Genetic Algorithm Based Personalized Ontological Model for Web Information Retrieval

Jayashri J. Patil

M.E (Comp.Engg.), KKWIEER, Nasik

Abstract

Nowadays there is dramatic increase in the number of Internet users and the number of accessible Web pages. It is becoming increasingly difficult for users to find relevant documents to their particular needs. The process of finding relevant document to user is becoming time consuming. In this paper, we report on research that adapts information retrieval based on a user profile. A user can create his own concept hierarchy and use them for web searching that attempts to reveal expected documents to user. Ontology models are widely used to represent user profiles in personalized web information retrieval. Many models have utilized only knowledge from either a global knowledge base or user local information for representing user profiles. A personalized ontology model is used for knowledge representation.This model uses ontological user profiles based on both a world knowledge base and user local instance repositories. This model makes use of GA (Genetic algorithm).It is observed that genetic alogithm based personalized ontological model approach improves the overall performance of web information retrieval.

Keywords: *genetic algorithm, local instance repository, ontology, personalization, semantic relations, user profiles, web Information gathering, world knowledge.*

1. Introduction

The web-based information available to the user has increased drastically and to gather useful information from the web is a challenging issue for the users. The web information gathering systems attempt to satisfy user requirements by creating user profiles.

User profiles represent the user concept models possessed by them while gathering web information. A concept model possessed by users is generated from their background knowledge. Many web ontologies have observed it in user behavior. When users read through a document, they can easily determine whether or not it is of their interest to them, a judgment that arises from their implicit concept models. If one can simulates user's concept model then a superior representation of user profiles can be built.

Ontologies are the models used for knowledge description formalization.To simulate user concept models ontologies are used in personalized web information gathering.These ontologies are called Personalized ontologies or ontological user profiles. Many researchers have attempted to discover user background knowledge through global or local analysis to represent user profiles.

1.1 Motivation

The basic objective for this project is to achieve high performance in web information retrieval using a personalized ontology model. Most of the times when user searches for some information with some ideas in mind, It is always the case that he didn't get the information exactly as he wants in first page. He has to go through different pages until he get the information exactly as per his concept. The basic idea is to create ontological user profiles from both a world knowledge base and user local instance repositories in order to have a fast information retrieval as per the concept model of the user.

1.2 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 it is limited by the quality of the used knowledge base.
Local analysis investigates user local information or observes user behavior in user profiles. Analyzed query logs to discover user background knowledge is used.
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. The discovered results may contain noisy and uncertain information.

1.3 Concept or seed idea

We proposed a Personalized Ontology model for web information retrieval to get high performances over the techniques used previously .It uses both knowledge 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 Exhaustivity is also introduced by considering the rapid explosion of web information and the growing accessibility of online documents.

2. Literature Review

This chapter will introduce us with the previous system and its analysis. It also includes the comparison of existing system with proposed system. This will give us the detail idea about the need of proposed system.

2.1 Ontology Learning

Many existing models used global knowledge bases to learn ontologies for web information retrieval.e.g Gauch and Sieg learned personalized ontologies developed from the Open Directory Project to specify users' preferences and interests in web search. On the basis of the Dewey decimal classification, King improved performance in distributed web information retrieval. Wikipedia was used by Downey to help understand user interests in queries.These discovered user background knowledge though performance was limited by the quality of the global knowledge base.Many works mined 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 can be helpful that uses pattern recognition and association rule mining techniques to discover knowledge. Li and Zhong used pattern recognition and association rule mining techniques to discover knowledge from user local documents to construct ontology. Tran translated keyword queries to Description Logics' conjunctive queries and used ontologies to represent user background knowledge. Zhong proposed domain ontology learning approach that employed various data mining and natural-language understanding techniques is introduced. One can learn to discover semantic concepts and relations from web documents. Web content mining techniques were used by to discover semantic knowledge from domain-specific text documents for ontology learning. Finally, Shehata captured user information needs at the sentence level rather than the document level, and represented user profiles by the Conceptual Ontological Graph.

The knowledge discovered in these works contained noise and uncertainties. Additionally, ontologies were used in many works to improve the performance of knowledge discovery. Using a fuzzy domain ontology extraction algorithm, a mechanism was developed in 2009 to construct concept maps based on the posts on online discussion forums. One can integrate data mining and information retrieval techniques to further enhance knowledge discovery. GLUE model was proposed by Doan and used machine learning techniques to find similar concepts in different ontologies. Dou proposed a framework for learning domain ontologies using pattern decomposition, clustering/classification, and association rules mining techniques that attempted to explore world knowledge more efficiently.

2.2 User Profiles

User profiles were created to capture 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 into two diagrams: the data diagram user profiles acquired by analyzing a database or a set of transactions, the information diagram user profiles acquired by using manual techniques, such as questionnaires and interviews or automatic techniques, such as information retrieval and machine learning.

Generic User Model was proposed by Van der Sluijs and Huben to improve the quality and utilization of user modeling.Wikipedia was also used to help discover user interests. In order to acquire a user profile, Chirita and Teevan used a collection of user desktop text documents and emails and cached web pages to explore user interests. Makris acquire 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 as interviewing, semi-interviewing, and non-interviewing. Interviewing user profiles are acquired by using manual techniques, such as questionnaires, interviewing users, and analyzing user classified training sets.e.g TREC Filtering Track training sets, which were generated manually. The users read each document and gave a positive or negative judgment to the document against a given topic. These training documents reflect user background knowledge accurately. In Semi-interviewing user profiles there is limited user involvement. These techniques provide users with a list of categories and ask users for interesting or non-interesting categories.e.g Web training set acquisition model, which extracts training sets from the web based on user fed back categories. Noninterviewing techniques do not involve users at all and captures user interests by user activity and behaviour to discover user background knowledge. e.g.OBIWAN, which acquires user profiles based on users' online browsing history.

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

3. Personalized Ontology

Web search can be made personalized by constructing personalized ontologies which describes and specifies user background knowledge from captured user interest.i.e from user profiles. While searching web,users might have different expectations for the same query. For example, for the search 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.

An assumption is formed on the basis of observation that web users have a personal concept model for their information needs and it may change according to different information needs. Here we introduce a model constructing personalized ontologies from users's concept models.

3.1 World Knowledge Representation

World knowledge is commonsense knowledge possessed by people and acquired through their experience and education. World knowledge is important for information gathering. We first need to construct the world knowledge base. It 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 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. LCSH is used as a means for enhancing subject access to knowledge management systems

TABLE1 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 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).

As shown in Table 1, the LCSH covers more topics than other taxonomies. It has a more specific structure, and it specifies more semantic relations. The classification quality is superior and well-defined with continuously refined cataloging rules. These features contribute LCSH an ideal world knowledge base for knowledge dicovery.The structure of LCSH is directed acyclic graph. It 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.

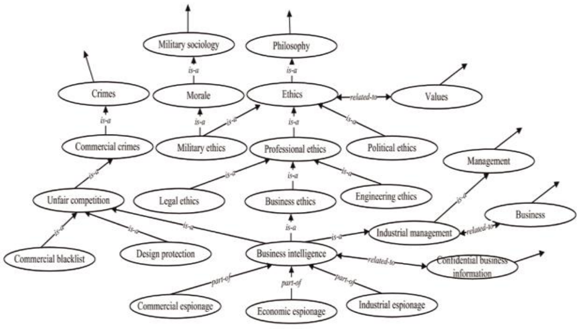


Fig.1 A sample part of the world knowledge base.

These subjects are formalized as follows:

Def.1 Let ***S*** is a set of subjects, an element *s* ∈***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
- ***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 are used to link the subjects to each other in the world knowledge base.The relations are formalized as follows:

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.***

3.2 Ontology Construction

Ontologies are constructed using a tool called OLE (Ontology Learning Environment).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 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.

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 hierarchical form. The negative subjects are the descendants of the positive subjects to the user. These are shown on the bottom-left panel. 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 which are not either positive or negative as per the fedback from the user, become the neutral subjects 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.

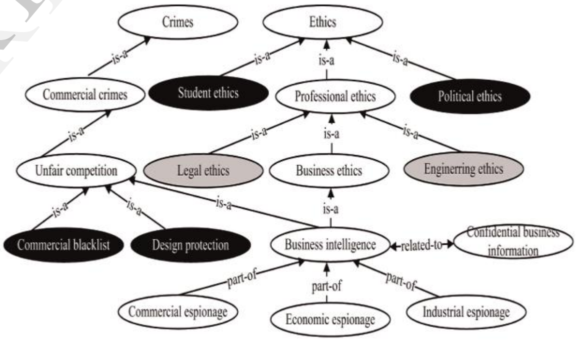


Fig.2 Ontology (partial) constructed for topic ‘Economic espionage’

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 subject 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 $type(e) = is\text{-}a$ or $part\text{-}of$, 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 user selects positive and negative subjects for personal preferences and interests, the constructed ontology is personalized.

4. Ontology Mining

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

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

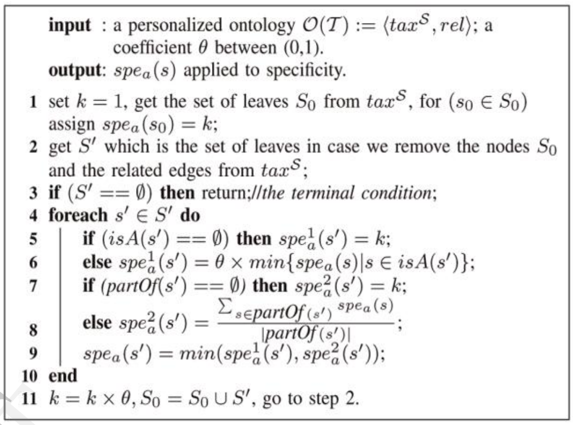
A subject's specificity is of two types 1) semantic specificity - on the referring-to concepts and 2) topic specificity - on the given topic.

4.1 Semantic Specificity

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

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

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



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

Algorithm1. Analysis of semantic relations for specificity.

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

4.2 Topic Specificity

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

4.2.1 User Local Instance Repository (LIR).

User background knowledge can be discovered from user local information collections, such as a user's browsed web pages, stored documents and composed or received emails .The ontology has only subject labels and semantic relations specified. Here we follow the ontology with the instances generated from user local information collection.A collection of user local information is called user's local instance repository (LIR).

To generate users LIRs is a challenging issue. 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 to match the concepts when different world knowledge bases are used e.g GLUE system.

For the documents that do not have such clear and direct references in Local instance repository (LIR), the different data mining techniques, clustering, and classification can be followed.

The clustering techniques group the documents into clusters based on the document features. These features, usually represented by terms, can be extracted from the clusters. These represent the user background knowledge discovered from the user LIR. The semantic similarity between these features and the subjects in $\mathcal{O}(\mathcal{T})$ can be measured and the references of these clustered documents to the subjects in $\mathcal{O}(\mathcal{T})$ can be established. The documents with a strong reference to the subjects in $\mathcal{O}(\mathcal{T})$ can then be used to populate these subjects.

The another strategy that can be applied is Classification that maps the unstructured/semi-structured documents in user LIRs to the representation in the global knowledge base. We can measure the semantic similarity between documents in the LIR and the subjects in $\mathcal{O}(\mathcal{T})$ by using the subject labels. The documents can then be classified into the different subjects based on their similarity, and become the instances of the subjects of which they belong to. Ontology mapping technique can be used to map the features discovered by using clustering and classification to the subjects in ontology $\mathcal{O}(\mathcal{T})$, if they are in different representations.

The WKB is encoded from the LCSH. The LCSH contains the content-related descriptors (subjects) in controlled vocabularies. Corresponding to these descriptors, the catalogs of library collections also contains the descriptive information of library-stored books and documents. The descriptive information, such as the title, table of contents, and summary, is provided by authors and librarians. This trustworthy information classified by the experts can be recognized as the extensive knowledge from the LCSH. A list of content-based descriptors cited on the bottom, 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 of a user is to be discovered from both the user's LIR and $\mathcal{O}(\mathcal{T})$.

4.2.2 Evaluation

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

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

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

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

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

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

4.3 Multidimensional Analysis of Subjects

The exhaustivity of a subject refers to the extent of its concept space dealing with a given topic. This space extends if a subject has more positive descendants regarding the topic. In contrast, if a subject has more negative descendants, its exhaustivity decreases. Based on this, let $desc(s)$ be a function that returns the descendants of s (inclusive) in $\mathcal{O}(\mathcal{T})$; we evaluate a subject's exhaustivity by aggregating the semantic specificity of its descendants:

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

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

$$\begin{aligned}\mathcal{S}^+ &= \{s|spe(s,\mathcal{T})>0, exh(s,\mathcal{T})>0, s\in\mathcal{S}\};\\ \mathcal{S}^- &= \{s|spe(s,\mathcal{T})<0, exh(s,\mathcal{T})<0, s\in\mathcal{S}\};\\ \mathcal{S}^\diamond &= \{s|s\in(\mathcal{S}-(\mathcal{S}^+\cup\mathcal{S}^-))\}.\end{aligned}$$

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

5. System Architecture

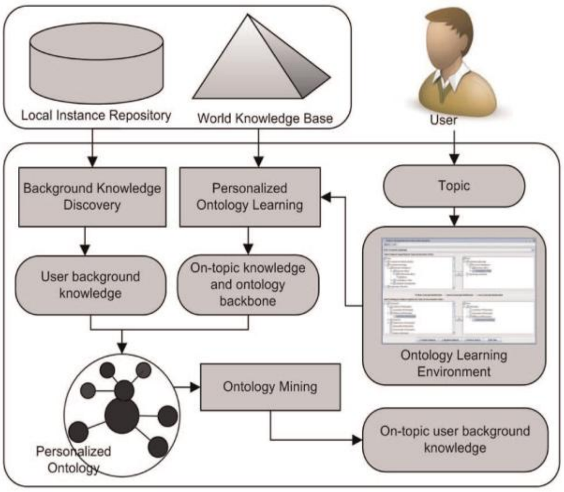

Local Instance Repository
World Knowledge Base
User
Background Knowledge Discovery
Personalized Ontology Learning
Topic
User background knowledge
On-topic knowledge and ontology backbone
Ontology Learning Environment
Ontology Mining
Personalized Ontology
On-topic user background knowledge

Fig.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.

6. Evaluation

6.1 Experiment Design

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.

The comparison is performed using a test set and a set of topics for the ontology model and that of TREC model. TREC model can be viewed as a benchmark model as the knowledge was manually specified by the users. In information gathering evaluation, a common batch style experiment is developed for the comparison of the models using a test set and a set of topics associated with relevant judgements. Our experiments followed this style and were performed under the experimental set up by the TREC-11 filtering track.This track evaluate the methods of persistant user profiles for separation as relevant and non-relevant documents.

In the experiments, user background knowledge was represented by user profiles. A user profile consisted of two sets of documents: a positive document set D+ containing the on-topic,interesting knowledge and a negative document set D⁻ containing the ambiguous, paradoxical concepts. For each document *d*, there is a s support value to the given topic.The baseline models in our experiments were carefully selected based on this representations.

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.

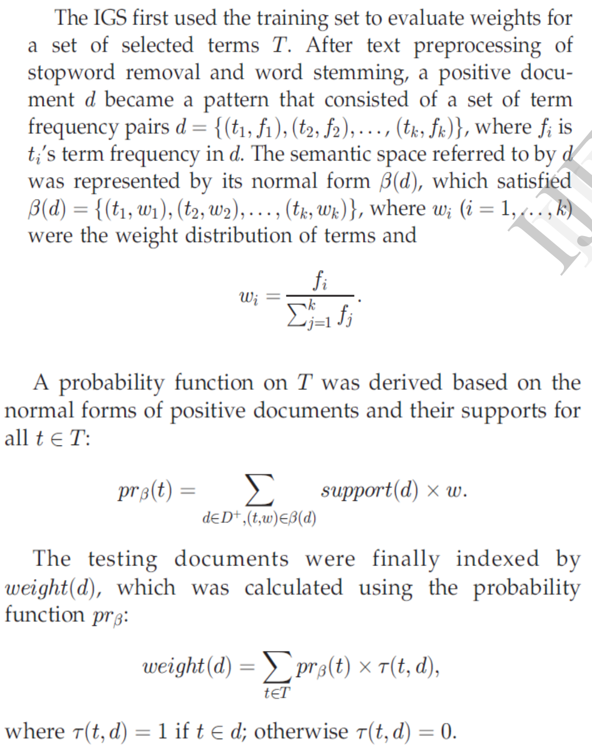
The TREC-11 Filtering Track testing set and topics were used in our experiments. The testing set was the Reuters Corpus Volume 1 (RCV1) corpus [21] that contains 806,791 documents and covers a great range of topics. This corpus consists of a training set and a testing set partitioned by the TREC. The 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. We have also further processed these documents by removing the stop-words, and stemming and grouping the terms.

In the experiments, we attempted to evaluate the proposed model in an environment covering a great range of topics. However, it is difficult to obtain an adequate number of users who have a great range of topics in their background knowledge. The TREC-11 Filtering Track provided a set of 50 topics specifically designed manually by linguists, covering various domains and topics. 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.

The titles of topics were used, based on the assumption that in the real world users often have only a small number of terms in their queries

6.2 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.



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

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

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

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

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

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

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

6.3 Proposed Model

6.3.1 Genetic algorithm based ontology Model.

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 support level to the topic.

The WKB was constructed based on the LCSH system. The constructed WKB contained multiple subjects covering a wide range of topics linked by semantic relations. The user's personalized ontologies were constructed based onuser interaction. Here the 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 1the 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)* ≤ 0 formed D- eventually.

Genetic algorithms to identify topic.

There are many concepts and terms in document. Our main problem is to distinguish the weight of each concept or term in topic of document. We represented the weight of concepts and terms as Wdi vector in previous section. Each concept or term can have a weight between 0 and 1. For simplifying our problem, we can consider weights as a binary number. That means the related concept or term belongs or doesn't belong to topic of document. A chromosome is defined as a list of concept or term weights which have real or binary numbers. The definition of a chromosome is represented as *J* = (*j*1, *j*2, . . . , *ji* , . . . , *jL*), where *ji* denotes the weight of the concept *i* and *L* is the number of concept to be considered. Each gene represents a concept or term weight. The genes of initial chromosomes are generated randomly and the range of weight values is from 0.0 to 1.0 for experiments.

Our Ontological model assumes genetic algorithm as a clustering technique of web pages Genetic Algorithm, the solutions are called chromosomes. After the initial population is generated randomly, different functions are applied e.g. selection and variation. These are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into the next generation. The quality of an individual is measured by a fitness function.These results in optimized web pages.

These results are again filtered on the basis of personalized information in LIR. The overall result improves the performance of our proposed model.

7. Results and analysis

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.

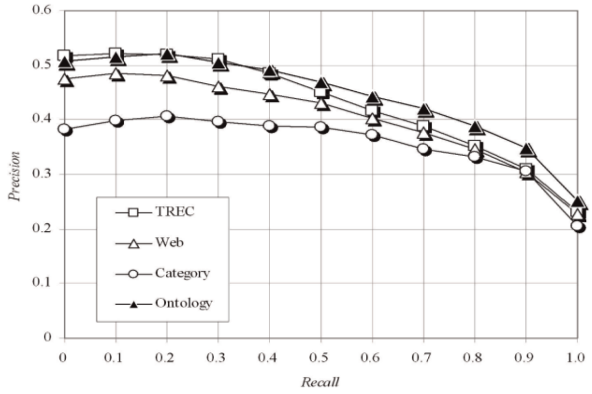


Fig.4 The 11SPR experimental results.

8. 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. A multidimensional ontology mining method, exhaustivity and specificity, is also applied to discover user background knowledge.The model was compared against 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 that shows significant improvement. The ontology model using knowledge with both is-a and part-of semantic relations works better than using only one of them.

The proposed ontology model is a single computational model that discovers background knowledge from both global and local knowledge. In future this model can be applied to the design of web information gathering systems to achieve high performance. Thus it contribute to the fields of Information Retrieval, Recommendation systems, web Intelligence and Information Systems.

The present work assumes data clustering techniques such as genetic algorithm 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 also hopeful to make use of social networking of users to find the area of interest that can help us to improve user profiles in our future work.

9. 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
application libraries Unlimited, 2005.

[6] P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," Proc. ACM SIGIR ('07), pp. 7-14, 2007.

[7] 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.

[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, "Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.

[13] R. Gligorov, W. ten Kate, Z. Aleksovski, and F. van Harmelen, "Using Google Distance to Weight Approximate Ontology Matches," Proc. 16th Int'l Conf. World Wide Web (WWW '07),
pp. 767-776, 2007.

[14] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.

[15] B.J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real Life Information Retrieval: A Study of User Queries on the Web," ACM SIGIR Forum, vol. 32, no. 1, pp. 5-17, 1998.

[16] X. Jiang and A.-H. Tan, "Mining Ontological Knowledge from Domain-Specific Text Documents," Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM '05), pp. 665-668, 2005.

[17] W. Jin, R.K. Srihari, H.H. Ho, and X. Wu, "Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 193-202, 2007.

[18] J.D. King, Y. Li, X. Tao, and R. Nayak, "Mining World Knowledge for Analysis of Search Engine Content," Web Intelligence and Agent Systems, vol. 5, no. 3, pp. 233-253, 2007.

[19] R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, and J.X. Hao, "Towards a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 6,
pp. 800-813, June 2009.

[20] K.S. Lee, W.B. Croft, and J. Allan, "A Cluster-Based Resampling Method for Pseudo-Relevance Feedback," Proc. ACM SIGIR '08,
pp. 235-242, 2008.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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), 767-776, 2007.

.