EXPLORATIONS ON SCIENCE LETTERS (ESL)
VOLUME 1, ISSUE 1 (2016):PP.29-36
SANA ACADEMIC PRESS

A Concretized Ontology Model for Web Information Caucus

Tamizharasi T[1]. Jisha Liju Daniel[2], S.Bridnha[3]
[1]*Research Scholar, St.Peter's University, Chennai.*
divyaelanchelian1990@gmail.com
[2] *Asst.Prof., Dept. of Computer Science & Applications, St.Peter's University, Chennai*
[3] *Asst.Prof., Dept. of Computer Science & Applications, St.Peter's University, Chennai*

Abstract- *Day by day, the number of Internet users and the number of attainable Web pages are ever accumulating. It is becoming tough for users to find correlated documents to their interests or needs. Thus whole process of finding relevant document is becoming time dominating. In this paper, we report on research that attempt information retrieval based on a user profile. A user can create his own concept hierarchy and use them for web searching which attempts to reveal expected documents to user. Ontology models are used to represent user profiles in personalized web information retrieval process. Many models make use of any one of the global knowledge base or user local information for representing user profiles. We attempt a personalized ontology model for knowledge representation. This model uses ontological user profiles based on a world knowledge base and user local instance repositories. It is observed that superior representation of user profiles can be built by using user concept models and it is found that the ontology model improves performance of web information retrieval.*

Index Terms: *LIR(local instance repository), World Knowledge Base, User Profile, ontology, personalization, semantic relations, user profiles, Web Information gathering.*

I.INTRODUCTION

The web-based information available to the user is ever increasing day by day and to gather useful information from the web is a really a challenging issue for the users. The web information gathering systems attempt to satisfy user requirements by creating user profiles of users.

User profiles represent the user concept models possessed by user while gathering web information. When users read through a document, they can easily determine whether it is of their interest and a judgment arises from their implicit concept models

Ontologies are the models that are widely used for knowledge description formalization. To simulate user concept models ontologies are used in personalized web information gathering system. These ontologies are named as personalized ontologies or ontological user profiles. Many researchers have attempted to discover user background knowledge from global or local analysis to represent user profiles.

i) Motivation

The objective for this project is to achieve high performance in web information retrieval that makes use of a personalized ontology model. Many times when user searches for some information with some ideas in mind, It is mostly the case that he didn't get the information exactly as he wants in first page. User has to go surf through different pages until he get the information exactly as per his interest. Here the basic idea is to create ontological user profiles from both a world knowledge base and user local instance repositories. This technique attempt fast information retrieval as per the concept model of the user.

ii) Existing systems

Commonly used knowledge bases include generic ontologies, thesauruses, and online knowledge bases. The global analysis produce effective performance for user background knowledge extraction but same is limited by the quality of the used knowledge base. Local analysis investigates user local information in user profiles. Analyzed query logs to discover user background knowledge is also used.

Users were provided with a set of documents and asked for feedback. User background knowledge was discovered from this feedback for user profiles. The discovered results may contain uncertain and noisy information.

iii) Concept or seed idea

We proposed a Personalized Ontology model for web information retrieval to get high performances over the techniques used previously that uses both global analysis as well as local analysis from LIR(Local Instance Repository). Here, we suggest some alternatives such as in a multidimensional ontology mining method, Specificity and Exhaustively is also introduced by considering the growth of web information and the growing accessibility of online documents. Further we use customized algorithm to calculate the clusters. These clusters minimizes the time for retrieval.

II.LITERATURE REVIEW

This chapter will introduce the previous system and its analysis. It includes the comparison of existing system with proposed system.

PREVIOUS ONTOLOGY LEARNING

Many existing models used global knowledge bases to learn ontologies in web information retrieval.e.g Gauch and Sieg learned personalized ontologies developed from the Open Directory Project which find users' preferences and interests in web search. Use of Dewey decimal classification, King attempted to improve performance in distributed web information retrieval. Downey used Wikipedia to understand user interests in queries. User background knowledge was discovered though performance and the same was limited by the quality of the global knowledge base. Many works attempted to find user background knowledge from user local information to learn personalized ontologies.

Ontologies can be constructed in different ways. Different data mining techniques lead to more user background knowledge being discovered e.g user local documents makes use of pattern recognition and association rule mining techniques to discover knowledge. Li and Zhong used pattern recognition and association rule mining techniques to find knowledge from user local documents to construct ontology. One can discover semantic concepts and relations from web documents. Web content mining techniques were used to discover semantic knowledge from domain-specific text documents for ontology learning. The knowledge discovered in these works contained noise and uncertainties. Additionally, ontologies were mostly used to improve the performance of knowledge discovery. Using a fuzzy domain ontology extraction algorithm, a mechanism was developed to construct concept maps based on the posts on online discussion forums.

User Profiles

User profiles were created to identify user information needs on the basis of interest of users in web information gathering that interpret the semantic meanings of queries. User profiles can be defined as the interesting topics of a user's information need.

User profiles can be categorized: interviewing, semi- interviewing, and non-interviewing. Interviewing user profiles are obtained by using manual techniques like questionnaires, interviewing and analyzing user classified training sets.e.g TREC Filtering Track training sets, which were generated manually. The users read document and gave a positive or negative judgment to the document against a given topic.

WORLD KNOWLEDGE REPRESENTATION

World knowledge is commonsense knowledge possessed by people that is acquired through their experience and education. It is important for information gathering. We first need to construct the world knowledge base. That must cover an exhaustive range of topics , since users may 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.The LCSH covers comprehensive and exhaustive topics of world knowledge. In addition, the LCSH is the most comprehensive non-specialized controlled vocabulary in English and it has become a de facto standard for subject cataloging and indexing. The LCSH system is superior than other world knowledge taxonomies. Table 1 shows a comparison of the LCSH with the Library of Congress

Classification (LCC), Dewey Decimal Classification (DDC) used and the reference categorization (RC).

TABLE 1 Comparison showing 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

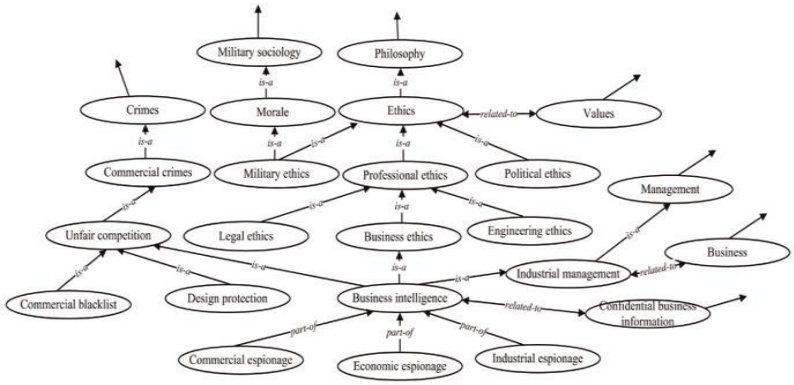
As shown in Table 1, the LCSH covers more topics than other. It has a more specific structure, and it specifies more semantic relations. The structure of LCSH is directed acyclic graph that contains three types of references: Broader term (BT), Used-for (UF), and Related term (RT).The BT references shows different levels of abstraction (or specificity).The primitive knowledge unit in our world knowledge base is subjects. They are encoded from the subject headings in the LCSH.

These subjects are formalized as follows:
Def.1 Let *S* is a set of subjects, an element *s* is represented as a tuple *s* = < *label, neighbor, ancestor, descendant* >
where

*label i*s the heading of *s* in the LCSH thesaurus;

neighbor is a function returning the subjects that have direct links to *s* in the world knowledge base

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



ancestor is a function returning the subjects that have a higher level of abstraction than *s* and link to *s* directly or indirectly in the world knowledge base;

descendant is a function returning the subjects that are more specific than s and link to *s* directly or indirectly in the world knowledge base.

The semantic relations of *is-a, part-of,* and *related-to* links the subjects to each other in the world knowledge base.The relations are formalized as following.

Def.2 Let *R* be a set of relations ,an element r□*R* is a tuple r= <*edge, type* > where

An *edge* that connects both subjects that hold a type of relation.
A *type* of relations is element from {*is-a, part-of, related-to*}

With Def 1 and 2, the WKB can be formalized as :

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

S is a set of subjects = { *s1,s2,….., sm* }
R is a set of semantic relations *R= { r1, r2, r3, …rn}*
linking the subjects in *S.*

Ontology Formation

A tool called OLE(Ontology Learning Environment) is used for constructing ontologies. The subjects of user interest are extracted from the WKB with user interaction. For a given topic, the interesting subjects consist of two subjects: positive subjects and negative subjects. Positive subjects are the concepts relevant to the topic, and negative subjects are the concepts not related to topic as per user need. Thus, for a given topic, the OLE provides users with a set of two candidates to identify positive and negative subjects. These subjects are extracted from the WKB.

For a given topic e.g. "economic" and "espionage", the user selects positive subjects for the topic. The positive subjects selected by user are presented on the top-right in hierarchy. The negative subjects are the descendants of the positive subjects. These are shown on the bottom-left side. From them user selects the negative subjects. These negative subjects to user re listed on the bottom-right panel (here "Political ethics" and "Student ethics"). 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 candidates that are not either positive or negative as per the fedback from the user, become the neutral to the topic specified.

Ontology is constructed for the given topic using user feedback subjects.It contains three types of Subjects: positive, negative, and neutral
subjects. The structure of the ontology is based on the semantic relations linking these subjects in the WKB.

Fig.2 An ontology (partial) constructed for topic 'Economic espionage'

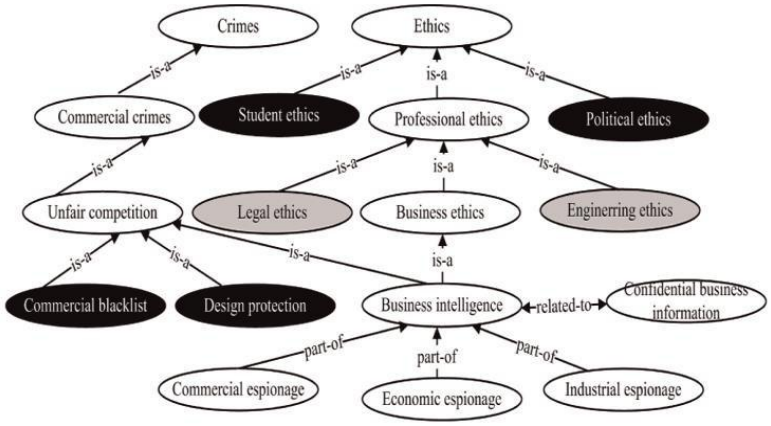


Fig.2 illustrates the ontology constructed for the sample topic "Economic espionage," where the white nodes are positive, the dark nodes are negative, and the gray nodes are neutral subjects.

Formalization of ontology constructed for a given topic is

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

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

The constructed ontology is personalized as the user selects positive and negative subjects for personal preferences and interests.

Multidimentional Ontology Mining

Multidimentional ontology mining method discovers on-topic and interesting knowledge from the user concepts, semantic relations, and instances in an ontology. Here a 2D ontology mining methods are : *specificity and Exhaustivity*. These methods observes the subjects and the strength of their associations in an ontology structure.

Specificity (denoted *spe*) shows a subject's focus on a given topic. Exhaustivity (denoted *exh*) shows a subject's semantic space dealing with the topic.These methods investigate the subjects and the strength of their associations in an ontology structure.

A subject's specificity is of two types

1) Semantic specificity - on the referring-to concepts and
2) Topic specificity - on the given topic.

III.SEMANTIC PRECISION

The semantic specificity is investigated from the structure of $\mathcal{O}(\mathcal{T})$ inherited from the world knowledge base. The strength of a focus is guided by the subject's locality in the taxonomic structure of $tax^{\mathcal{S}}$ of is a graph that links semantic relations. The semantic specificity is measured by hierarchical semantic relations *(is-a* and *part-of)* held by that subject and its neighbors in taxonomic structure $tax^{\mathcal{S}}$ $\mathcal{O}(\mathcal{T})$ $\mathcal{O}(\mathcal{T})$

,As subjects have a fixed locality on the $tax^{\mathcal{S}}$ of sen $tax^{\mathcal{S}}$ $\mathcal{O}(\mathcal{T})$ ecificity can be described as absolute specificity and can be denoted by $spe_a(s)$.

The subjects located at upper bound levels and toward the root are more abstract than those of at lower bound levels toward the "leaves." The semantic specificity of a lower bound subject is greater than that of an upper bound subject.

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.

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

Topic Specificity
User background knowledge that uses user's local information is used to analyse the topic specificity of a subject.

User Local Instance Repository (LIR)
User background knowledge can be discovered from user local information collections like user's browsed web pages, stored documents and composed or received emails .The ontology has only subject labels and semantic relations specified.We follow the ontology with the instances generated from user local information collection. A collection of user local information is reffered as user's local instance repository (LIR).
To generate users LIRs is a challenging part. The documents in LIRs may be of different types semi-stuctured like the browsed HTML and XML web documents or unstructured like the stored local DOC and TXT documents. From this one has to generate LIR. Some semi-structured web documents has content-related descriptors specified in the metadata sections. These descriptors have direct references to the concepts specified in WKB. These documents are ideal to generate the instances for ontology population e.g. the infoset tags in XML documents. Ontology mapping can be used when different world knowledge bases are used e.g GLUE system.
For the documents without clear and direct references in Local instance repository (LIR), the different data mining techniques, clustering and classification can be followed.
The WKB is encoded from the LCSH. The LCSH contains the content-related descriptors (subjects) in controlled vocabularies.User background knowledge of a user is to be discovered from both the user's LIR and O(T).

IV.SYSTEM ARCHITECTURE

Figure 3 Architecture of ontology model

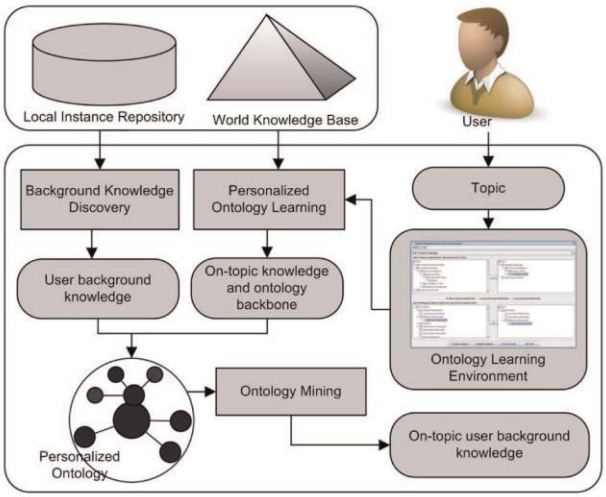


Fig. shows architecture of the ontology model that discovers user background knowledge and learns personalized ontologies to represent user profiles.

A personalized ontology is constructed for the given topic that uses two knowledge resources, the global world knowledge base and the user's local instance repository. The WKB (world knowledge base) provides the taxonomic structure for the personalized ontology. The user background knowledge is then discovered from the user LIR(Local Instance Repository). For the given topic by the user, the specificity and exhaustivity of subjects are investigated to discover user background knowledge discovery.

The input to the proposed ontology model is a topic and the output is user background knowledge which is computationaly discovered.User profile consisting of positive documents and negative documents .Each document *d* is associated with a *Support(d)* value indicating its support level to the topic.

V.EXPERIMENTAL-EVALUATION

The principal experimental design of the evaluation was to compare the effectiveness of an information gathering system (IGS) for the different sets of user background knowledge.

User profiles can be broadly classified into three groups: interviewing, semi-interviewing, and non-interviewing. That each is used by the TREC model , Web model , and Category model respectively.We compare the proposed ontology model to the typical model.

1. The TREC model that represented the perfect interviewing user profiles and user background knowledge was manually specified by users.
2. The Category model that represented the noninter- viewing user profiles.
3. The Web model that represented the semiinterviewing user profiles.
4. The Ontology model that we have implemented as the proposed ontology model. Here user background knowledge is computationally discovered.

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 duplicated documents, normalization of dateline and byline formats, addition of copyright statements, and so on.

Web Information Gathering System

The IGS was an implementation of a model developed by Li and Zhong that uses user profiles for web information gathering. The input support values associated with the documents in user profiles affected the IGS's performance.Experiments here assumes Li and Zhong's model that uses support values of training documents for web information gathering.

Proposed Model: Ontology Model

The input to ontology model was a topic and the output was a user profile consisting of positive documents (D^+) and negative documents (D^-). Each document was associated with a *support(d)* value indicating support level to the topic for document *d*.

Here the WKB was constructed based on the LCSH system.

The constructed WKB contained multiple subject covering a wide range of topics linked by semantic relations. The user's personalized ontologies were constructed by user interaction. Authors played the user role to select positive and negative subjects for ontology construction, for
each topic T, the ontology mining method was performed on the

constructed $\mathcal{O}(\mathcal{T})$ and the user LIR to discover interesting concepts. The user provided documents was preprocessed by removing the stopwords, and stemming and grouping the terms. 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 the different users and for different topics experiment was performed. The semantic relations of *is-a* and *part-of* were analyzed in the ontology mining for interesting knowledge discovery. As per algorithm 1,the coefficient θ in some preliminary tests had been conducted for various values of the coefficient θ such as 0.5, 0.7, 0.8, and 0.9. As a result, θ = 0.9 gave the best performance and was chosen in the experiments.

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

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

Various thresholds of *support(d)* were tested to classify positive and negative documents. As constructed ontologies were personalized and focused on various topics, we could not find a universal threshold that worked for all topics. Hence we set the threshold as *support(d)=0,* following the nature of positive and negative defined.

The documents with *support(d)* > 0 formed D+, and those with negative *support(d)* $\leq$ O formed D− eventually.

VI.EXPERIMENTAL SETUP

The performance of the experimental models was measured by the precision averages at 11 standard recall levels (11SPR). Precision is the ability of a system to retrieve only relevant documents and 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 it by the number of topics:

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

where N = number of topics

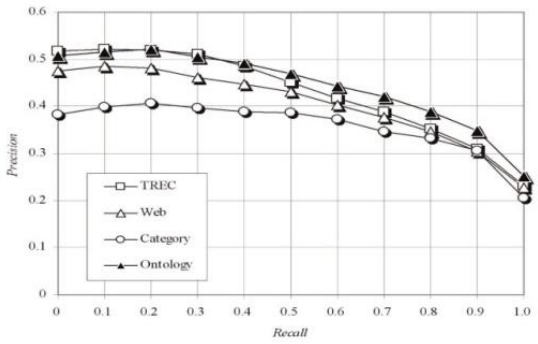
λ = cutoff points where precisions are interpolated.

At each λ point, an average precision value over N topics is calculated. These average precisions then link to a curve describing the recall-precision performance. The experimental 11SPR results are plotted in Fig. 4, where the 11SPR curves show that the Ontology model was the best, followed by the TREC model, the web model.

The average precision for each topic is the mean of the precision obtained after each relevant document is retrieved.

As per graph TREC model was the best, followed by the Ontology model, and then the web.

Fig.4 The 11SPR experimental results.



VII.CONCLUSION

An ontology model is evaluated that represents user background knowledge in personalized web information retrieval. The model discovers user background knowledge from LIR and builds userwise personalized ontologies extracting world knowledge from LCSH.The model was compared against benchmark models such as TREC model and WEB model. The results shows that our model is promising model in web information gathering that attempts to retrieve documents as per user interest that obviously improves performance of web information retrieval system. It is found that the use of both i.e global and local knowledge performs better than using any one.

The present work assumes that all user local instance repositories uses data mining techniques such as web page clustering that improves results and extends the applicability of the ontology model to the majority of the existing web documents and increase the contribution and significance of the present work. We are hopeful to find different areas to improve background knowledge of user and thus improving user personal profile.

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]. A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," Proc. ACM SIGIR ('07), pp. 7-14, 2007.
[6]. A Personalized Ontology model for web Information gathering. Xiaohui Tao, Yuefeng Li, and Ning Zhong, Senior Member, IEEE, IEEE Transactions on knowledge and data engineering, Vol 23, No 4,April 2011.
[7]. 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.
[8]. 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.
[9]. 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.
[10]. 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.
[11]. S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.
[12]. R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen, "Using Google Distance to Weight Approximate Ontology Matches," Proc. 16th Int'l Conf. World Wide Web (WWW '07),pp. 767-776, 2007.
[13]. J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[14]. 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.
[15].