IJDCST @October Issue- V-1, I-6, SW-11
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Web Information collection using Personalized Ontology Model

Mohana Prasad.Mendu[1], Kalla Kiran[2]

M.Tech Student, Dept. Of Computer Science and Engineering, Rama Chandra College of Engineering, Eluru, India [1]

Assistant Professor, Dept. Of information Technology, Rama Chandra College of Engineering, Eluru, India [2]

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.

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

1 INTRODUCTION

The huge amount of web-based information available has increased dramatically in last few decades. But the users are face a challenging issue is how to gather useful information from the web. Current web informationgathering systems attempt to satisfy user requirements by capturing their information needs. For this purpose, user profiles are created for user background knowledge description.

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. 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, and then a superior representation of user profiles can be built.

To simulate user concept models, ontologies—a knowl-edge description and formalization model—are utilized in personalized web information gathering. Such ontologies are called ontological user profiles or personalized ontologies. 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) 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, Word Net was reported as helpful in capturing user interest in some areas but useless for others.

Local analysis investigates user local information or observes user behavior in user profiles. For example, Li and Zhong discovered taxonomical patterns from the users' local text documents to learn ontologies for user profiles. Some groups learned personalized ontologies adaptively from user's browsing history. Alternatively, Sekine and Suzuki analyzed query logs to discover user background knowledge. In some works, such as 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, occa-sionally 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.

2 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' concept models is introduced.

2.1 World Knowledge Representation

World knowledge is important for information gathering. According to the definition provided by L.A. Zadeh, world knowledge is commonsense knowledge possessed by people and acquired through their experience and education. Also, as pointed out by Nirenburg and Raskin , "world knowledge is necessary for lexical and referential disambiguation, including establishing co-reference relations and resolving ellipsis as well as for establishing and maintaining connectivity of the discourse and adherence of the text to the text producer's goal and plans." In this proposed model, user background knowledge is extracted from a world knowledge base encoded from the Library of Congress Subject Headings (LCSH). We first need to construct the world knowledge base. The world knowledge base must cover an exhaustive range of topics, since users may come from different backgrounds. For this reason, the LCSH system is an ideal world knowledge base. The LCSH was developed for organizing and retrieving information from a large volume of library collections. For over a hundred years, the knowledge contained in the LCSH has undergone continuous revision and enrichment. The LCSH represents the natural growth and distribution of human intellectual work, and covers comprehensive and exhaustive topics of world knowledge by L.M.Chan. 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 .

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 , the Dewey Decimal Classification (DDC) used by Wang and Lee and King et al. , and the reference categorization (RC) developed by Gauch et al. using online categorizations. As shown in Table 1, the LCSH coversmore 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. 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). 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.

MATHEMATICAL GROUNDWORK

Definition 1: Let S be a set of subjects, an element s € S is formalized as a 4-tuple s :< label; neighbor; ancestor; descendant >, where label is the heading of s in the LCSH thesaurus; neighbor is a function returning the subjects that have direct links to s in the world knowledge base;

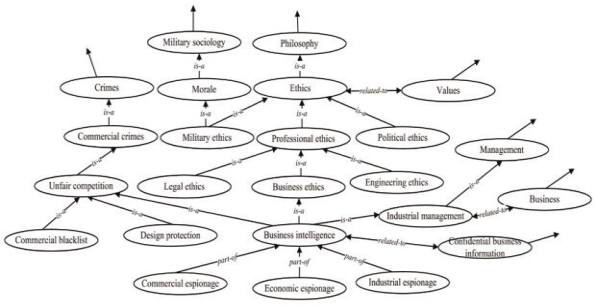

Fig. 1. A sample part of the world knowledge base

- 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 an 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 IR be a set of relations, an element r € IR is a 2-tuple r: = <edge; type>, 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 relations, and can be formally defined as a 2-tuple taxonomy constructed as a directed acyclic graph. The WKB consists of a set of subjects linked by their semantic WKB :=<S; IR>, where

- S is a set of subjects S :={s1, s2, . . . , sm};
- IR is a set of semantic relations IR: = {r1; r2; . . . ; rn} linking the subjects in 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.)

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.

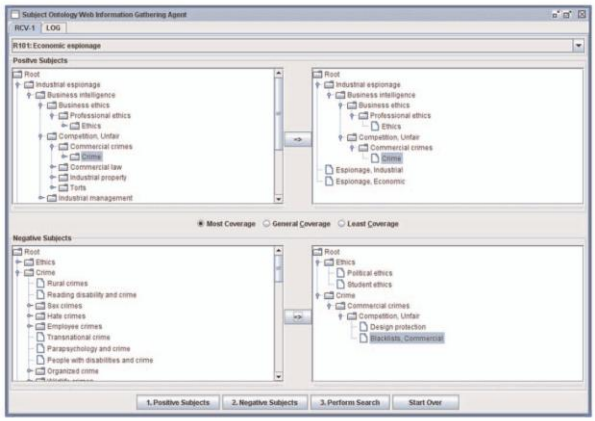

Fig. 2. Ontology learning environment

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

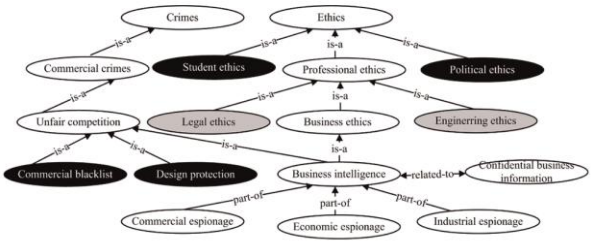

Fig: 3 An ontology (partial) constructed for topic "Economic Espionage."

Definition 4: The structure of an ontology 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 φ (T):= <S; tax^s ; rel>, where

- S is a set of subjects consisting of three subsets S^+, S^- , and S^0, where S^+ is a set of positive subjects regarding T , s^-uuu s is negative, and s^0 uuu is neutral;
- Tax^S is the taxonomic structure of φ (T), which is a noncyclic and directed graph(s, ε). For each edge e Є ε and type(e)= *is-a* or *part-of*, iff $<s_1 \rightarrow s_2>$ Є ε , 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 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.

PROPOSED SYSTEM ARCHITECTURE

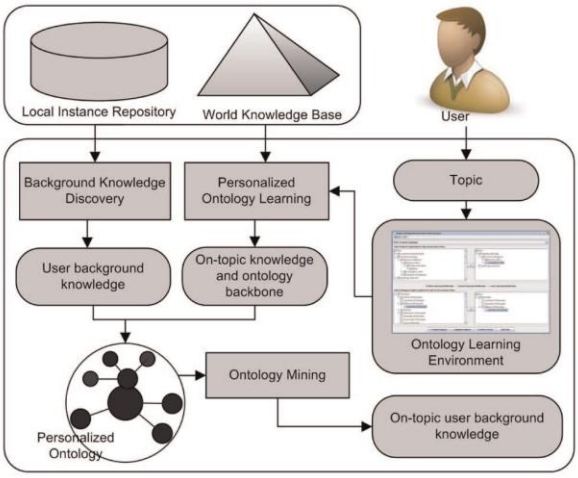

Fig. 4. Architecture of the ontology model.

The proposed ontology model aims to discover user background knowledge and learns personalized ontologies to represent user profiles. Fig. 4 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.

PROPOSED WORK

A. Experiment Design

The proposed ontology model was evaluated by objective experiments. Because it is difficult to compare two sets of knowledge in different representations, the principal design of the evaluation was to compare the effectiveness of an information gathering system (IGS) that used different sets of user background knowledge for information gathering. The knowledge discovered by the ontology model was first used for a run of information gathering, and then the knowledge manually specified by users was used for another run. The latter run set up a benchmark for the evaluation because the knowledge was manually specified by users. Under the same experimental conditions, if the IGS could achieve the same (or similar) performance in two different runs, we could prove that the discovered knowledge has the same quality as the user specified knowledge. The proposed ontology model could then be proven promising to the domain of web information gathering. User profiles can be categorized into three groups: interviewing, semi-interviewing, and non-inter viewing profiles, as previously discussed in Section 2. In an attempt to compare the proposed ontology model to the typical models representing these three group user profiles, four models were implemented in the experiments:

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

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

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

We have also further processed these documents by removing the stop words, and stemming and grouping the terms.

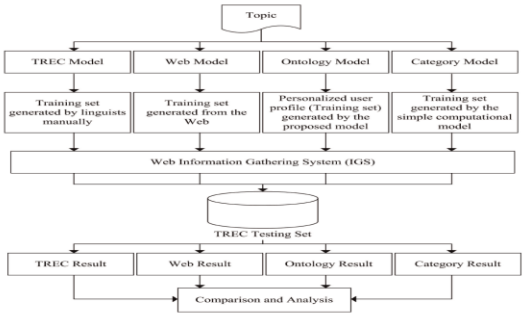

Fig. 5. Experiment design.

B. Web Information Gathering System

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

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 = $\{(t_1, f_1), (t_2, f_2), \ldots\ldots (t_k, f_k)\}$ where fi is t_i's term frequency in d. The semantic space referred to by d was represented by its normal form β(d), which satisfied $\beta(d) = \{(t_1, w_1), (t_2, w_2), \ldots\ldots (t_k, w_k)\}$ where w_i (i = 1, . . . ; k) were the weight distribution of terms and $wi = f_i / \sum_{j=1}^{k} f_j$

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

$pr_\beta(t) = \sum support(d) * wd \in D^+, (t, w) \in \beta(d)$

The testing documents were finally indexed by weight(d), which was calculated using the probability function pr_β:

$weight(d) = \sum pr_\beta(t) * T(t, d)$,

where T(t, d) =1 if t € d; otherwise T (t, d)= 0.

C. Proposed Model: Ontology Model

This model was the implementation of the proposed ontology model. As shown in Fig. 5, 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 jwj 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.

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

from this catalog data set. As a result, there were about 1,111 instances existing in one LIR on average.

The semantic relations of is-a and part-of were also analysed 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, Θ= 0:9 gave the testing model the best performance and was chosen in the experiments.

Finally, a document d in the user profile was enerated from an instance i in the LIR. The d held a support value support(d) to the T , which was measured by support(di) = str(i, T)* ∑ spe(s, T), s€ n (i)

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

D. Golden Model: TREC Model

The TREC model was used to demonstrate the interviewing user profiles, which reflected user concept models perfectly. As previously described, the RCV1 data set consisted of a training set and a testing set. The 50 topics were designed manually by linguists and associated with positive and negative training documents in the RCV1 set . 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 non-relevant to the topic. If a document d was judged relevant, it became a positive document in the TREC user profile and support(d) = $1/|D^+|$; otherwise, it became a negative document and support(d) = 0. The TREC user profiles perfectly reflected the users' personal interests, as the relevant judgments were provided by the same people who created the topics as well, following the fact that only users know their interests and preferences perfectly. Hence, the TREC model was the golden model for our proposed model to be measured against. The modeling of a user's concept model could be proven if our proposed model achieved the same or similar performance to the TREC model.

E. Baseline Model: Category Model

This model demonstrated the noninterviewing user profiles, in particular Gauch et al.'s OBIWAN 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 LIRs. However, in this model, a document's support(d) value was determined by the number of positive subjects cited by d. Thus, more positive subjects cited by d would give the document a stronger support (d) value.

There was no negative training set generated by this model, as it was not required by the OBIWAN model.

F. Baseline Model: Web Model

The web model was the implementation of typical semiinterviewing user profiles. It acquired user profiles from the web by employing a web search engine. For a given topic, a set of feature terms {t\ t €T^+ } and a set of noisy terms {t\ t €T^- } were first manually

identified. The feature terms referred to the interesting concepts of the topic. The noisy terms referred to the paradoxical or ambiguous concepts. Also identified were the certainty factors CF(t) of the terms that determined their supporting rates ([-1, 1]) to the topic.

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

These positive and negative documents were filtered by their certainty degrees CD. The CD(d) was determined by the document's index ind(d) on the returned list from Google and Google's precision rate γ The γ was set as 0.9, based on the preliminary test results using experimental topics. If a document d was in the cutoff k and γ_k= 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

$$CD(d) = \gamma_k * \frac{K - ind(d) \bmod (k) + 1}{K} * \sum_{t \in (T \cap d)} |CF(t)|$$

where K is a constant number of 10 for the number of documents in each cutoff k, 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)≤ 0.

V.EXPERIMENTAL RESULT

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

A. Experimental Results

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

An 11SPR value is computed by summing the interpolated precisions at the specified recall cutoff, and then dividing by the number of topics:

$\sum_{i=1}^{N\ precision} \lambda/N$; λ= {0.0,0.1,0.2,…..1.0}

where N denotes the number of topics, and _ ¼ indicates the cutoff points where the precisions are interpolated. At each _ 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. The average precision for each topic is the mean of the precision obtained after each relevant document is retrieved. The MAP for the 50 experimental topics is then the mean of the average precision scores of each of the individual topics in the experiments. Different from the 11SPR measure, the MAP reflects the performance in a non-interpolated recall-precision curve. The experimental MAP results are presented in Table 2. As shown in this table, the TREC model was the best, followed by the Ontology

model, and then the web and the Category models.

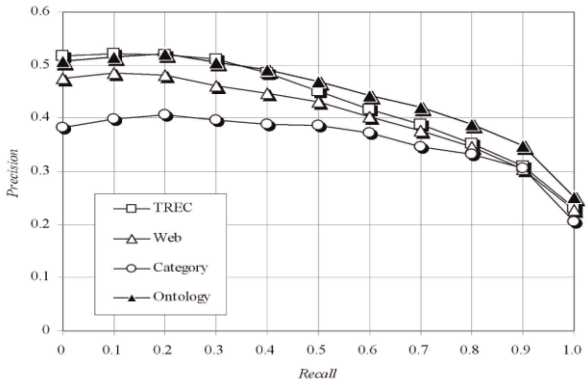

Fig. 6. The 11SPR experimental results.

Table 2 also presents the average macro-F1 and micro-F1 Measure results. The F1 Measure is calculated by

$$F1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Where precision and recall are evenly weighted. For each topic, the macro-F1 Measure averages the precision and recall and then calculates F1 Measure, whereas the micro-F1 Measure calculates the F1 Measure for each returned result and then averages the F1 Measure values. The greater F1 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 for the probability that a value could have occurred under a given null hypothesis). 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 to compute the difference in MAP and F1 Measure results occurred between the Ontology model and a target model. It is calculated by

$$\%Chg = 1/N * \sum_{i=1}^{n} \frac{\text{result ontology- result target}}{\text{result target}} * 100\%$$

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 model achieved substantial improvements over other models in the experiments.

TABLE 2
The MAP and F1 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

	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

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.

B. 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 F1 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 serves a broad spectrum of communities [7]. Thus, the acquired user profiles had satisfactory topic coverage.

TABLE 4
The Design of Experimental Models in the Sensitivity Test

	is-a only	*part-of* only	*is-a* and *part-of*	non-relationship specified
$\mathcal{LIR}$s	-	-	-	Loc
$\mathcal{WKB}$	GI	GP	GIP	-
$\mathcal{LIR}$s + $\mathcal{WKB}$	GLI	GLP	Ontology	-

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.

C. 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,

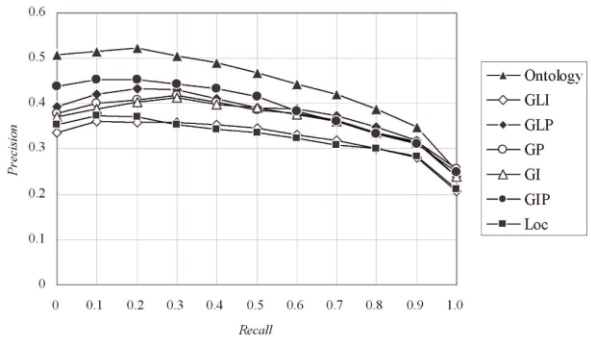

Fig. 9. The 11SPR results of sensitivity test.

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 theWKB and theWKB 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, whereas GI treated all relations as is-a.

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

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.

VI. CONCLUSIONS

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.

ACKNOWLEDGMENT

This paper is a part of our M.Tech Project. Igrateful thank to my H.O.D Smt HimaBindu garu, my project guide Sri Kalla Kiran garu and myfriend D.Ganga Raju for giving valuable guidance, suggestions, comments and contribution .

REFERENCES

1. L.M. Chan, Library of Congress Subject Headings: Principle andApplication. Libraries Unlimited, 2005.

2. S. Gauch, J. Chaffee, and A. Pretschner, “Ontology- BasedPersonalized Search and Browsing,” Web Intelligence and AgentSystems, vol. 1, nos. 3/4, pp. 219-234, 2003.

3.W. Jin, R.K. Srihari, H.H. Ho, and X. Wu,“Improving KnowledgeDiscovery in DocumentCollections through Combining TextRetrieval and Link Analysis Techniques,” Proc. Seventh IEEE Int’lConf. Data Mining (ICDM ’07), pp. 193-202, 2007.

4.J.D. King, Y. Li, X. Tao, and R. Nayak, “MiningWorld Knowledgefor Analysis of Search Engine Content,” Web Intelligence and AgentSystems, vol. 5, no. 3, pp. 233-253, 2007.

5. 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.

6. R. Navigli, P. Velardi, and A. Gangemi, “Ontology Learning andIts Application to Automated Terminology Translation,” IEEEIntelligent Systems, vol. 18, no. 1, pp. 22-31, Jan./Feb. 2003.

7.S. Nirenburg and V. Rasin, Ontological Semantics. The MIT Press,2004.

8.L.A. Zadeh, “Web Intelligence and World Knowledge—TheConcept of Web IQ (WIQ),” Proc. IEEE Ann. Meeting of the NorthAmerican Fuzzy Information Soc. (NAFIPS ’04), vol. 1, pp. 1-3, 2004.

9. N. Zhong, “Representation and Construction of Ontologies forWeb Intelligence,” Int’l J. Foundation of Computer Science, vol. 13,no. 4, pp. 555-570,2002.

10. C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen“Discovering Personally Meaningful Places: An Interactive ClusteringApproach,” ACM Trans. Information Systems, vol. 25, no. 3,article no. 12, July 2007.

Mohana Prasad.Mendu received his M.scdegree in computer science from Acharya NagarjunaUniversity, Guntur, Andhra Pradesh, in 2010 and pursuing his post graduation fromJawaharlal Nehru Technological University, Kakinada in computer science and Engineering. His areas of interest Information Security and database administration.

Kalla Kiran received his post graduation degree M.Tech in computer science and Engineering From Acharya Nagarjuna university .He has 14 years of experience, he is currently working an Assistant Professor in Department of Information Technology in *Rama Chandra College of Engineering, Eluru.* His areas of interest Information Security and Cryptography, Data Mining.