instances and subjects. Each i maps to a set of subjects in \mathcal{S} , and each s is cited by a set of instances in Ω . Each pair, (i, s) , is associated with a strength value defined by (1). Fig. 5 presents a sample mapping related to the topic "Business intelligence."

4.2.2 Evaluating Topic Specificity

From Definition 4, an $\mathcal{O}(T)$ contains a set of positive subjects, a set of negative subjects, and a set of neutral subjects, pertaining to a topic T . Based on the mapping of (2), if an instance refers only to positive subjects, the instance fully supports the T . If it refers only to negative subjects, it is strongly against the T . Hence, we can measure the strength of an instance to the T by utilizing (1) and (2):

$$str(i, T) = \sum_{s \in (\eta(i) \cap \mathcal{S}^+)} str(i, s) - \sum_{s \in (\eta(i) \cap \mathcal{S}^-)} str(i, s). \quad (4)$$

If $str(i, T) > 0$, i contains knowledge relevant to the T . Otherwise, i is against the T .

The topic specificity of a subject is evaluated based on the instance-topic strength of its citing instances. With respect to the absolute specificity, the topic specificity can also be called relative specificity and denoted by $spe_r(s, T, \mathcal{LIR})$. A subject's $spe_r(s, T, \mathcal{LIR})$ is calculated by

$$spe_r(s, T, \mathcal{LIR}) = \sum_{i \in \eta^{-1}(s)} str(i, T). \quad (5)$$

Because the $str(i, T)$ from (4) could be positive or negative values, the $spe_r(s, T, \mathcal{LIR})$ values from (5) could be positive or negative as well.

As discussed previously, a subject's specificity has two focuses: semantic specificity and topic specificity. Therefore, the final specificity of a subject is a composition of them and calculated by

$$spe(s, T) = spe_a(s) \times spe_r(s, T, \mathcal{LIR}). \quad (6)$$

Based on (6), the lower bound subjects in the ontology would receive greater specificity values, as well as those cited by more positive instances.

4.3 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 $\mathcal{O}(T)$; we evaluate a subject's exhaustivity by aggregating the semantic specificity of its descendants:

$$exh(s, T) = \sum_{s' \in desc(s)} \sum_{i \in \eta^{-1}(s')} str(i, T) \times spe_a(s', T). \quad (7)$$

Subjects are considered interesting to the user only if their specificity and exhaustivity are positive. The subject sets of \mathcal{S}^+ , \mathcal{S}^- , and \mathcal{S}° , originally defined in Section 3.2, can be refined after ontology mining for the specificity and exhaustivity of subjects:

$$\mathcal{S}^+ = \{s | spe(s, T) > 0, exh(s, T) > 0, s \in \mathcal{S}\}; \quad (8)$$

$$\mathcal{S}^- = \{s | spe(s, T) < 0, exh(s, T) < 0, s \in \mathcal{S}\}; \quad (9)$$

$$\mathcal{S}^\circ = \{s | s \in (\mathcal{S} - (\mathcal{S}^+ \cup \mathcal{S}^-))\}. \quad (10)$$

A few theorems can be introduced, based on the subject analysis of specificity and exhaustivity.

Theorem 1. *A leaf subject in an ontology has the same value of specificity and exhaustivity.*

Proof 1. As s is a leaf subject, we have $desc(s) = \{s\}$, from (7), we have

$$\begin{aligned} exh(s, T) &= \sum_{s' \in desc(s)} \sum_{i \in \eta^{-1}(s')} str(i, T) \times spe_a(s', T) \\ &= spe_a(s', T) \times \sum_{i \in \eta^{-1}(s)} str(i, T) \\ &= spe_a(s', T) \times spe_r(s, T, \mathcal{LIR}) \\ &= spe(s, T). \end{aligned}$$

□

Theorem 2. Let s_1, s_2 be two different subjects in the S^+ of $\mathcal{O}(T)$, $s_1 \in desc(s_2)$, and $\eta^{-1}(s_1) = \eta^{-1}(s_2)$; we always have $spe(s_1, T) \geq spe(s_2, T)$.

Proof 2. From (5) and (6), we have

$$\begin{aligned} & spe(s_1, T) - spe(s_2, T) \\ &= spe_a(s_1) \times spe_r(s_1, T, \mathcal{LIR}) - spe_a(s_2) \times spe_r(s_2, T, \mathcal{LIR}) \\ &= spe_a(s_1) \times \sum_{i \in \eta^{-1}(s_1)} str(i, T) - spe_a(s_2) \times \sum_{i \in \eta^{-1}(s_2)} str(i, T) \\ &= (spe_a(s_1) - spe_a(s_2)) \times \sum_{i \in \eta^{-1}(s_1)} str(i, T) \end{aligned}$$

Because there exists a path from s_1 to s_2 :

$$s_1 \rightarrow s' \rightarrow \dots \rightarrow s'' \rightarrow s_2,$$

From Algorithm 1, we have

$$spe_a(s_1) \geq spe_a(s'), \dots, spe_a(s'') \geq spe_a(s_2);$$

Therefore $spe_a(s_1) \geq spe_a(s_2)$ and

$$spe(s_1, T) - spe(s_2, T) \geq 0.$$

□

Theorem 3. Let s_1, s_2 be two subjects in $\mathcal{O}(T)$, and $s_1 \in desc(s_2)$.

1. If $desc(s_2) \subseteq S^+$, we always have $exh(s_1, T) \leq exh(s_2, T)$;
2. If $desc(s_2) \subseteq S^-$, we always have $exh(s_1, T) \geq exh(s_2, T)$.

Proof 3. From (7), we have

$$\begin{aligned} & exh(s_2, T) - exh(s_1, T) \\ &= \sum_{s' \in desc(s_2)} \sum_{i \in \eta^{-1}(s')} str(i, T) \times spe_a(s', T) \\ &\quad - \sum_{s'' \in desc(s_1)} \sum_{i \in \eta^{-1}(s'')} str(i, T) \times spe_a(s'', T) \\ &= \sum_{s''' \in (desc(s_2) - desc(s_1))} \sum_{i \in \eta^{-1}(s''')} str(i, T) \times spe_a(s''', T) \\ &= \sum_{s''' \in (desc(s_2) - desc(s_1))} spe_r(s''', T, \mathcal{LIR}) \times spe_a(s''', T) \\ &= \sum_{s''' \in (desc(s_2) - desc(s_1))} spe(s''', T) \end{aligned}$$

Because from (8), for $\forall s''' \in desc(s_2)$

and $desc(s_2) \subseteq S^+ \Rightarrow spe(s''', T) > 0$

Therefore $exh(s_2, T) - exh(s_1, T) \geq 0$;

Analogically, from (9), for $\forall s''' \in desc(s_2)$

and $desc(s_2) \subseteq S^- \Rightarrow spe(s''', T) < 0$

Therefore $exh(s_2, T) - exh(s_1, T) \leq 0$, if $desc(s_2) \subseteq S^-$.

□

These theorems restrict the use of specificity and exhaustivity in ontology mining. Theorem 1 describes the leaf subjects in terms of specificity and exhaustivity.

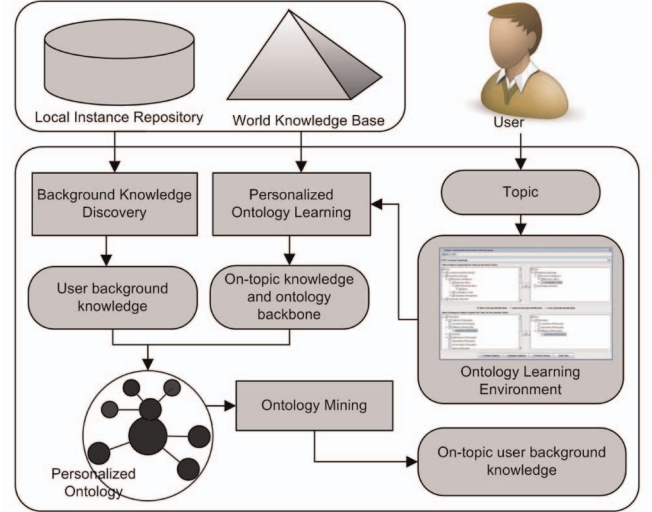


Fig. 6. Architecture of the ontology model.

Theorem 2 guarantees that, if two positive subjects hold the same strengths to T , the one at a lower level must be more specific than the other one. Theorem 3 constrains the influence of positive and negative subjects to exhaustivity. With respect to T , a subject in $\mathcal{O}(T)$ may be highly exhaustive but not specific. Similarly, a subject may be highly specific but may deal with only a limited semantic extent referred to by T .

5 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.

6 EVALUATION

6.1 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 batch-style 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 noninterviewing 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 noninterviewing 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

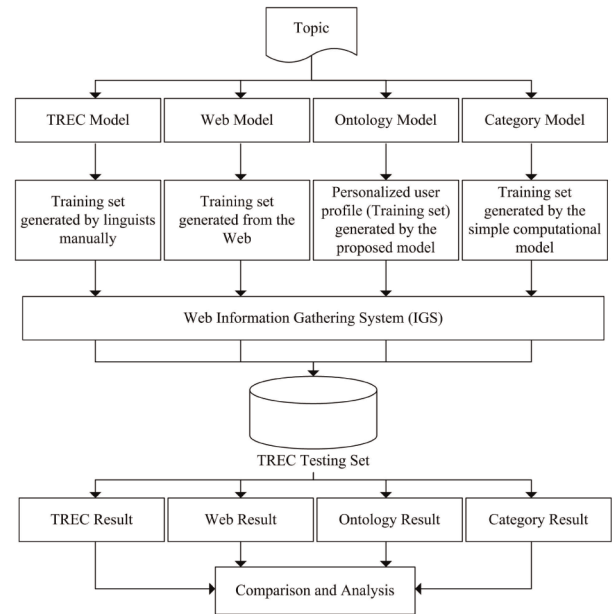


Fig. 7. Experiment design.

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 stop-words, 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].

6.2 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.

3. Text REtrieval Conference, <http://trec.nist.gov/>.

The IGS first used the training set to evaluate weights for a set of selected terms T . After text preprocessing of stopword removal and word stemming, a positive document d became a pattern that consisted of a set of term frequency pairs $d = \{(t_1, f_1), (t_2, f_2), \dots, (t_k, f_k)\}$, where f_i is t_i 's term frequency in d . The semantic space referred to by d was represented by its normal form $\beta(d)$, which satisfied $\beta(d) = \{(t_1, w_1), (t_2, w_2), \dots, (t_k, w_k)\}$, where w_i ($i = 1, \dots, k$) were the weight distribution of terms and

$$w_i = \frac{f_i}{\sum_{j=1}^k f_j}.$$

A probability function on T was derived based on the normal forms of positive documents and their supports for all $t \in T$:

$$pr_{\beta}(t) = \sum_{d \in D^+, (t,w) \in \beta(d)} support(d) \times w. \quad (11)$$

The testing documents were finally indexed by $weight(d)$, which was calculated using the probability function pr_{β} :

$$weight(d) = \sum_{t \in T} pr_{\beta}(t) \times \tau(t, d), \quad (12)$$

where $\tau(t, d) = 1$ if $t \in d$; otherwise $\tau(t, d) = 0$.

6.3 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 $\mathcal{O}(T)$ and the user \mathcal{LIR} to discover interesting concepts, as discussed in Section 4. The user \mathcal{LIR} s 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 \mathcal{LIR} s. For each one of the 50 experimental topics, and thus, each one of the 50 corresponding users, the user's \mathcal{LIR} was extracted from this catalog data set. As a result, there were about 1,111 instances existing in one \mathcal{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 θ in Algorithm 1, some preliminary tests had been conducted for various values (0.5, 0.7, 0.8, and 0.9). As a result, $\theta = 0.9$ 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 \mathcal{LIR} . The d held a support value $support(d)$ to the T , which was measured by

$$support(d_i) = str(i, T) \times \sum_{s \in \eta(i)} spe(s, T), \quad (13)$$

where $s \in \mathcal{S}$ of $\mathcal{O}(T)$, $str(i, T)$ 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 $support(d) \leq 0$ formed D^- eventually.

6.4 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.

6.5 Baseline Model: Category Model

This model demonstrated the noninterviewing 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 reranks 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 *LTRs*. 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.

6.6 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 $\{t|t \in T^+\}$ and a set of noisy terms $\{t|t \in T^-\}$ 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 document's index $ind(d)$ on the returned list from Google and Google's precision rate φ . The φ was set as 0.9, based on the preliminary test results using experimental

topics. If a document d was in the cutoff κ and $\varphi_\kappa = 0.9$, the Google's confidence on d would be 0.9. Together with the $CF(t)$ values of the feature terms and noisy terms, we had a $CD(d)$ calculated by

$$CD(d) = \varphi_\kappa \times \frac{K - ind(d)(mod \kappa) + 1}{K} \times \sum_{t \in (T \cap d)} |CF(t)|, \quad (14)$$

where K is a constant number of 10 for the number of documents in each cutoff κ , T refers to T^+ or T^- , depending on the positive or negative set that d is in.

The support value of a document was finally determined by $support(d) = CD^+(d) - CD^-(d)$. The positive training set was then generated by the documents with $support(d) > 0$, and the negative set by the documents with $support(d) \leq 0$.

7 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.

7.1 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 F_1 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:

$$\frac{\sum_{i=1}^N precision_\lambda}{N}; \quad \lambda = \{0.0, 0.1, 0.2, \dots, 1.0\}, \quad (15)$$

where N denotes the number of topics, and $\lambda =$ 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.

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 noninterpolated 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.

4. <http://www.google.com>.

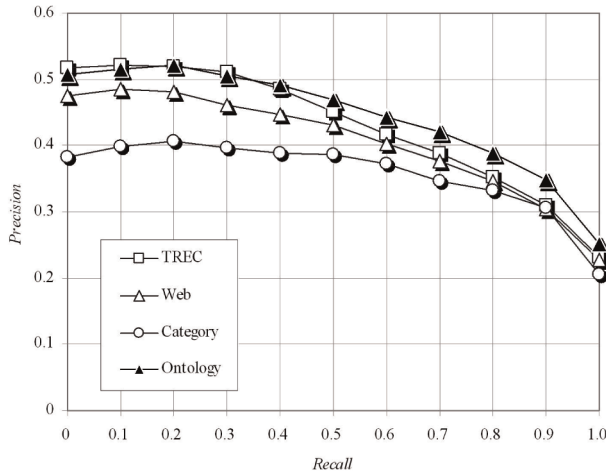


Fig. 8. The 11SPR experimental results.

Table 2 also presents the average macro- F_1 and micro- F_1 Measure results. The F_1 Measure is calculated by

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (16)$$

where precision and recall are evenly weighted. For each topic, the *macro- F_1* Measure averages the precision and recall and then calculates F_1 Measure, whereas the *micro- F_1* Measure calculates the F_1 Measure for each returned result and then averages the F_1 Measure values. The greater F_1 values indicate the better performance. According to the results, the Ontology model was the best, followed by the TREC model, and then the web and the Category models.

The statistical tests were also performed for the reliability of the evaluation. Usually, a reliable significance test concerns the difference in the mean of a measuring metric (e.g., MAP) and the significance level (e.g., p -value computed under a given *null hypothesis*) [2], [36]. Following this guide, we used the *percentage change in performance* and *Student's Paired T-Test* for the significance test.

The *percentage change in performance* is used to compute the difference in MAP and F_1 Measure results occurred between the Ontology model and a target model. It is calculated by

$$\%Chg = \frac{1}{N} \times \sum_{i=1}^N \frac{\text{result}_{\text{Ontology}} - \text{result}_{\text{target}}}{\text{result}_{\text{target}}} \times 100\%. \quad (17)$$

A larger $\%Chg$ value indicates more significant improvement achieved by the Ontology model. Table 3 presents the average $\%Chg$ results in our test. As shown, the Ontology

TABLE 2
The MAP and F_1 Measure Experimental Results

	TREC	Web	Category	Ontology
MAP	0.2901	0.2775	0.2612	0.2886
Micro-FM	0.3559	0.3458	0.3288	0.3622
Macro-FM	0.3875	0.3759	0.3554	0.3941

TABLE 3
Significance Test Results

Ontology vs.	MAP		Macro-FM		Micro-FM	
	$\%Chg$	p -value	$\%Chg$	p -value	$\%Chg$	p -value
TREC	7.66%	0.882	7.00%	0.551	6.69%	0.519
Web	9.25%	0.026	8.57%	0.006	8.28%	0.005
Category	20.42%	0.0002	18.40%	0.0001	16.93%	0.0002

model achieved substantial improvements over other models in the experiments.

In terms of our *Student's paired T-Test*, the typical *null hypothesis* is that no difference exists in comparing two models. When two tests produce highly different significance levels (p -value < 0.05), the *null hypothesis* can be rejected, and the significant difference between two models can be proven. In contrast, when two models produce nearly equivalent significance levels (p -value > 0.1), there is little practical difference between two models. The T-Test results are also presented in Table 3. The p -values show that the Ontology model has achieved significant improvement from the web and Category models, and has little practical difference from the TREC model.

Based on these, we can conclude that the Ontology model is very close to the TREC model, and significantly better than the baseline models. These evaluation results are promising and reliable.

7.2 Discussion

7.2.1 Experimental Result Analysis

The TREC user profiles have weaknesses. Every document in the training sets was read and judged by the users. This ensured the accuracy of the judgments. However, the topic coverage of TREC profiles was limited. A user could afford to read only a small set of documents (54 on average in each topic). As a result, only a limited number of topics were covered by the documents. Hence, the TREC user profiles had good precision but relatively poor recall performance.

Compared with the TREC model, the Ontology model had better recall but relatively weaker precision performance. The Ontology model discovered user background knowledge from user local instance repositories, rather than documents read and judged by users. Thus, the Ontology user profiles were not as precise as the TREC user profiles. However, the Ontology profiles had a broad topic coverage. The substantial coverage of possibly-related topics was gained from the use of the *WKB* and the large number of training documents (1,111 on average in each *LIR*). As a result, when taking into account only precision results, the TREC model's MAP performance was better than that of the Ontology model. However, when considering recall results together, the Ontology model's F_1 Measure results outperformed that of the TREC model, as shown in Table 2. Also, as shown on Fig. 8, when counting only top indexed results (with low recall values), the TREC model outperformed the Ontology model. When the recall values increased, the TREC model's performance dropped quickly, and was eventually outperformed by the Ontology model.

The web model acquired user profiles from web documents. Web information covers a wide range of topics and

TABLE 4
The Design of Experimental Models in the Sensitivity Test

	<i>is-a</i> only	<i>part-of</i> only	<i>is-a</i> and <i>part-of</i>	non-relationship specified
<i>LIRs</i>	-	-	-	Loc
<i>WKB</i>	GI	GP	GIP	-
<i>LIRs</i> + <i>WKB</i>	GLI	GLP	Ontology	-

serves a broad spectrum of communities [7]. Thus, the acquired user profiles had satisfactory topic coverage. However, using web documents for training sets has one severe drawback: web information has much noise and uncertainties. As a result, the web user profiles were satisfactory in terms of recall, but weak in terms of precision.

Compared to the web data used by the web model, the *LIRs* used by the Ontology model were controlled and contained less uncertainties. Additionally, a large number of uncertainties was eliminated when user background knowledge was discovered. As a result, the user profiles acquired by the Ontology model performed better than the web model, as shown in Fig. 8 and Table 2.

The Category model specified only the knowledge with a relation of *superclass* and *subclass*. In contrast, the Ontology model moved beyond the Category model and had more comprehensive knowledge with *is-a* and *part-of* relations. Furthermore, *specificity* and *exhaustivity* took into account subject localities, and performed knowledge discovery tasks in deeper technical level compared to the Category model. Thus, the Ontology model discovered user background knowledge more effectively than the Category model. As a result, the Ontology model outperformed the Category model in the experiments.

7.2.2 Sensitivity Analysis

The sensitivity analysis conducted in this paper aims to clarify the impacts made by different components in the Ontology model. As the architecture shows in Fig. 6, two knowledge resources, the global *WKB* and the *LIRs*, are used in the proposed model for user background knowledge discovery. In the constructed ontologies, knowledge with two different semantic relations, *is-a* and *part-of*, are used for *specificity* and *exhaustivity* and ontology mining. In this sensitivity study, we called these (*WKB*, *LIR*, knowledge with *is-a* and with *part-of*) as contributors and clarified their significance impact to the proposed model. In particular, the study was to answer the following questions:

- Q1. Does the model using all contributors have better performance than those using only one (or subcombination) of the four contributors?
- Q2. Which one is more important to the Ontology model, the *is-a* or *part-of* knowledge?
- Q3. Which knowledge resource is more important to the ontology model, the *WKB* or *LIRs*?

In an attempt to answer these questions, six submodels of the experimental Ontology model were evaluated, each one employing one or more contributors. Let "G" for the use of global *WKB*, "L" or "Loc" for user *LIRs*, "I" for the knowledge with *is-a*, and "P" for the knowledge with *part-of* relations, the design of six submodels is presented in Table 4,

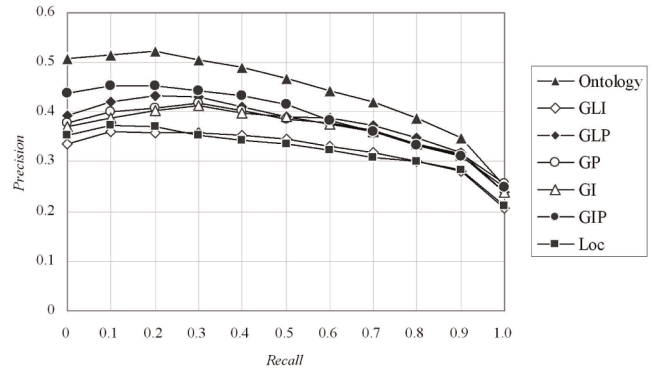


Fig. 9. The 11SPR results of sensitivity test.

along with the Ontology model employing all the contributors. We were not able to remove the unrequested relations from the taxonomy because this would ruin the ontology structure and made Algorithm 1 impossible to run. Thus, in the GI, GP, GLI, and GLP models, all semantic relations were treated as the same type (*is-a* or *part-of* as requested). The Loc model did not have any semantic relations specified because the relations were encoded from the *WKB* and the *WKB* was not employed. The comparison between the Ontology model and all the submodels was designed to answer Q1. The comparison between the GLI and GLP models (and assisted by the comparison of the GI and GP models) was to address Q2, and the comparison between the GIP and Loc models was to answer Q3. Except for the employment of different contributors, all implementation and experiment details were the same as those described in Section 6 and Fig. 7 for the Ontology model.

The overall sensitivity test results are presented in Fig. 9 and Table 5. These results demonstrate that the Ontology model significantly outperformed all six submodels. Based on this, Q1 is answered: the combination usage of all contributors makes the Ontology model outperform those using any one (or subcombination) of the contributors. This significant outperformance is also confirmed by the T-Test results presented in Table 6, where the bold *p-values* indicate substantial differences between the comparing models.

The Ontology model outperformed the GLP and GLI models under the same condition of using both the global *WKB* and local *LIRs*. This indicates that the use of knowledge with both *is-a* and *part-of* relations makes the model more effective than those using only one of them. This indication is confirmed by the comparisons of the GIP model with the GP and GI models, where only the global *WKB* is used.

Both the GP and GI models used only the *WKB*. However, the GP model treated all relations as *part-of*,

TABLE 5
The Average MAP and F-Measure Results of Sensitivity Test

	Ontology	GIP	GLP	GP	GI	GLI	Loc
MAP	0.288	0.269	0.265	0.264	0.264	0.247	0.246
Micro-FM	0.362	0.337	0.335	0.332	0.332	0.313	0.309
Macro-FM	0.394	0.365	0.362	0.359	0.359	0.338	0.334

TABLE 6
T-Test Statistic Results for Sensitivity Test

		Ontology	GIP	GLP	GP	GI	GLI
GIP	MAP	0.002					
	Mic-FM	9.53E-05					
	Mac-FM	1.11E-05					
GLP	MAP	3.95E-06	0.425				
	Mic-FM	5.16E-06	0.756				
	Mac-FM	4.47E-06	0.674				
GP	MAP	1.59E-04	0.106	0.899			
	Mic-FM	2.46E-05	0.23	0.702			
	Mac-FM	1.86E-05	0.159	0.653			
GI	MAP	8.49E-05	0.137	0.841	0.846		
	Mic-FM	1.58E-05	0.268	0.688	0.998		
	Mac-FM	1.11E-05	0.177	0.625	0.927		
GLI	MAP	1.23E-08	0.006	9.89E-04	0.029	0.022	
	Mic-FM	1.33E-09	0.005	2.53E-04	0.028	0.020	
	Mac-FM	7.77E-10	0.004	2.52E-04	0.028	0.022	
Loc	MAP	1.80E-08	0.007	0.007	0.041	0.046	0.864
	Mic-FM	3.51E-08	0.008	0.001	0.036	0.035	0.555
	Mac-FM	3.46E-08	0.007	0.001	0.042	0.042	0.611

whereas GI treated all relations as *is-a*. In the experiments, the GP model had similar performance as GI. Their little practical difference is also indicated by their high T-Test *p-value* shown in Table 6. This suggests that the knowledge with *is-a* and with *part-of* relations have similar impacts to the Ontology model. However, the significance of *part-of* knowledge was amplified when user *LIRs* were used together. As a result, the GLP model treating all as *part-of*, significantly outperformed that treating all as *is-a* (GLI), as shown in Table 6. Thus, in terms of the proposed ontology model using both the *WKB* and *LIRs*, the *part-of* knowledge is more important than that of the *is-a* knowledge. Q2 is answered.

The Ontology model, using both the *WKB* and *LIRs*, outperformed the GIP model (using only the *WKB*) and the Loc model (using only the *LIRs*). This result indicates that the combined usage of both the global *WKB* and local *LIRs* is significant for the proposed Ontology model. Missing any one of them may degrade the performance of the proposed model.

However, which one is more important: the *WKB* or *LIRs*? The Loc model using only user *LIRs* had substantially low performance, compared with the GP, GI, and GIP models using only the *WKB* (as shown in Table 6). Thus, Q3 is answered: the *WKB* is more important than user *LIRs*. In addition, the GP, GI, and GIP models using the *WKB* also have the knowledge with *is-a* and/or *part-of* semantic relations. The Loc model, however, has no such relations specified. Hence, it is reasonable to conclude that a part of the improvement achieved by the GP, GI, and GIP models is due to the *is-a* and/or *part-of* knowledge. We then have an extensive finding: the knowledge with *is-a* and/or *part-of* relations is an important component of the ontology model.

8 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.

ACKNOWLEDGMENTS

This paper presents the extensive work of, but significantly beyond, an earlier paper [39] published in WI '07. The authors thank the *Library of Congress* and *QUT Library* for the use of the LCSH and library catalogs. The authors also thank the anonymous reviewers for their valuable comments. Thanks also go to M. Carey-Smith, P. Delaney, and J. Beale, for their assistance in proofreading and editing the paper. The work presented in this paper was partly supported by Grant DP0988007 from the Australian Research Council.

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 Approach," *ACM Trans. Information Systems*, vol. 25, no. 3, article no. 12, July 2007.



Xiaohui Tao is a lecturer of computing in the Department of Mathematics and Computing, Faculty of Sciences at the University of Southern Queensland (USQ), Australia. Before working at USQ, he was a research associate with the Computer Science Discipline, Faculty of Science and Technology at Queensland University of Technology (QUT), Australia, after receiving his PhD degree from QUT in 2009. His research interests include ontology learning and mining, knowledge engineering, web intelligence, data mining, sentiment analysis and opinion mining, machine learning, and information retrieval.



Yuefeng Li is the leader of the Web Intelligence and Data Mining Group, and a professor of Computer Science Discipline, Faculty of Science and Technology at Queensland University of Technology, Australia. He has published more than 120 refereed papers (including 35 journal papers). He has also coauthored a book and edited five books. He has supervised six PhD students and four master by research students to successful completion. He is an associate

editor of the *International Journal of Pattern Recognition and Artificial Intelligence* and an associate editor of the *IEEE Intelligent Informatics Bulletin*. He has established a strong reputation internationally in the fields of web intelligence, ontology learning, and text mining, and has been awarded three Australian Research Council grants.



Ning Zhong is currently the head of the Knowledge Information Systems Laboratory, and is a professor in the Department of Life Science and Informatics, Maebashi Institute of Technology, Japan. He is also an adjunct professor and the director of the International WIC Institute (WICI), Beijing University of Technology. He has conducted research in the areas of knowledge discovery and data mining, rough sets and granular-soft computing, web intelligence, intel-

ligent agents, brain informatics, and knowledge information systems, with more than 200 journal and conference publications and 20 books to his credit. He is the editor-in-chief of *Web Intelligence and Agent Systems* and serves as associate editor/editorial board member for several international journals and book series. He is the cochair of Web Intelligence Consortium (WIC), chair of the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), and chair of IEEE Computational Intelligence Society Task Force on Brain Informatics. He is a senior member of the IEEE and the IEEE Computer Society.

▷ **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.**