An Epistemology for Online Information Gathering

C. Arun Kumar[1], Rajeswari Palaniswamy[2]

[1]Sun College Of Engineering, Anna University, Chennai, India
dr.acs.arunkumar@gmail.com

[2]S Einstein college of Engineering, Anna University, Chennai, India
rajessvarib4u@email.com

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 user local information. In this paper, a personalized epistemology 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 epistemology model is evaluated by comparing it against benchmark models in web information gathering. The results show that this epistemology model is successful.***

Keywords: Epistemology, semantic relations, world knowledge, local instance repository

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 knowledge description and formalization models 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 epistemology model should produce a superior representation of user profiles for web information gathering

In this paper, an epistemology 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 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 epistemology mining method, Specificity and exhaustively, 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 epistemology model is evaluated by comparison against some benchmark models through experiments using a large standard data set. The evaluation results show that the proposed epistemology 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. 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 epistemology 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. 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. Inter-viewing 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 pro-files 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 inter-viewing user profiles can also be viewed as manual, semiautomatic, and automatic profiles, respectively.

3. Personalized Epistemology 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' 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 conference 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 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.

	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

Figure 1

Comparison of different 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 of user interest are extracted from the WKB via user interaction. A tool called Epistemology 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.

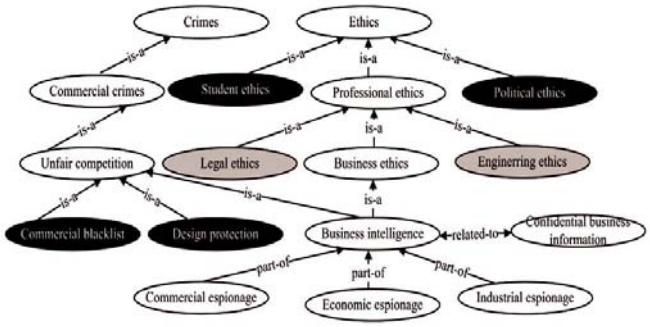
Fig. 2 is a screen-shot of the OLE for the sample topic “Economic espionage.” The structure of an epistemology that describes and specifies topic T is a graph consisting of a set of subject nodes. The structure can be formalized as a 3-tuple OðT Þ:¼ hS; taxS; reli, where S is a set of subjects consisting of three subsets Sþ, S_, and S_, where Sþ is a set of positive subjects regarding T , S_ _ S is negative, and S_ _ S is neutral; taxS is the taxonomic structure of OðT Þ , which is a noncyclic and directed graph ðS; EÞ. For each edge e 2 E and typeðeÞ ¼ is-a or part-of, iff hs1 ! s2i 2 E, taxðs1 ! s2Þ ¼ True m e a n s s1 is-a o r i s a part-of s2;. rel is a boolean function defining the related-to 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.

The remaining candidates, who are not fed back as either positive or negative from the user, become the neutral

subjects to the given topic.

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



The constructed epistemology 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 epistemology constructed, compared to those selected and constructed by a leisure user planning for a holiday.

4. Multidimensional Epistemology Mining

Epistemology mining discovers interesting and on-topic knowledge from the concepts, semantic relations, and instances in an epistemology. In this section, a 2D epistemology mining method is introduced: Specificity and Exhaustivity. Specificity (de-noted 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 epistemology.

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 OðT Þ 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 OðT Þ (this is also argued by [42]). As 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 taxS of OðT Þ, semantic specificity is also called absolute specificity and denoted by speaðsÞ.

The determination of a subject's spea is described in Algorithm 1. The isAðs0Þ and partOfðs0Þ are two functions in the algorithm satisfying isAðs0Þ \ partOfðs0Þ ¼; The isAðs0Þ returns a set of subjects s 2 taxS that satisfy taxðs ! s0 Þ ¼ True and typeðs ! s0Þ ¼ is _ a. The partOfðs0Þ returns a set of subjects s 2 taxS that satisfy taxðs ! s0Þ ¼ True and typeðs ! s0Þ ¼ part _ of. Algorithm 1 is efficient with the complexity of only OðnÞ, where n ¼ jSj. The algorithm terminates eventually because taxS is a directed acyclic graph, as defined in Definition 4.

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

Referring to concepts and the highest spe_aðsÞ. 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_aðsÞ of 1, and the root subject of tax^S has the weakest spe_a ðsÞ and the smallest value in (0, 1]. Toward the root of tax^S, the spe_aðsÞ decreases for each level up. A coefficient _ is applied to the spe_aðsÞ analysis, defining the decreasing rate of semantic specificity from lower bound toward upper bound levels. (_ ¼ 0:9 was used in the related experiments presented in this paper.) From the leaf subjects toward upper bound levels in taxS, 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 speaðsÞ value of a parent subject as the smallest speaðsÞ of its is a child subjects, applying the decreasing rate _.

If a subject has part-of child subjects, the speaðsÞ 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 spea.

1. In this analysis, the related-to semantic relations are not considered because they are non taxonomic. 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 spea value of its part-of child subjects applying _.

2. If a subject has direct child subjects mixed with is-a and part-of relationships, a spe1a and a spe2a are addressed separately with respect to the is-a and part-of child subjects.

The approaches to calculate spe1a and spe2a are the same as described previously. Following the principle that specificity decreases for the subjects located toward upper bound levels, the smaller value of spe1a or spe2a 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 taxS of OðT Þ. 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 epistemology OðT Þ constructed in Section 3 has only subject labels and semantic relations specified. In this section, we populate the epistemology 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 epistemology population. When different global knowledge bases are used, epistemology 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 (non predefined) 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 OðT Þ, the references of these clustered documents to the subjects in OðT Þ can be established and the strength of each reference can be scaled by using methods like Non latent Similarity [4]. The documents with a strong reference to the subjects in OðT Þ can then be used to populate these subjects.

Classification is another strategy to map the unstructured / semi structured 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 OðT Þ. The documents can then be classified into the subjects based on their similarity, and become the instances populating the subjects they belong to. Epistemology mapping techniques can also be used to map the features discovered by using clustering and classification to the subjects in OðT Þ, if they are in different representations.

Because epistemology 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 OðT Þ. In particular, we use the information items in the catalogs of the QUT library2 as user LIR to populate the OðT Þ 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 OðT

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.

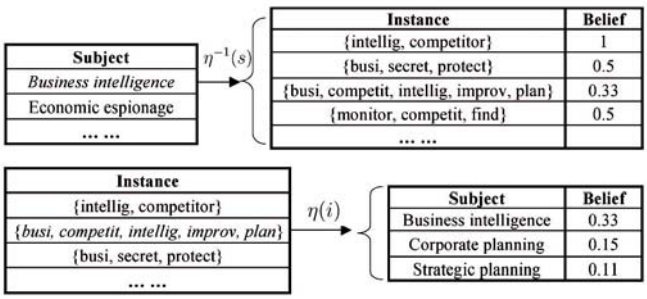


Figure 5: Mappings of subjects and instances

4.2.2 Evaluating Topic Specificity

From Definition 4, an OðT Þ 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):

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

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 regard-ing the topic. In contrast, if a subject has more negative descendants, its exhaustivity decreases. Based on this, let descðsÞ be a function that returns the descendants of s (inclusive) in OðT Þ; we evaluate a subject's exhaustivity by aggregating the semantic specificity of its descendants:

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

5. Evaluation

5.1 Experiment Design

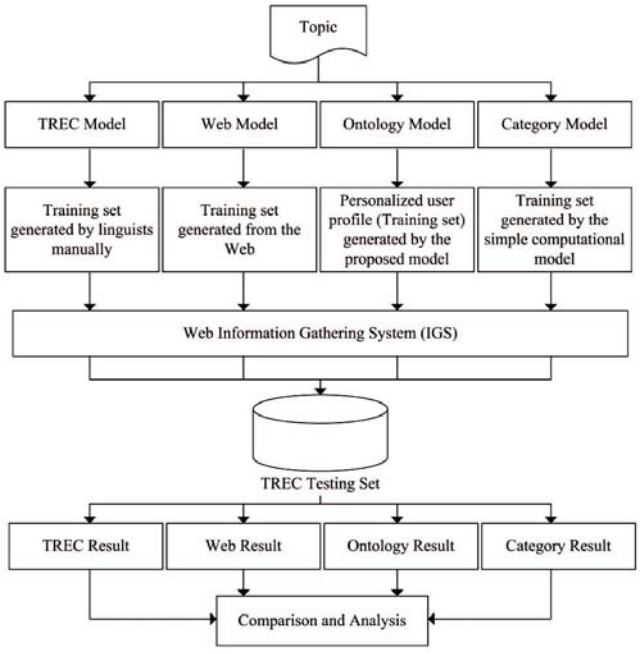
The proposed epistemology model was evaluated by objective experiments. Because it is difficult to compare two sets of knowledge in different representations, the principal design of the evaluation was to compare the effectiveness of an information gathering system (IGS) that used different sets of user background knowledge for information gathering. The knowledge discovered by the epistemology 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. 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 non relevant documents in an incoming stream [32].

A user profile consisted of two document sets: a positive document set $D^{þ}$ containing the on-topic, interesting knowledge, and a negative document set D^{-} containing the paradoxical, ambiguous concepts. Each document d held a support value supportðdÞ 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 epistemology model to the typical models representing these three group user profiles, four models were implemented in the experiments:

1. The Epistemology model that implemented the proposed epistemology 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.

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.



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.

7. Results and Discussions

The experiments were designed to compare the information gathering performance achieved by using the proposed (Epistemology) 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 F1 Measure. These are modern methods based on precision and recall, the standard methods for information gathering evaluation [1], [3].

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 Epistemology model, and then the web and the Category models.

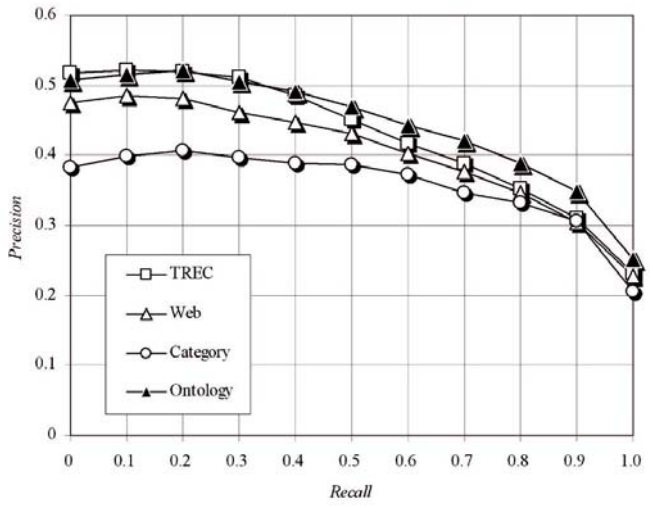


Table 2 also presents the average macro-F_1 and micro-F_1 Measure results. Where precision and recall are evenly weighted. For each topic, the macro-F_1 Measure averages the precision and recall and then calculates F_1 Measure, whereas the micro-F_1 Measure calculates the F_1 Measure for each returned result and then averages the F_1 Measure values. The greater F_1 values indicate the better performance. According to the results, the Epistemology 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 for the probability that a value could have occurred 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 Epistemology model and a target model. It is calculated by

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

	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 2: The MAP and F_1 Measure Experimental Results

	MAP		Macro-FM		Micro-FM	
Ontology vs.	*%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

Table 3: Significance Test Results

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

Based on these, we can conclude that the Epistemology 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 Epistemology model had better recall but relatively weaker precision performance. The Epistemology model discovered user background knowledge from user local instance repositories, rather than documents read and judged by users. Thus, the Epistemology user profiles were not as precise as the TREC user profiles. However, the Epistemology 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 Epistemology model. However, when considering recall results together, the Epistemology model's

F_1 Measure results out-performed 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 out-performed the Epistemology model. When the recall values increased, the TREC model's performance dropped quickly, and was eventually outperformed by the Epistemology model.

The web model acquired user profiles from web documents. Web information covers a wide range of topics and 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 Epistemology 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 Epistemology 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 super class and subclass. In contrast, the Epistemology 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 Epistemology model discovered user background knowledge more effectively than the Category model. As a result, the Epistemology model outperformed the Category model in the experiments.

7.2.2 Sensitivity Analysis

Along with the Epistemology model employing all the contributors. We were not able to remove the unrequested relations from the taxonomy because this would ruin the epistemology 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 Epistemology model and all the sub models 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 Epistemology model.

The overall sensitivity test results are presented in Fig. 9 and Table 5. These results demonstrate that the Epistemology model significantly outperformed all six sub models. Based on this, Q1 is answered: the combination usage of all contributors makes the Epistemology model outperform those using any one (or sub combination) 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 Epistemology 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

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

Where as 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 Epistemology 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 epistemology 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 Epistemology 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 Epistemology model. Missing any one of them may degrade the performance of the proposed model.

8 Conclusions and Future Work

In this paper, an epistemology 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 epistemology mining method, exhaustivity and specificity, is also introduced for user background knowledge discovery. In evaluation, the standard topics and a large test bed were used for experiments. The model was compared against bench-mark 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 epistemology 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 epistemology 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 epistemology model.

The proposed epistemology 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, how-ever, a large volume of documents existing on the web may not have such content-based descriptors. These strategies will be investigated in future work to solve this problem. The investigation will extend the applicability of the epistemology model to the majority of the existing web documents and increase the contribution and significance of the present work.

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 Experi-menters. 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 [30] A.-M. Popescu and O. Etzioni, “Extracting Product Features 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. Opinions from Reviews,” Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT ’05), pp. 339-346, 2005.

[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, “Under-standing the Relationship between Searchers’ Queries and Information Goals,” Proc. 17th ACM Conf. Information and Knowl-edge 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, “Epistemology-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 Epistemology Matches,” Proc. 16th Int’l Conf. World Wide Web (WWW ’07), 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 Epistemology Extraction Method for Adaptive e-Learning,” IEEE Trans. Knowledge and Data Eng., vol. 21, no. 6, 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, 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 Epistemology 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, "Epistemology 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. Dynamic Grammars," Proc. IEEE Computational Systems Bioinfor-matics Conf. (CSB '04), pp. 495-496, 2004.
[30] S.E. Robertson and I. Soboroff, "The TREC 2002 Filtering Track Report," Proc. Text REtrieval Conf., 2002.
[31] 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.
[32] 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.
[33] 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.
[34] M.D. Smucker, J. Allan, and B. Carterette, "A Comparison of Statistical Significance Tests for Information Retrieval Evalua-tion," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 623-632, 2007.
[35] 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.
[36] 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.
[37] X. Tao, Y. Li, N. Zhong, and R. Nayak, "Epistemology Mining for Personalized Web Information Gathering," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 351-358, 2007.
[38] 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.

Author Profile

Dr. Arun Kumar is currently associated with Sun Group of Institutions of Technical Education. He is known for his research in soft computing. He has a total of 10 years of research experience; he has served Energy systems under Government of Gujarat too. He has authored several books on the research.