Ontology Mining for Personalized Web Information Gathering

Xiaohui Tao, Yuefeng Li, Ning Zhong*, Richi Nayak
Faculty of Information Technology, Queensland University of Technology, Australia
*Department of Systems and Information Engineering, Maebashi Institute of Technology, Japan
{x.tao, y2.li, r.nayak}@qut.edu.au, *zhong@maebashi-it.ac.jp

Abstract

It is well accepted that ontology is useful for personalized Web information gathering. However, it is challenging to use semantic relations of "kind-of", "part-of", and "related-to" and synthesize commonsense and expert knowledge in a single computational model. In this paper, a personalized ontology model is proposed attempting to answer this challenge. A two-dimensional (Exhaustivity and Specificity) method is also presented to quantitatively analyze these semantic relations in a single framework. The proposals are successfully evaluated by applying the model to a Web information gathering system. The model is a significant contribution to personalized ontology engineering and concept-based Web information gathering in Web Intelligence.

1. Introduction

Ontology is a formal description and specification of knowledge. It provides a common understanding of topics to be communicated between users and systems [2]. By using an ontology, information systems are expected to be able to understand the semantic meaning of words and phrases, and be able to compare information items by concepts instead of keywords [9]. Ontology is deemed by the Web Intelligence community as one of the most useful techniques for Web information gathering.

Over the last decade, many attempts have been suggested to learn ontology in order to describe and specify the knowledge possessed by humans. Li & Zhong [7] proposed to discover the backbone of an ontology based on the patterns found in documents. Gauch et al. [3] proposed to learn personalized ontology based on the online portals. King et al. [4] proposed to learn ontology based on the Dewey Decimal Classification (DDC)[1]. However, these existing works only specify "super-" and "sub-class" relations in the ontology, and do not extend beyond the ontology learning framework proposed by Maedche [10]. Maedche's ontology learning framework consists of four phases: Import, Extract, Prune, and Refine. The Maedche's framework, however, has a pitfall of relying on the manpower of ontology engineers heavily, e.g. for an incoming lexical entry, the engineer needs to manually determine either assigning it to an existing concept or defining a new concept for it. Consequently, these ontology methods are either incomprehensive or expensive in knowledge acquisition.

Web users possess a concept model in the process of information gathering. Usually, users can easily determine if a Web page interests them or not while they read through the content. The rationale behind this is that users implicitly possess a concept model based on their knowledge, although they may not be able to express it [7]. There exists a potential that by describing and specifying this concept model, the semantic meaning of a user's information need can be well interpreted. In this paper, a personalized ontology model is proposed, which extract the commonsense knowledge possessed by the user in her concept model and the expert knowledge revising the concept model. The model synthesizes these two kinds of knowledge and formally specifies the semantic relations of "kind-of", "part-of", and "related-to" in a single computational model, instead of simple "super-" and "sub-class" in the existing models [3, 4, 7]. In this paper, a two dimensional method, *Specificity* and *Exhaustivity*, is also presented to analyze these semantic relations in order to discover knowledge from the learnt personalized ontology and to use the ontology for personalized Web information gathering. The proposed model is evaluated by assessing its applications to a system that gathers information from a large corpus. The model is a significant contribution to personalized ontology engineering and concept-based personalized Web information gathering in Web Intelligence.

The paper is organized as follows. Section 2 is the problem statement. Section 3 introduces the personalized ontology learning method attempting to formally ontologize a user's concept model. Section 4 presents the two dimen-

[1] Dewey Decimal Classification, http://www.oclc.org/dewey/.

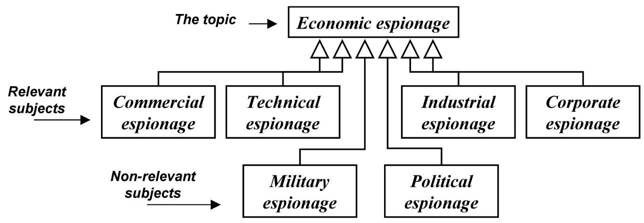


Figure 1. A User Concept Model

sional method for ontology mining. Section 5 describes the related user profiling for personalized Web information gathering, and Section 6 discusses the evaluation. Finally, Section 7 presents the related work, and Section 8 makes the conclusions.

2. The Research Problem

The personalized Web information gathering is a difficult task. A great challenge is that the semantic meaning of a user's information need (called a topic in this paper) is difficult to interpret. One example is "Economic espionage", which is a topic generated by the linguists in TREC[2]. It comes with a description of "*What is being done to counter economic espionage internationally?*" and a narrative of "*Documents which identify economic espionage cases and provide action(s) taken to reprimand offenders or terminate their behavior are relevant. Economic espionage would encompass commercial, technical, industrial or corporate types of espionage. Documents about military or political espionage would be irrelevant*". A concept model for the topic may be manually constructed, as illustrated in Fig. 1, consisting of various relevant or non-relevant subjects, according to these linguist generated specifications. However, it is hard for general users to specify such adequate description and narrative. Even this manually generated concept model could still be incomplete, because some important subjects may be missed out, and some semantic relations between the subjects may be overlooked, e.g. the semantic relation existing between "technical espionage" and "industrial espionage" in Fig. 1. The personalized Web information gathering is a challenge in Web Intelligence.

The research work presented in this paper attempts to answer the challenge by proposing a personalized ontology learning and mining framework. The framework consists of five phases: (i) Building a taxonomic world knowledge base; (ii) Constructing the personalized ontology backbone by interacting with a user; (iii) Extracting expert knowledge to revise the ontology automatically; (iv) Mining the ontology by analyzing the semantic relations; and (v) Generating the personalized user profile.

3. Personalized Ontology Learning for a Specific Topic

The personalized ontology can describe different concept models for different users, although they may have the same topic. In order to do so, we argue that two kinds of knowledge are required: world knowledge covering large number of topics so that the user's individual information need can best match, and expert knowledge revising the concept model. World knowledge is the commonsense knowledge possessed by humans and "the kind of knowledge that humans acquire through experience and education" [19]. Expert knowledge is the kind of knowledge classified by the people who hold expertise in that domain. The difficulty in world knowledge extraction is the topic coverage and semantic relation specification, whereas in expert knowledge extraction is the efficiency, since by traditional means expert knowledge is extracted by experts reading a set of documents manually. In this section, we are going to propose a method to extract the world knowledge and expert knowledge automatically.

3.1 World Knowledge Representation

A taxonomic world knowledge base with great coverage of topics is superior of backbone learning for an ontology. The Library of Congress Subject Headings[3] (LCSH) classification is a system developed for organizing large volume of information stored in a library. It comprises a thesaurus of subject headings exhaustively covering a large number of topics in the world (contains 299,000 records according to the retrospective of 1986-2006). The LCSH system specifies the semantic relations existing in the subject headings, and facilitates the user's perspectives in accessing the information items in a library catalogue. Based on the LCSH system, a taxonomic world knowledge base can be constructed by forming each subject heading a class node and using the specified semantic relations as the links between the nodes. The taxonomic knowledge base is formalized as follows.

Definition 1. *Let $OntoBASE$ be a taxonomic ontology base. An ontology base is formally defined as a 2-tuple $OntoBASE :=< \mathbb{S}, \mathbb{R} >$, where*

- *$\mathbb{S}$ is a set of subjects $\mathbb{S} := \{s_1, s_2, \cdots, s_m\}$;*
- *$\mathbb{R}$ is a set of relations $\mathbb{R} := \{r_1, r_2, \cdots, r_n\}$.*

Definition 2. *A subject $s \in \mathbb{S}$ is formalized as a 3-tuple $s :=< label, instanceSet, \sigma >$, where*

- *$label$ is a label assigned by linguists to a subject s in the LCSH system. The label of s is denoted by $label(s)$;*

[2] Text REtrieval Conference, http://trec.nist.gov/.

[3] The Library of Congress, http://www.loc.gov/.

- *$instanceSet$ is a set of objects associated to a subject s, in which each element specifies a semantic meaning referring by s and is called an $instance$ (see Definition 5 for more details);*

- *σ is a signature mapping ($\sigma : s \rightarrow 2^s$) that defines a set of relevant subjects to a given s.*

Definition 3. *A relation $r \in \mathbb{R}$ is a 2-tuple $r :=< type, r_\nu >$, where*

- *$type$ is a set of relationships, $type = \{kindOf, partOf, relatedTo\}$;*

- *$r_\nu \subseteq \mathbb{S} \times \mathbb{S}$. For each $(x, y) \in r_\nu$, y is the subject who holds the $type$ of relation to s, e.g. s_y is $kindOf$ s_x.*

$KindOf$ is a directed relation in which one subject is in different form of another subject. The property of $kindOf$ is Transitivity and Asymmetry. Transitivity means if s_1 is a kind of s_2 and s_2 is a kind of s_3, then s_1 is a kind of s_3 as well. Asymmetry means if s_1 is a kind of s_2 and $s_1 \neq s_2$, s_2 may not be a kind of s_1 necessarily. One example is that "Business ethics" is a $KindOf$ "Professional ethic", and "Professional ethic" is a $KindOf$ "Ethics". Then "Business ethics" is also a $KindOf$ "Ethics" as well. However, these relations can not be inverse.

$PartOf$ is a directed relation used to describe the relationship held by a compound subject class and its component classes, e.g. subject s_1 forms a part of s_2. The $partOf$ relationship also holds the properties of transitivity and asymmetry. If s_1 is a part of s_2 and s_2 is a part of s_3, then s_1 is also a part of s_3. If s_1 is a part of s_2 and $s_1 \neq s_2$, s_2 is definitely not a part of s_1. One example in the knowledge based is that "Economic espionage" is a $partOf$ of "Business intelligence". The latter can not be a $partOf$ the former.

$RelatedTo$ is a non-taxonomic relation describing the relationship held by two subjects that overlap in their semantic spaces. $relatedTo$ holds the property of symmetry. If s_1 is related to s_2, s_2 is also related to s_1. One example in the knowledge base is "Business intelligence" and "Confidential business information".

A personalized ontology facilitating a user's concept model needs to be dynamically constructed in response to the change of information need. For this purpose, a tool called ontology learning environment interacting with the user is developed to help study a specific information need. The tool analyzes a specific topic, retrieves the possible relevant subjects from the knowledge base and presents them to the user. The user interacts with the tool and identifies the positive and negative (ambiguous) subjects according to the topic and the possessed concept model. The subject based personalized ontology is then built based on the user feedback and the taxonomic knowledge base. Fig. 2 shows an incomplete ontology constructed for the topic "Economic espionage", where the white nodes are the positive subjects, the dark gray are the negative, and the light gray are the unlabelled subjects. The unlabelled subjects are those in the volume of a positive subject but not being identified by the user as either positive or negative. The semantic relations existing between the subjects are addressed by different type of lines. The personalized ontology is formalized by the following definition.

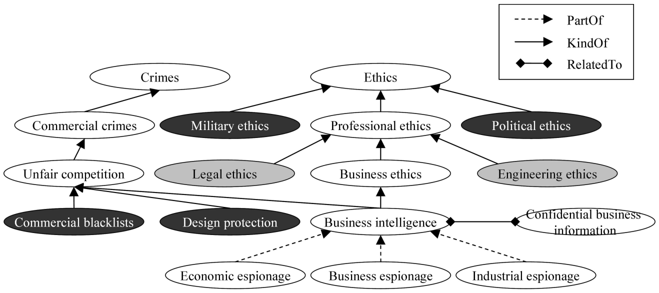


Figure 2. A Constructed Ontology

Definition 4. *The structure of an ontology that formally describes and specifies topic $\mathcal{T}$ is a 5-tuple $\mathcal{O}(\mathcal{T}) := \{\mathcal{S}, \mathcal{R}, tax^{\mathcal{S}}, rel, \mathcal{A}^{\mathcal{O}}\}$, where*

- *$\mathcal{S}$ is a set of subjects and $\mathcal{S} \subseteq \mathbb{S}$. $\mathcal{S}$ has three subsets, where $\mathcal{S}^+ \subseteq \mathcal{S}$ is a set of positive subjects to $\mathcal{T}$, $\mathcal{S}^- \subseteq \mathcal{S}$ is a set of negative subject to $\mathcal{T}$, and $\mathcal{S}^\diamond \subseteq \mathcal{S}$ is a set of unlabelled subjects to $\mathcal{T}$;*

- *$\mathcal{R}$ is a set of relations and $\mathcal{R} \subseteq \mathbb{R}$;*

- *$tax^{\mathcal{S}}$: $tax^{\mathcal{S}} \subseteq \mathcal{S} \times \mathcal{S}$ is a taxonomic backbone of the ontology, which consists of two directed relations $kindOf$ and $partOf$;*

- *rel is a function defining non-taxonomic relations;*

- *$\mathcal{A}^{\mathcal{O}}$ is a set of rules mined from $\mathcal{O}$.*

Given a pair of subjects (s_1, s_2), its $dom(s_1, s_2)$ refers to their least common ancestor subject in $tax^{\mathcal{S}}$. Given a subject s, its $vol(s)$ refers to the union of all subjects in its volume. For $partOf(tax^{\mathcal{S}}) = (s_1, s_2)$ one may also write $partOf(s_1, s_2)$, which means that s_1 is a part of s_2. For $kindOf(tax^{\mathcal{S}}) = (s_1, s_2)$ one may also write $kindOf(s_1, s_2)$, which means that s_1 is a kind of s_2.

3.2 Expert Knowledge Discovery

An ontology requires expert knowledge to fill the taxonomic backbone by instances [1]. Usually, each information item in a library catalogue is described by some brief information, e.g. title and table of content provided by the author and the summary generated by the linguists. The author is the one who produces this information item, and the linguists are experts who are trained to summarize these information items. The expert knowledge is underlying in these brief information. In a library using the LCSH system for organization, each stored information item is also assigned with one or more LCSH subject headings specifying the topics discussed by the information item. These subject headings provide a bridge connecting the expert classified information and the taxonomic world knowledge base, and thus for automatic expert knowledge extraction. The library catalogue provides a great resource of expert knowledge discovery.

Each information item in a library catalogue forms an instance $inst$ in the ontology. The instance is represented by a term vector after stopword removal, word stemming and grouping. The belief bel of an $inst$ to a s can be determined by:

$$bel(inst, s) = \frac{1}{index(s) \times n} \tag{1}$$

where n is the number of subjects associated to the instance, $index(s)$ is the index of a subject starting from 1 in a list of subjects associated. Based on this, the more subjects being assigned to an information item will decrease the item's belief to each assigned subject, due to the loss of focus. A subject being assigned to an item at a higher index will increase the belief of the item to the subject, since the item is more relevant to the subject than others with lower indexes.

The *instanceSet* associated to a subject can be formalized as follows, based on the subjects in the ontology backbone and the associated information items:

Definition 5. *Given a subject s of $\mathcal{O}$, its instanceSet $\eta(s) \subseteq \Omega$ is a set of related instances that are relevant to the subject in certain extent, which satisfies the following mapping, where Ω is the set of all instances and min_bef is the minimum value of $bel(inst, s)$:*

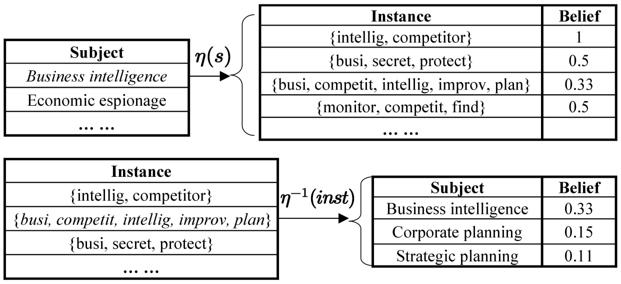
$$\eta(s) = \{inst \in \Omega | bel(inst, s) \geqslant min\_bef\}. \tag{2}$$

Based on this definition, given a subject s, a set of instances (e.g. positive documents) can be extracted. The *instanceSet* also has its reverse mapping $\eta^{-1}(inst)$:

$$\eta^{-1}(inst) = \{s \in \mathcal{S} | bel(inst, s) \geqslant min\_bef\}. \tag{3}$$

Based on the reverse mapping, given an instance, a set of subjects can be identified. Fig. 3 illustrates a sample mapping related to the topic "Economic espionage", along with the belief assigned to the instances.

The expert knowledge referred by a topic can be extracted based on the ontology and the related mappings. Given a topic $\mathcal{T}$, a set of instances that support the topic can be identified based on the user identified positive subjects $\mathcal{S}^+$. The belief of an instance to a topic relies on the number of its mapping positive and negative subjects and their beliefs, and can be determined by:



Subject
Business intelligence
Economic espionage
… …

Instance	Belief
{intellig, competitor}	1
{busi, secret, protect}	0.5
{busi, competit, intellig, improv, plan}	0.33
{monitor, competit, find}	0.5
… …	

Instance
{intellig, competitor}
{*busi, competit, intellig, improv, plan*}
{busi, secret, protect}
… …

Subject	Belief
Business intelligence	0.33
Corporate planning	0.15
Strategic planning	0.11

Figure 3. Mappings of Subject and Instance

$$\begin{aligned} bel(inst, \mathcal{T}) = & \sum_{s \in \eta^{-1}(inst), s \in \mathcal{S}^+} bel(inst, s) \\ & - \sum_{s \in \eta^{-1}(inst), s \in \mathcal{S}^-} bel(inst, s) \end{aligned} \tag{4}$$

If $bel(inst, \mathcal{T}) > 0$, the instance supports the topic. Otherwise, it is against the topic or makes the topic more confusing. The instances associated to an unlabelled subject count nothing to the topic because there is no evidence that they appreciate any site of positive or negative. Similarly, the belief of a subject supporting (or against) a topic depends on the beliefs of its associated instances to the topic:

$$bel(s, \mathcal{T}) = \sum_{inst \in \eta(s)} bel(inst, \mathcal{T}) \tag{5}$$

If $bel(s, \mathcal{T}) > 0$, the subject supports the topic. Otherwise, it is against the topic or makes it more confusing. Eq. (5) amplifies the strength of an instance supporting (or against) a subject according to a topic. The greater $bel(s, \mathcal{T})$ value makes the support (or confusion) stronger. Again, the unlabelled subjects hold belief value of 0 to the topic because their beliefs can not be clarified. Comparing to the positive and negative subjects identified by a user previously, these subjects may be called "confirmed" positive and negative subjects, and those user identified subjects may be called "candidate" subjects. The confirmed subjects revise the candidates by adding the expert knowledge into the ontology learning.

4. Personalized Ontology Mining

Ontology mining is a process of discovering knowledge from the ontology backbone and the associated instances. A two dimensional method is introduced here for mining an ontology. *Exhaustivity* (*exh* for short) describes the semantic extent covered by a subject referring to a topic; and

Specificity (*spe* for short) describes the semantic focus of a subject referring to a topic. The two dimensional method aims to analyze the semantic relations held by the subjects existing in the ontology referring to a topic. A subject in the ontology may be deemed highly exhaustive, although it may be not specific to the topic. In contrast, a subject may be highly specific, although it may deal with only a few aspects of the topic.

A subject's exhaustivity is affected by the number of subjects that are covered in its volume and the belief of these subjects to the topic:

$$exh(s,\mathcal{T}) = \sum_{s^{'} \in vol(s)} bel(s^{'},\mathcal{T}) \tag{6}$$

The semantic extent spreads if more subjects appear in its volume and more details these subjects hold. A subject with the positive exhaustivity value makes the semantic meaning of the topic clearer, and a subject with the negative exhaustivity value makes it more confusing. Exhaustivity can be used to refine the process of expert knowledge extraction for a topic, e.g. the positive exhaustive subjects for the extraction of positive training set, and the negative exhaustive subjects for the negative training set.

The specificity of a subject is affected by some factors. Firstly, the specificity increases if more instances refer to the subject, and if greater belief of these instances are to the topic. Secondly, the specificity decreases if a subject locates at a higher level in the taxonomy, since its description becomes more abstractive, e.g. from "Economic espionage" to "Business intelligence" in Fig. 2. Thirdly, a subject's semantic relations with its peers may impact the specificity. If a subject s is combined by a number of n subjects (each one holds the semantic relation $partOf(s_i, s)$ with s, $i = 1 \ldots n$), it holds only one nth of focus held by s_i, e.g. "Business intelligence" holds less focus than "Economic espionage". Based on these, the specificity of a subject is defined by:

$$spe(s,\mathcal{T}) = bel(s,\mathcal{T}) \times \delta(s) \tag{7}$$

where $s \in \mathcal{S}$ in $\mathcal{O}(\mathcal{T})$, and δ is a relative parameter between (0,1] applied by the semantic relation held by s and its peers. For a leaf subject, $\delta(s) = 1$; for $PartOf(s_i, s)$ relation and $i = 1 \ldots n$, $\delta(s) = \frac{1}{n} \times \delta(s_i)$; for $kindOf(s_1, s)$ or $relatedTo(s_1, s)$ relation, $\delta(s) = \delta(s_1) \times \theta$, where θ is a threshold between (0,1] determining how much specificity inherited or overlapped from s_1, e.g. 0.9. If $s = dom(s_1, s_2)$, or s has a mix of $kindOf$, $partOf$, and $relatedTo$ relations with the peers, the least $\delta(s)$ takes place. The specificity is used to determine the strength of a subject supporting or against a topic.

A few theorems can then be proposed based on the definitions of specificity and exhaustivity.

Theorem 1. *Let* $\{s_1, s_2\} \subseteq \mathcal{S}$ *in* $\mathcal{O}(\mathcal{T}), s_1 \in vol(s_2)$*, and* $bel(s_1,\mathcal{T}) \geqslant bel(s_2,\mathcal{T})$*, we always have*

$$spe(s_1,\mathcal{T}) > spe(s_2,\mathcal{T}).$$

Proof. From Eq. (7), let $\{s_1, s_2\} \subseteq \mathcal{S}$, s_1 and s_2 hold the relation of $kindOf(s_1, s_2)$ or $partOf(s_1, s_2)$, and $bel(s_1,\mathcal{T}) \geqslant bel(s_2,\mathcal{T})$, we have:

$$\begin{aligned}
&spe(s_1,\mathcal{T}) - spe(s_2,\mathcal{T}) \\
=&(bel(s_1,\mathcal{T}) \times \delta(s_1)) - (bel(s_2,\mathcal{T}) \times \delta(s_2)) \\
\because\ &(kindOf(s_1,s_2) \text{ or } partOf(s_1,s_2)) \Rightarrow \delta(s_1) > \delta(s_2); \\
&\&\ bel(s_1,\mathcal{T}) \geqslant bel(s_2,\mathcal{T}) \\
\therefore\ &spe(s_1,\mathcal{T}) - spe(s_2,\mathcal{T}) > 0.
\end{aligned}$$

Theorem 2. *Let* $\{s_1, s_2\} \subseteq \mathcal{S}, vol(s_1) \subset vol(s_2) \subseteq S^+$ *in* $\mathcal{O}(\mathcal{T})$*, we always have*

$$exh(s_1,\mathcal{T}) < exh(s_2,\mathcal{T}).$$

Proof. From Eq. (6), let $\{s_1, s_2\} \subseteq \mathcal{S}, vol(s_1) \subset vol(s_2) \subseteq S^+$ in $\mathcal{O}(\mathcal{T})$, let $\wedge$ denotes the logic "and", we have:

$$\begin{aligned}
&exh(s_2,\mathcal{T}) - exh(s_1,\mathcal{T}) \\
=&\sum_{s^{'} \in vol(s_2)} bel(s^{'},\mathcal{T}) - \sum_{s^{''} \in vol(s_1)} bel(s^{''},\mathcal{T}) \\
=&\sum_{s^{'''} \in vol(s_2) \wedge s^{'''} \notin vol(s_1)} bel(s^{'''},\mathcal{T}) \\
\because\ &(s^{'''} \in vol(s_2)) \wedge (vol(s_2) \subseteq S^+) \Rightarrow bel(s^{'''},\mathcal{T}) > 0 \\
\therefore\ &exh(s_2,\mathcal{T}) - exh(s_1,\mathcal{T}) > 0
\end{aligned}$$

Theorem 3. *Let* $\{s, s^{'}\} \subseteq \mathcal{S}$ *and* $s^{'} \in vol(s)$ *in* $\mathcal{O}(\mathcal{T})$*, we have that* $exh(s,\mathcal{T})$ *increases by the way* $spe(s^{'},\mathcal{T})$ *increases, and vise versa.*

Proof. From Eq. (6) and (7), let $\{s, s^{'}\} \subseteq \mathcal{S}$ and $s^{'} \in vol(s)$ in $\mathcal{O}(\mathcal{T})$, we have:

$$\begin{aligned}
exh(s,\mathcal{T}) &= \sum_{s^{'} \in vol(s)} bel(s^{'},\mathcal{T}) \\
&= \sum_{s^{'} \in vol(s)} \frac{bel(s^{'},\mathcal{T}) \times \delta(s^{'})}{\delta(s^{'})} \\
&= \sum_{s^{'} \in vol(s)} \frac{spe(s^{'},\mathcal{T})}{\delta(s^{'})}
\end{aligned}$$

$\because$ The position of $s^{'}$ is fixed in $vol(s)$, and $\delta(s^{'}) > 0$
$\therefore$ $exh(s,\mathcal{T})$ increases where $spe(s^{'},\mathcal{T})$ increases, and $exh(s,\mathcal{T})$ decreases where $spe(s^{'},\mathcal{T})$ decreases.
A subject is of high exhaustivity value if the subjects in its volume are highly specific.

5. User Profiling for Web Information Gathering

A user profile is the descriptions of the concept model possessed by the user [7]. In terms of Web information gathering, a user profile is the semantic interpretation of a topic based on the user possessed concept model. The subject based personalized ontology provides a basis for the user profile generating.

The user profile is represented by a set of training documents in this paper, instead of a set of keywords or patterns by traditional means [3, 7]. Training sets are commonly used in Web data mining and text classification to represent knowledge [16]. A training set usually consists of a set of positive documents, a set of negative documents, and sometimes a set of unlabelled documents. Traditionally, the experts are needed to read a set of text documents and provide feedbacks of either positiveness or negativeness of each document according to the given topic. This technique is expensive because of manual effort involved. In this paper, a training set is generated to represent the user profile by using the proposed personalized ontology model. Each document d_{inst} in the training set is generated by an instance $inst$, which holds a specific value of supporting, against, or unlabelled to the given topic. The specificity of d_{inst} is determined by:

$$spe(d_{inst}) = bel(inst, \mathcal{T}) \times \sum_{s \in \eta^{-1}(inst)} spe(s, \mathcal{T}) \qquad (8)$$

where $s \in \mathcal{S}$ in $\mathcal{O}(\mathcal{T})$. The documents with a positive value go to the positive set, with a negative value go to the negative set, and with zero go to the unlabelled training set. The benefit of this representation is that the underlying expert knowledge can be mined later on by the system using the user profile. The system can have the maximal flexibility of choosing any method (information retrieval, text classification, or data mining techniques, etc.) to mine the expert knowledge from the training set in order to achieve the best performance.

6. Evaluation

The proposed model is evaluated by assessing the success of its application to a Web information gathering system. In response to a given topic, the user profiles (training sets) are generated by the proposed model and the state-of-the-art baselines. The profiles are input into a common system and used to train the system for information gathering. The performance of the system is determined by the quality of input training sets, where the information gathering method remains the same. By comparing the gathering results, the proposed model can then be evaluated quantitatively.

The experiment design is as follows. The Web information gathering system is implemented based on Li & Zhong's model (see [7] for technical details), including the basic text processing (e.g. stopword removal, word stemming and grouping). For generating the training sets, three models are implemented:

TREC model The training sets are manually generated by the TREC linguists who read each document and mark it either positiveness or negativeness according to a topic [13]. These training sets reflect a user's concept model perfectly, and may be deemed as the "perfect" sets;

Web model The training sets are automatically generated from the Web (see [16] for technical details). The model analyzes a given topic and identifies the relevant subjects, then uses the subjects to gather a set of Web documents by using a selected Web search engine (Google is chosen for the experiments as it has become the most popular search engine nowadays[4]). The model then measures the certainty of each document supporting/against the topic and assigns a float type of positive (or negative) judgment to the document. These documents then become the input training set to the Web information gathering system;

Ontology model The training sets are generated as described in Section 5, by using the personalized ontology model proposed in this paper. A large volume (138MB) of information stored in the catalogue of a library[5] is used, which contains 448,590 documents and 162,751 unique terms.

The Reuters Corpus Volume 1 (RCV1) is used as the testbed, which is the official testbed used in TREC-11 2002 and an archive of 806,791 documents. TREC-11 has topics designed by linguists and associated with the training sets and testing sets. These topics (R101-115) are used in the experiments.

The performances achieved by the Web information gathering system by applying the three models are compared and analyzed quantitatively. Two schemes are applied in the evaluation: the precision averages at 11 standard recall levels [18] and F_1 Measure [5]. The former is used by TREC and computes each recall-precision point by:

$$\frac{\sum_{i=1}^{N} precision_{\lambda}}{N} \qquad (9)$$

where $\lambda = \{0.0, 0.1, 0.2, \ldots, 1.0\}$ and N denotes the number of topics. Fig. 4 illustrates the recall-precision average results of the three models. The perfect TREC model

[4] http://www.comscore.com/press/release.asp?press=873.

[5] The Queensland University of Technology Library, http://www.library.qut.edu.au/.

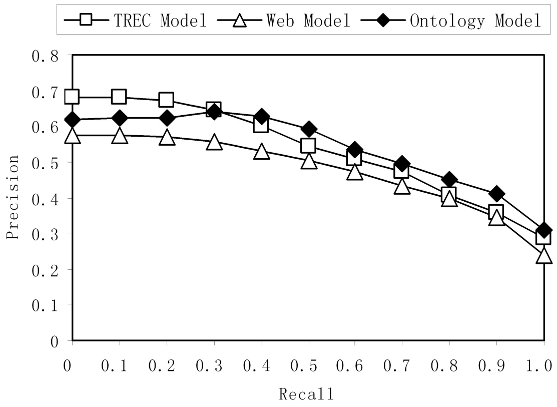


Figure 4. The Recall-Precision Results

	Macro-*F1* Measure			Micro-*F1* Measure		
Topic	TREC	Web	Onto	TREC	Web	Onto
R101	0.7333	0.6555	0.5978	0.6660	0.5982	0.5428
R102	0.7285	0.5588	0.5754	0.6712	0.5179	0.5327
R103	0.3600	0.3347	0.3859	0.3242	0.3059	0.3445
R104	0.6441	0.6162	0.6280	0.5851	0.5662	0.5786
R105	0.5548	0.5662	0.5782	0.5092	0.5163	0.5293
R106	0.2324	0.2433	0.2794	0.2223	0.2270	0.2586
R107	0.2297	0.2028	0.2057	0.2061	0.1866	0.1936
R108	0.1794	0.1520	0.1388	0.1676	0.1424	0.1295
R109	0.4508	0.6564	0.6659	0.4205	0.6026	0.6119
R110	0.2176	0.1560	0.2801	0.2019	0.1466	0.2568
R111	0.1082	0.0905	0.1267	0.1017	0.0863	0.1218
R112	0.1940	0.1745	0.1987	0.1800	0.1631	0.1813
R113	0.3152	0.2126	0.3519	0.2867	0.1975	0.3252
R114	0.4128	0.4247	0.4192	0.3732	0.3892	0.3840
R115	0.5063	0.5395	0.5079	0.4523	0.4831	0.4551
Avg.	0.3911	0.3722	0.3960	0.3579	0.3419	0.3630

Table 1. The Detailed Experiment Results

slightly outperforms others before reaching recall cutoff 0.3, and then the Ontology model becomes the best since that on. This may indicate that the perfect TREC training sets are more precise than others, but does not cover as much relevant semantic space as the Ontology model. As a result, the Ontology model's precision catches it up while the recall value increases. The other evaluation scheme, F_1 Measure, is calculated by:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \qquad (10)$$

Precision and recall are evenly weighted in F_1 Measure. The *macro-*F_1 averages each topic's precision and recall values then calculates the F_1 Measure, whereas the *micro-*F_1 calculates the F_1 Measure for each returned result for a topic and then averages the F_1 values. The greater F_1 values indicate the better performance. The detailed F_1 Measure results are presented in Table. 1. In average, the Ontology model performs best. The highlighted rows are the topics that the Ontology model outperforms the perfect TREC model. This is because the TREC model employs the manpower of linguists to read every single document in the training set, which is perfect but expensive. As a result, the number of documents included in a TREC training set is limited (about 70 documents per topic in average), and some semantic meanings contained by the topic are not fully covered by the TREC training set. In contrast, the knowledge in the Ontology model is extracted from the LCSH and a large volume of expert classified information in library catalogue. The broad semantic coverage is the Ontology model's strength. As a result, the Ontology model has about 1730 documents per topic in average covering much broader semantic extent than the TREC training set. Based on the experiments, the proposed ontology learning and mining model is evaluated and its success is confirmed.

7. Related Work

Much effort has been invested in ontology learning or mining for semantic interpretation. Staab & Studer [14] formally define an ontology as a 4-tuple of a set of concepts, a set of relations, a set of instances and a set of axioms. Slightly different, Maedche [10] has another definition which differentiates the relations to hierarchical and plain relations. They also proposed an ontology learning framework for the Semantic Web. The framework extends typical ontology engineering environments by using semi-automatic ontology construction tools with human intervention, and constructs ontologies adopting the paradigm of balanced cooperative modelling.

Zhong & Hayazaki [21] defined two major stages of ontology development: conceptual relationship analysis and ontology prototype generation. Zhong [20] proposed a learning approach for task (or domain-specific) ontology, which employs various mining techniques and natural-language understanding methods. Li & Zhong [7] proposed an semi-automatic ontology learning method, in which a class is called compound concept assembled by primitive classes that are the smallest concepts and can not be divided any further.

Singh et al. [8] developed *ConceptNet* ontology and tried to specify common sense knowledge. However, *ConceptNet* does not count expert knowledge. Navigli et al. built an ontology called *OntoLearn* [12] to mine the semantic relations among the concepts from Web documents. Gauch et al. [3] used reference ontology built based on the categorization of online portals and proposed to learn personalized ontology for users. Developed by King et al. [4], *IntelliOnto* is built based on DDC system, and attempts to describe the world knowledge. Unfortunately, these works cover only a small number of concepts and do not specify the semantic relationships of "part-of" and "kind-of" existing in the concepts but only "super-class" and "sub-class".

In terms of user profiling, it is common that a user profile is generated by asking user questions explicitly or observing her activity implicitly [11, 17], or by analyzing the user log data [11, 15]. In some recent researches, ontology is used as a basis for the user profile generating, and the user profile is represented by a set of keywords or patterns [3, 6, 7, 17].

8. Conclusions

In this paper, a personalized ontology model is proposed aiming to synthesize world knowledge and expert knowledge for specific topics. The model extracts world knowledge from the LCSH system and discovers expert knowledge from a large volume of specified information in the library catalogue. The proposed model attempts to facilitate the user possessed concept model and to generate the personalized user profile for Web information gathering.

It is a challenge to use semantic relations of "kind-of", "part-of", and "related-to" in a single computational model. During literature review, we did not find any mathematic model that can well formalize these three relations together. In this paper, the proposed ontology model is an attempt to specify these semantic relations in a single framework. A two-dimensional method (Exhaustivity and Specificity) is also presented in the paper to quantitatively analyze these three semantic relations. The proposals are successfully evaluated by comparing knowledge extracted by the personalized ontology model, against knowledge generated manually by linguists. The proposed model is a significant contribution to personalized ontology engineering and to concept-based Web information gathering in Web Intelligence.

Acknowledgements

We would like to thank the *Library of Congress* and *Queensland University of Technology Library* for authorizing the use of MARC and the catalogue records. We would also like to thank Prof. Peter Bruza and Prof. Kerry Raymond for their valuable comments to the research work.

References

[1] G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. The MIT Press, 2004.

[2] K. Curran, C. Murphy, and S. Annesley. Web intelligence in information retrieval. In *Proc. of WI' 03*, pages 409 – 412, 2003.

[3] S. Gauch, J. Chaffee, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3-4):219–234, 2003.

[4] J. D. King, Y. Li, X. Tao, and R. Nayak. Mining World Knowledge for Analysis of Search Engine Content. *Web Intelligence and Agent Systems*, 5(3):233–253, 2007.

[5] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proc. of the 18th intl. ACM SIGIR conf. on Res. and development in inf. retr.*, pages 246–254. ACM Press, 1995.

[6] Y. Li and N. Zhong. Web Mining Model and its Applications for Information Gathering. *Knowledge-Based Systems*, 17:207–217, 2004.

[7] Y. Li and N. Zhong. Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.

[8] H. Liu and P. Singh. ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal*, 22(4):211–226, 2004. Kluwer Academic Publishers.

[9] J. Liu. New Challenges in the World Wide Wisdom Web (W4) Research. *Lecture Notes in Computer Science*, 2871:1–6, Jan 2003.

[10] A. D. Maedche. *Ontology Learning for the Semantic Web*. Kluwer Academic Publisher, 2002.

[11] S. E. Middleton, N. R. Shadbolt, and D. C. D. Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, 2004.

[12] R. Navigli, P. Velardi, and A. Gangemi. Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE*, 18:22–31, 2003.

[13] S. E. Robertson and I. Soboroff. The TREC 2001 filtering track report. In *Text REtrieval Conference*, 2001.

[14] S. Staab and S. R., editors. *Handbook on Ontologies*. Springer-Verlag Berlin Heidelberg, 2004.

[15] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proc. of the 13th intl. conf. on World Wide Web*, pages 675–684, USA, 2004.

[16] X. Tao, Y. Li, N. Zhong, and R. Nayak. Automatic Acquiring Training Sets for Web Information Gathering. In *Proc. of the IEEE/WIC/ACM Intl. Conf. on Web Intelligence*, pages 532–535, HK, China, 2006.

[17] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proc. of RIAO 2004*, pages 380–389, France, 2004.

[18] E. Voorhees. Overview of TREC 2002. In *The Text REtrieval Conference (TREC)*, 2002. Retrieved From: http://trec.nist.gov/pubs/trec11/papers/OVERVIEW.11.pdf and http://trec.nist.gov/pubs/trec11/appendices/MEASURES.pdf.

[19] L. Zadeh. Web intelligence and world knowledge - the concept of Web IQ (WIQ). In *Processing of NAFIPS '04.*, volume 1, pages 1–3, 27-30 June 2004.

[20] N. Zhong. Representation and construction of ontologies for Web intelligence. *International Journal of Foundation of Computer Science*, 13(4):555–570, 2002.

[21] N. Zhong and N. Hayazaki. Roles of ontologies for web intelligence. In *Proceedings of Foundations of Intelligent Systems : 13th International Symposium, ISMIS 2002,*, volume 2366, page 55, Lyon, France, June 27-29 2002.