Chapter 1
ONTOGRAPHY MINING FOR PERSONALIZED SEARCH

Yuefeng Li\textsuperscript{1} and Xiaohui Tao\textsuperscript{2}

Abstract Ontology mining is a cutting edge research field, aims to discover the useful knowledge from a set of data by using an ontology. In this paper, we propose an ontology mining model to discover the topics of a user’s interests from the user’s personalized ontology, in order to interpret and satisfy a user information need. A user information need can be better interpreted if the user’s interests and preferences are specified. In the past, many methods have been proposed to interpret user information needs by using ontologies. However, they are either incomplete or computationally expensive, and need to improve. Our proposed ontology mining model attempts to fill this gap by using world knowledge and a user’s local instance repository. The proposed model is evaluated by applying to a Web information gathering system, and the results are promising.

1.1 Introduction

In the past decades the information available on the World Wide Web has exploded rapidly. Web information covers a great range of topics and serves a broad spectrum of communities. How to gather needed information from the Web, however, becomes a challenging issue.

Personalized Web search is a possible solution to this challenge. It argues that the key to satisfy an information seeker is to understand the seeker, including her (or his) background and information needs. Usually Web users implicitly use concept models to judge the relevance of a document, although they may not know how to represent the models \[9\]. To obtain such a concept model and rebuild it for a user, most people use training sets for knowledge discovery in personalized Web search.

The current methods for acquiring training sets can be grouped into three categories: the interviewing, non-interviewing and pseudo-relevance feedback strate-
gies. The first category is manual techniques and usually involve great efforts by users, e.g. questionnaire and interview. The downside of such techniques is the cost of time, money and users’ patience. The second category, non-interviewing techniques, attempt to guess a user’s interests by observing the user behavior or mining knowledge from the records of the user’s browsing history. These techniques are automated, but the generated user profiles lack accuracy, as too many uncertainties exist in the records. The third category techniques perform a search using a given query first and assume the top retrieved documents as a user’s positive feedback. These documents are then used to generate a user profile. However, these documents still contain a lot of uncertainties. As a result, the user profile is not accurate. In summary, these current techniques need to improve.

In this paper we propose an ontology mining model for personalized Web search by using world knowledge and a user’s local instance repository (LIR). World knowledge is the commonsense knowledge possessed by humans [20], and is also called user background knowledge. A local instance repository (LIR) is a personal collection of information items that were recently visited by a user. These information items include such as text documents, emails, etc, that implicitly cite the knowledge specified in the world knowledge base. Corresponding to a user’s information need, we construct an ontology for the user based on world knowledge, and personalize the ontology by using the user’s LIR. We then perform ontology mining on the personalized ontology and discover the topics of the user’s recent interests related to the information need. The ontology mining model aims to better interpret the semantic meanings of an information need and so to improve the Web search performance. The proposed model is evaluate by applying to a Web information gathering system, against some baselines. The evaluation results are promising. Our ontology mining model contributes to the development and utilization of ontologies and to the improvement of the personalized Web search systems.

The paper is organized as follows. Section 1.2 presents related work and our motivation. The architecture of the proposed ontology mining model is presented in Section 1.3. Section 1.4 presents the background information including world knowledge base and LIRs. In Section 1.5, we describe how to discover knowledge from data and construct an ontology, and in Section 1.6 we present how to mine the topics of user interests from the ontology. The ontology mining model is evaluated and the related issues are discussed in Section 1.7. Finally, Section 1.8 makes conclusions.

1.2 Related Work

Much effort has been invested in semantic interpretation of user topics and concept models. Chirita et al. [1] and Teevan et al. [16] used a collection of a user’s desktop text documents, emails, and cached Web pages, for query expansion or exploration of user interests. Many other works are focused on using user profiles. A user profile is defined by Li and Zhong [9] as the topics of interests related to a user information.
need. They classified Web user profiles into two diagrams: the data diagram and information diagram. A data diagram profile is usually generated by analyzing a database or a set of transactions, e.g., user log data [2, 9, 10, 13]. An information diagram profile is generated by using manual techniques such as questionnaires and interviews or by using the information retrieval techniques and machine-learning methods [10,17]. These profiles can be used for personalized Web services including personalization search and recommendations [2, 3, 9, 17, 21].

Ontologies represent an information diagram profile by a sub-taxonomy of a predefined hierarchy of concepts. Ontologies can provide a basis for the match of initial behavior information and the existing concepts and relations [2, 17]. Li, et al. [7–9, 19] used ontology mining techniques to discover interesting patterns from positive documents, and ontologize the meaningful descriptions to generate a user profile. Gauch et al. [2] used reference ontology based on the categorization of online portals and proposed to learn personalized ontology for users. Sieg et al. [12] modelled a user’s context as an ontological profile and assigned interest scores to the existing concepts in a domain ontology. Developed by King et al. [4], IntelliOnto is built based on the Dewey Decimal Classification and attempts to describe a user’s background knowledge. Unfortunately, these works cover only a small volume of concepts, and do not specify the semantic relationships of partOf and kindOf existing in the concepts but only superClass and subClass.

In summary, there still remains a research gap in semantic study of a user’s interests by using ontologies. Filling this gap in order to better interpret and satisfy a user information need motivates our research work presented in this paper.

1.3 Architecture

The proposed ontology mining model aims to discover the useful knowledge from a set of data by using an ontology. In order to better interpret and satisfy a user information need, we need to obtain a user’s interests and preferences. These knowledge are believed contained in the user’s LIR. However, how to explore an LIR and discover the knowledge from it remain a challenging issue. Firstly, an LIR is just a collection of unstructured or semi-structured text data. There are many noisy data and uncertainties in the collection. Secondly, not all the knowledge contained in an LIR are useful for user information need interpretation. Only the knowledge relevant to the information need are needed. The ontology mining model is adopted to discover such useful knowledge to benefit Web search.

The architecture of the model is presented in Fig. 1.1. The ontology mining model consists of two parts. A given query from a user is the starting point of the ontology mining, as a query is generated by a user as a brief description of his (or her) information need. In the first part, an ontology is constructed based on the world knowledge base corresponding to a given query. The ontology is personalized by using the user’s LIR. Thus, the ontology can describe the user’s background knowledge. However, as aforementioned, not all knowledge described and specified
in the ontology are useful. The second part is to discover the useful knowledge from the constructed ontology, and uses the discovered topics of the user's interests to improve the performance of Web search.

1.4 Background

1.4.1 World Knowledge Base

A world knowledge base is a knowledge frame that formally describes and specifies world knowledge. In this paper, we use the Library of Congress Subject Headings\(^1\) (LCSH) for the base. The LCSH system is a taxonomic classification developed for organizing the large volumes of library collections and for retrieving information from the library. It aims to facilitate users' perspectives in accessing the information items stored in a library. The system is comprised of a thesaurus containing about 400,000 subject headings that cover an exhaustive range of topics. The LCSH is ideal for a world knowledge base as it has semantic subjects and relations specified.

A subject heading in the LCSH is transformed into a knowledge unit, and the LCSH structure forms the taxonomic backbone of the world knowledge base. The LCSH system defines three kinds of references, *BT* (Broader Term) and *NT* (Narrower Term), *UF* (Used-Tor), and *RT* (Related Term). *BT* and *NT* references are for two subjects describing the same entity but at different levels of abstraction (or concretion). They are transformed into the *kindOf* relationships in the world knowledge base. *UF* references describe compound subjects and the subjects subdivided

---

by others, and are transformed into the partOf relationships. KindOf and partOf are both transitive and asymmetric. RT references are for two subjects related in some manner other than by hierarchy, and are transformed into the relatedTo relationships. RelatedTo is symmetric but not transitive. The world knowledge base is formalized as follows:

**Definition 1.** Let $\text{WKB}$ be a taxonomic world knowledge base. It is formally defined as a 2-tuple $\text{WKB} := \langle S, R \rangle$, where

- $S$ is a set of subjects $S := \{s_1, s_2, \cdots, s_m\}$, in which each element is a 2-tuple $s := \langle \text{label}, \sigma \rangle$, where \text{label} is a label assigned by linguists to subject $s$ and is denoted by label($s$), and $\sigma(s)$ is a signature mapping defining a set of relevant subjects to $s$ and $\sigma(s) \subseteq S$;
- $R$ is a set of relations $R := \{r_1, r_2, \cdots, r_n\}$, in which each element is a 2-tuple $r := \langle \text{type}, r_\nu \rangle$, where \text{type} is a relation type of kindOf, partOf, or relatedTo and $r_\nu \subseteq S \times S$. For each $(s_x, s_y) \in r_\nu$, $s_y$ is the subject who holds the type of relation to $s_x$, e.g. $s_x$ is kindOf $s_y$.

### 1.4.2 Local Instance Repository

A local instance repository (LIR) is a collection of Web documents that were recently visited by a user. These documents implicitly cite the knowledge specified in the world knowledge base. In this demonstrated model, we use a set of the library catalogue information items that were accessed by a user recently to represent a user’s LIR. Each item in the catalogue has a title, a table of contents, a summary, and a list of subject headings. An instance in an LIR is represented by a set of terms that are generated from these information after text pre-processing, e.g. stopword removal and word stemming.

The subject headings build the bridge connecting an LIR to the world knowledge base. We denote an information item in an LIR by an instance $i$, and a listed subject heading as a subject $s$. For a given query, let $I = \{i_1, i_2, \cdots, i_p\}$ be an LIR, and $S \subseteq S$ be a set of subjects corresponding to the LIR. The relationships between $S$ and $I$ can be described as the following mappings:

$$
\eta : I \rightarrow 2^S, \quad \eta(i) = \{s \in S | s \text{ is used to describe } i\} \subseteq S; \\
\eta^{-1} : S \rightarrow 2^I, \quad \eta^{-1}(s) = \{i \in I | s \in \eta(i)\} \subseteq I;
$$

where $\eta^{-1}(s)$ is a reverse mapping of $\eta(i)$. These mappings aim to explore the semantic matrix existing between the subjects and instances.

The belief of an instance $i \in I$ to a subject $s \in S$ is measurable. The listed subject headings to an information item are indexed by their importance. Thus, the total

\footnote{Although there are three semantic relationships specified in the LCSH system, in the model presented in this paper we are focused on just kindOf and partOf, since specifying and utilizing these two semantic relationships are a current challenging issue in the related research fields.}
number and the indexes of assigned subjects matter for the belief measuring of a subject. Sometimes, one listed subject heading is formed by several divisions, for instance, “Business intelligence – History – Congresses”. This may cause a problem in matching a listed heading to a subject in the knowledge base. Let $\sigma(s)$ be the accurate level of matching a heading to a subject in the knowledge base. If the heading matches a subject’s label $label(s)$ perfectly, no information is lost and $\sigma(s) = 1$. Starting from the weakest division on the right hand side, if the heading can match a label $(s)$ with one division being cut off, e.g. from “Business intelligence – History – Congresses” to “Business intelligence – History”, $\sigma(s)$ increases by 1. The greater $\sigma(s)$ value indicates more information lost. Thus, let $\xi(i)$ be the number of headings assigned to $i$, $\iota(s)$ be the index of a $s$ on the list (starting from 1), the belief of $i$ to subject $s$ can be calculated by:

$$bel(i,s) = \frac{1}{\iota(s) \times \xi(i) \times \sigma(s)} \quad (1.3)$$

Greater $bel(i,s)$ indicates stronger support of $i$ to $s$.

1.5 Knowledge Discovering from Data

A subject ontology is built based on the world knowledge base and focused on an information need. In Web information search, a query is usually a set of terms generated by a user as a brief description of an information need. For an incoming query $q$, the relevant subjects are extracted from the $S$ in $WKB$ by using the syntax-matching mechanism. We use $sim(s,q)$ to specify the relevance of a subject $s \in S$ to $q$, which is calculated by the size of overlapping terms between $s$ and $q$. If $sim(s,q) > 0$, $s$ is deemed as a positive subject. The $s$’s ancestor subjects in the world knowledge taxonomy, along with their associated semantic relationships $r \in R$, are extracted for ontology construction. By:

$$\mathcal{I}^+ = \{ s | sim(s,q) > 0, s \in \mathcal{I} \}; \quad (1.4)$$

$$\mathcal{I}^- = \{ s | sim(s,q) = 0, s \in \mathcal{I} \}; \quad (1.5)$$

$$\mathcal{R} = \{ < r, (s_1, s_2) > \mid < r, (s_1, s_2) > \in \mathbb{R}, (s_1, s_2) \in \mathcal{I} \times \mathcal{I} \}; \quad (1.6)$$

we can construct a subject ontology $\mathcal{O}(q)$ to simulate a user’s concept model focusing on the given $q$. The formalization of a subject ontology $\mathcal{O}(q)$ is as follows:

**Definition 2.** The structure of a personalized ontology that formally describes and specifies query $q$ is a 4-tuple $\mathcal{O}(q) := \{ \mathcal{I}, \mathcal{R}, \text{tax}, \text{rel} \}$, where

- $\mathcal{I}$ is a set of subjects ($\mathcal{I} \subseteq S$) which includes a subset of positive subjects $\mathcal{I}^+ \subseteq \mathcal{I}$ relevant to $q$, and a subset of negative subjects $\mathcal{I}^- \subseteq \mathcal{I}$ non-relevant to $q$;
- $\mathcal{R}$ is a set of relations and $\mathcal{R} \subseteq \mathbb{R}$;
• *tax*: \( \text{tax} \subseteq \mathcal{I} \times \mathcal{I} \) is called the backbone of the ontology, which is constructed by two directed relationships *kindOf* and *partOf*;
• *rel* is a relation between subjects, where \( \text{rel}(s_1, s_2) = \text{True} \) means \( s_1 \) is *relatedTo* \( s_2 \) and \( s_2 \) is *relatedTo* \( s_1 \) as well.

A sample ontology is constructed corresponding to a query “Economic espionage”\(^3\). A part of the ontology is illustrated in Fig 1.2, where the subjects in dark color are the positive subjects in \( \mathcal{S}^+ \), and the rest are negative in \( \mathcal{S}^- \).

\[ \text{Fig. 1.2 A Constructed Ontology (Partial) for “Economic Espionage”} \]

### 1.5.1 Personalizing an Ontology

In order to interpret a user information need, the constructed ontology needs to be personalized, since a user information need is individual. Our subject ontology is constructed corresponding to a user’s given query. The ontology can be further personalized by using the user’s LIR. A user’s LIR is a collection of information items that were recently accessed by the user, and reflects the user’s recent interests. By using a user’s LIR, we can discover the topics related to the user’s interests, and make the ontology personalized.

\[ \text{Fig. 1.3 The Relationships Between Subjects and Instances} \]

Let \( \mathcal{I} \) be the compliment set of \( \mathcal{S}^+ \) and \( \mathcal{I} = \mathcal{S} - \mathcal{S}^+ \). Based on the mapping Eq. (1.1), each \( i \) maps to a set of subjects. Some instances may map to the subjects in

\(^3\) A query generated by the linguists in Text REtrieval Conference (TREC), http://trec.nist.gov/.
different sets of $\mathcal{F}^+$ and $\mathcal{F}$. Fig. 1.3 illustrates the situation, where $s'_1 \in \mathcal{F}$ overlaps $s_2 \in \mathcal{F}^+$ by its entire mapping instance set ($\{i_3, i_4\}$), and $s'_2 \in \mathcal{F}$ overlaps $s_2$ by part of the instance set ($\{i_4\}$) only. Two subjects can be considered specifying the same semantic topic, in case of mapping to the same instances. Similarly, if their mapping instances overlap, the semantic topics specified by the two subjects may overlap as well. Based on these, we can refine the $\mathcal{F}^+$ and $\mathcal{F}^-$, and have $\mathcal{F}^+$ expanded:

$$\mathcal{F}^+ = \mathcal{F}^+ \cup \{s'|s' \in \mathcal{F}, \eta^{-1}(s') \cap (\bigcup_{s \in \mathcal{F}^+} \eta^{-1}(s)) \neq \emptyset\};$$

$$\mathcal{F}^- = \mathcal{F} - \mathcal{F}^+. \tag{1.7}$$

This expansion is also illustrated in the example displayed in Fig. 1.2, in which the gray subjects are transferred from $\mathcal{F}$ to $\mathcal{F}^+$ by having instances referred by some of the dark subjects. This expansion is based on the semantic study of a user’s LIR, and thus personalizes the constructed ontology for the user.

We also need to evaluate the relevance of the expanded positive subjects to $q$. The evaluation starts with calculating the instances’ coversets:

$$\text{coverset}(i) = \{s|s \in \mathcal{F}, \text{sim}(s, q) > 0, s \in \eta(i)\} \tag{1.8}$$

We then have $\text{sim}_{\text{exp}}$ for the relevance of an expanded positive subject $s'$ by:

$$\text{sim}_{\text{exp}}(s', q) = \sum_{i \in \eta^{-1}(s')} \sum_{s \in \text{coverset}(i)} \frac{\text{sim}(s, q)}{|\eta^{-1}(s)|} \tag{1.9}$$

where $s$ is a subject in the initialized but not the expended $\mathcal{F}^+$. The value of $\text{sim}_{\text{exp}}(s', q)$ largely depends on the $\text{sim}$ values of subjects in $\mathcal{F}^+$ that overlap with $s'$ in their mapping instances.

### 1.6 Useful Knowledge Discovery from Ontology

The constructed personalized ontology describes the implicit concept model possessed by a user related to an information need. The topics of user interests can be discovered from the ontology in order to better interpret the information need.

We use the ontology mining method of Specificity introduced in [15] for the semantic study of a subject in an ontology. Specificity describes a subject’s semantic focus on an information need. The specificity value $\text{spe}$ of a subject $s$ increases if the subject is located on a lower level of an ontology’s taxonomic backbone.

Algorithm 1 presents a recursive method $\text{spe}(s)$ for assigning the specificity value to a subject in an ontology. A subject’s specificity value depends on its child subjects. Specificity aims to assess the strength of a subject in a user’s personalized ontology.

Based on the semantic study of a subject, the strength $\text{sup}$ of an instance $i$ supporting a given query $q$ can be measured by:
preferences. The techniques that are used to generate a user profile can be category.

A user profile is used in personalized Web search to describe a user’s interests and become the key to generate the user’s profile. These topics reflect a user’s recent interests from the user’s personalized ontology. These topics are referred to the topics of the user’s interests, whereas the instances with their sup(i,q) less than supmin refer to the non-relevant topics to q. Therefore, we can have two instance sets I+ and I−, which satisfy

\[ I^+ = \{ i | sup(i,q) > sup_{min}, i \in I \} \]
\[ I^- = \{ i | sup(i,q) < sup_{min}, i \in I \}. \]

Let \( R = \sum_{i \in I} sup(i,q) \), \( r(t) = \sum_{i \in I^+ \setminus I} sup(i,q) \), \( N = |I| \), and \( n(t) = |\{ i \mid i \in I, t \in i \}| \).

We have the following modified probabilistic formula to choose a set of terms from the set of instances to represent the user’s topics of interests:

\[
weight(t) = \log \frac{r(t) + 0.5}{R - r(t) + 0.5} \frac{R - r(t) + 0.5}{(N - n(t)) - (R - r(t)) + 0.5}
\]

The presented ontology mining method discovers the topics of a user’s interests from the user’s personalized ontology. These topics reflect a user’s recent interests and become the key to generate the user’s profile.

1.7 Evaluation

1.7.1 Experiment Design

A user profile is used in personalized Web search to describe a user’s interests and preferences. The techniques that are used to generate a user profile can be catego-
rized into three groups of the interviewing, the non-interviewing, and the pseudo-relevance feedback. A user profile generated by using the interviewing techniques can be called a “perfect” profile, as it is generated manually, and perfectly reflects a user’s interests. One representative of such “perfect” profiles is the training sets used in the TREC-11 2002 Filtering Track (see: http://trec.nist.gov/). They are generated by linguists reading each document through and providing a judgement of positive or negative to the document [11]. The non-interviewing techniques do not involve user efforts directly. Instead, they observe and mine knowledge from a user’s activity and behavior in order to generate a training set to describe the user’s interests [17]. One representative is the OBIWAN model proposed by Gauch et al [2]. Different from the interviewing and non-interviewing techniques, the pseudo-relevance feedback profiles are generated by semi-manual techniques. These group of techniques perform a search first and assume the top 10 (or 20) returned documents as the positive information feedback by a user. The Web training set acquisition method introduced by [14] is a representative of such techniques, which analyzes the retrieved URLs using a belief based method to obtain approximation training sets.

Our proposed model will be compared with the baselines implemented for these representative models. The implementation of the proposed model is called “Onto-based”. Three competitor models are: (i) the TREC model generating the “perfect” user profiles and representing the manual interviewing techniques. It sets a target to mark the achievement of our proposed model; (2) the Web model for the Web training set acquisition method [14] and representing the semi-automated pseudo-relevance feedback methods; and (iii) the Category model for the OBIWAN [2] and representing the automated non-interviewing profiling techniques. Fig. 1.4 il-
lustrates the experiment design. The experimental queries go into the four models, and produce different profiles. A produced user profile is represented by a training set consisting of a positive subset and a negative subset of documents. Each document in a training set is assigned a value indicating the support level of the document to a given query. The user profiles are used by the same Web information gathering system to retrieve relevant documents from the testing data set. The retrieval results are compared and analyzed for evaluation of the proposed model.

The Reuters Corpus Volume 1 (RCV1) [6] is used as the testbed in the experiments. The RCV1 is a large data set of 806,791 documents with great topic coverage. The RCV1 is also used in the TREC-11 2002 Filtering track for experiments. TREC-11 provides a set of topics defined and constructed by linguists. Each topic is associated with some positive and negative documents judged by the same group of linguists [11]. The titles of these topics (R101-125) are used as queries in our experiments.

1.7.2 Details of Experimental Models

The common information gathering system is implemented, based on a model that tends to effectively gathering information by using user profiles [9]. We choose this model in this paper because it is suitable for both perfect training sets and approximation training sets.

Each document in this model is represented by a pattern $P$ which consists of a set of terms ($T$) and the distribution of term frequencies ($w$) in the document ($\beta(P)$).

Let $PN$ be the set of discovered patterns. Using these patterns, we can have a probability function:

$$pr_{\beta}(t) = \sum_{P \in PN, (t,w) \in \beta(P)} \text{support}(P) \times w$$

(1.14)

for all $t \in T$, where $\text{support}(P)$ is used to describe the percentage of positive documents that can be represented by the pattern for the perfect training sets, or the sum of the supports that are transferred from documents in the approximation training sets, respectively.

In the end, for an incoming document $d$, its relevance can be evaluated as

$$\sum_{t \in T} pr_{\beta}(t) \tau(t,d), \quad \text{where} \quad \tau(t,d) = \begin{cases} 1 & \text{if } t \in d \\ 0 & \text{otherwise.} \end{cases}$$

(1.15)

1.7.2.1 Proposed Model: Onto-based Model

The taxonomic world knowledge base is constructed based on the LCSH system, as described in Section 1.4.1. For each query, we extract an LIR through searching the
subject catalogue of Queensland University of Technology Library by using the query, as described in Section 1.4.2. These library information are available to the public and can be accessed for free.

We treat each incoming query as an individual user, as a user may come from any domain. For a given query, the model constructs an ontology first. Only the ancestor subjects away from a positive subject within three levels are extracted, as we believe that any subjects in more than that distance are no longer significant and can be ignored. The Onto-based model then uses Eq. (1.7) and (1.9) to personalize the ontology. The model mines the ontology by using the specificity method and calculates the support values for the corresponding instances by using Eq. (1.10). The $s_{\text{min}}$ is set as 0 in the experiments. The modified probabilistic method (Eq. (1.13)) is used to choose 150 terms to represent the user’s topics of interests. At last, the model determines the training set by using the top 100 retrieved URLs by Google search (using the query). The positive documents are the top 20 ones that are weighted by the total probability function, and the rest URLs group to the set of negative documents.

1.7.2.2 Baseline Models

**TREC Model:** The training sets are manually generated by the TREC linguists. For a coming query, the TREC linguists read a set of documents and marked either positive or negative against each document [11]. Since the queries are also generated by these linguists, the TREC training sets perfectly reflect a user’s concept model, and the support value of each positive document is assigned with 1, and negative with 0. These training sets are thus deemed as “perfect” training sets.

The “perfect” model marks the research goal that our proposed model attempts to achieve. A successful retrieval of user interests and preferences can be confirmed if the performance achieved by the proposed model can match or is close to the performance of the “perfect” TREC model.

**Category Model:** This model represents a typical model using the non-interviewing techniques to generate user profiles. In this model, a user profile is a set of topics related to the user’s interests. Each topic is represented by a vector of terms trained from a user’s browsing history using the $tf \cdot idf$ method. While searching, the cosine similarity value of an incoming document to a user profile is calculated, and higher similarity value indicates that the document is more interesting to the user. In order to make the comparison fair, we used the same LIRs in the Onto-based model as the collection of a user’s Web browsing history in this model.

**Web Model:** This model represents a typical model using the pseudo-relevance feedback mechanism to generate a user profile. In this experimental model, the user profiles (training sets) are automatically retrieved from the Web by employing a Web search engine. For each incoming query, a set of positive concepts and a set of negative concepts are identified manually. By using Google, we retrieved a set of positive and a set of negative documents (100 documents in each set) using the

---

4 [http://library.qut.edu.au](http://library.qut.edu.au)
identified concepts (the same Web URLs are also used by the Onto-based model). The support value of a document in a training set is defined based on (i) the precision of the chosen search engine; (ii) the index of a document on the result list, and (iii) the belief of a subject supporting or against a given query. This model attempts to use Web resources to benefit information retrieval. The technical details can be found in [14].

1.7.3 Performance Assessment Methods

The performance is assessed by two methods: the precision averages at eleven standard recall levels, and $F_1$ Measure. The former is used in TREC evaluation as the standard for performance comparison of different information filtering models [18]. A recall-precision average is computed by summing the interpolated precisions at the specified recall cutoff and then dividing by the number of queries:

$$\frac{\sum_{i=1}^{N} \text{precision}_\lambda}{N}. \quad (1.16)$$

$N$ denotes the number of experimental queries, and $\lambda = \{0.0, 0.1, 0.2, \ldots, 1.0\}$ indicates the cutoff points where the precisions are interpolated. At each $\lambda$ point, an average precision value over $N$ queries is calculated. These average precisions then link to a curve describing the precision-recall performance. The other method, $F_1$ Measure [5], is well accepted by the community of information retrieval and Web information gathering. $F_1$ Measure is calculated by:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1.17)$$

Precision and recall are evenly weighted in $F_1$ Measure. The macro-$F_1$ Measure averages each query’s precision and recall values and then calculates $F_1$ Measure, whereas the micro-$F_1$ Measure calculates the $F_1$ Measure for each returned result in a query and then averages the $F_1$ Measure values. The greater $F_1$ values indicate the better performance.

1.7.4 Experiment Results and Discussions

In terms of precision and recall, the experimental results are displayed in Fig. 1.5, which is a chart for the precision averages at eleven standard recall levels. Table 1.1 presents the $F_1$ Measure results comparison results. The figures in “Improvement” are calculated by using the average $F_1$ Measure results of the Onto-based model to minus the baselines. The percentages displayed in “% Change” indicate the significance level of improvement achieved by the Onto-based model over the baselines,
<table>
<thead>
<tr>
<th>Model</th>
<th>Macro-F1 Measure</th>
<th>Micro-F1 Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Improvement</td>
</tr>
<tr>
<td>TREC</td>
<td>0.3944</td>
<td>-0.0061</td>
</tr>
<tr>
<td>Web</td>
<td>0.382</td>
<td>0.0063</td>
</tr>
<tr>
<td>Category</td>
<td>0.3715</td>
<td>0.0168</td>
</tr>
<tr>
<td>Onto-based</td>
<td>0.3883</td>
<td>-</td>
</tr>
</tbody>
</table>

which is calculated by:

\[
\% \text{ Change} = \frac{F_{\text{Onto-based}} - F_{\text{baseline}}}{F_{\text{baseline}}} \times 100\%.
\]  

Fig. 1.5 The 11 Standard Recall-Precision Results

Fig. 1.6 The Percentage Changes Achieved by the Onto-based Model over the Baseline Models
The experiments conducted for the Onto-based and TREC models are to compare the user interests discovered by the proposed model to the knowledge completely specified by linguists manually in the TREC model. According to the results illustrated in Fig. 1.5, the Onto-based model has achieved the same performance as the perfect TREC model at most of the cutoff points (0-0.2, 0.4-0.6, 0.9-1.0). The “perfect” TREC training sets are more precise in comparison with the Onto-based training sets, as they are generated manually. However, the TREC training sets may not cover the substantial relevant semantic space than the Onto-based training sets. The Onto-based model has about average 1000 documents in an LIR/per query for the discovery of interest topics. In contrast, The number of documents included in each TREC training set is very limited (about 60 documents per query in average). Some semantic meanings referred by a given query are not fully covered by the TREC training set. In comparison, the Onto-based model training sets cover much broader semantic extent, although the expert knowledge contained by TREC sets is more precise. As a result, the Onto-based model’s precision-recall performance has been close to the TREC model.

The close performance to the perfect TREC model achieved by the Onto-based model is confirmed by the $F_1$ Measure results. According to the average $F_1$ Measure results presented in Table. 1.1 and Fig. 1.6, the TREC model outperforms the Onto-based model slightly by only about 1.55% in Macro-$F_1$ and 1.72% in Micro-$F_1$ Measure. The performance of proposed model is close to the target model. Considering that the TREC model employs the human power of linguists to read every single document in the training set, which reflects a user’s concept model perfectly, the close performance to the TREC model achieved by the Onto-based model is still promising.

The experiments for the Onto-based and Category models are to compare the proposed model to the automated user profiling techniques using ontology. According to Fig. 1.5, 1.6 and Table 1.1, the Onto-based model outperforms the Category model. In average, the Onto-based model improves from the Category model by 4.52% in Macro-$F_1$ and 3.69% in Micro-$F_1$ Measure. The Onto-based model specifies the concepts in the personalized ontology by using the complex semantic relations of $kindOf$ and $partOf$, and analyzes the subjects by using the ontology mining method. In contrast, the Category model specifies only the simple relation of $superClass$ and $subClass$. The Onto-based model performs in more technical depth in comparison with the Category model, and moves far beyond the simple $superClass$ and $subClass$ specification. Furthermore, the specificity ontology mining method appreciates the subject’s location in the ontology backbone, which is more realistic. Based on these, we can say that the proposed model can describe knowledge more accurately than the Category model.

The experiment for the Onto-based and Web models are to compare the world knowledge extracted by the proposed method to the commonsense knowledge extracted by the Web model. As shown in Fig. 1.5, 1.6 and Table 1.1, the Onto-based model outperforms the Web model slightly. On average, the improvement achieved by the Onto-based model from the Web model are 1.65% in Macro-$F_1$ and 1.46% in Micro-$F_1$ Measure. The Web model’s training sets are extracted from the Web.
The Web documents, however, are not formally specified. Comparing to the Web model training sets, the Onto-based training sets integrate the world knowledge and the user interests from the LIRs. Considering both two models extract world knowledge, the user interests contained in the LIRs leverages the Onto-based model’s performance. Based on these, we conclude that the proposed model can improve the performance of Web information gathering from the Web model.

Based on the experimental results, we can conclude that the proposed personalized ontology model is promising.

1.8 Conclusions

Ontology mining is a cutting edge research field, aims to discover the useful knowledge from a set of data by using an ontology. In this paper, we have proposed a ontology mining model to discover the useful knowledge from a user’s LIR. The discovered knowledge is the topics related to a user’s interests, and is used to better interpret a user information need. Corresponding to a user’s given query, an ontology is constructed. The ontology is then personalized by using the user’s recently accessed documents in the LIR. By mining the personalized ontology we discover the useful knowledge as the topics related to the user’s interests. The model is evaluated and the results are promising. The ontology mining model benefits the interpretation of an information need and so Web search, and contributes to the knowledge discovery using ontologies.

References

15. _, *Ontology mining for personalized web information gathering*, Proc. of the IEEE/WIC/ACM intl. conf. on Web Intelligence (Silicon Valley, USA), Nov. 2007, pp. 351–358.