An Ontology-based Mining Approach for User Search Intent Discovery

*Yan Shen, Yuefeng Li, Yue Xu, *Renato Iannella, Abdulmohsen Algarni*

Computer Science Discipline, Faculty of Science and Technology
Queensland University of Technology, Australia

*Semantic Identity

{*y12.shen,y2.li,yue.xu,r.iannella,a1.algarni*}*@qut.edu.au*

Xiaohui Tao

Department of Mathematics and Computing, Faculty of Sciences
The University of Southern Queensland, Australia

xtao@usq.edu.au

Abstract *Discovering proper search intents is a vital process to return desired results. It is constantly a hot research topic regarding information retrieval in recent years. Existing methods are mainly limited by utilizing context-based mining, query expansion, and user profiling techniques, which are still suffering from the issue of ambiguity in search queries. In this paper, we introduce a novel ontology-based approach in terms of a world knowledge base in order to construct personalized ontologies for identifying adequate concept levels for matching user search intents. An iterative mining algorithm is designed for evaluating potential intents level by level until meeting the best result. The propose-to-attempt approach is evaluated in a large volume RCV1 data set, and experimental results indicate a distinct improvement on top precision after compared with baseline models.*

Keywords Ontology mining, Search intent, LCSH, World knowledge

1 Introduction

For an effective search engine, retrieving desired information is one primary objective that motivates many researchers to invest more than several decades. To improve the existing search capabilities, a series of advanced algorithms and processes along with the solid experimental supports have been developed [3]. However, due to the increasing complexity of the internet, recent search engines suffer from an emerging issue - ambiguity [8] [10]. It is caused by the fact that the majority of present search techniques are highly sensitive on vocabulary [1] and lack of personalization [2]. As a result, many personalized methods by considering user profiles are studied to alleviate this problem. However, these methods are either expensive in extraction or inaccurate in description. In order to avoid the discussed drawbacks and enrich the inference capacity of Web personalization, more and more people [19] [11] have taken advantage of using ontologies. Note that ontologies play an important role on machine learning and Information Retrieval (IR) [4]. With an integration of specified concepts and semantic relations, many retrieval systems can perform higher accuracy and automated characteristics. The ontologies classify all the knowledge into a well-structured way, which facilitate users to assess information items.

The paper aims to take advantage of ontologies to obtain accurate user search intent. Search intent is a significant object that contains what user needs. It can be studied into two means: the specificity and exhaustivity intent. Specificity describes the focusing extent of a topic, i.e., user's interests have a narrow and focusing goal, whereas exhaustivity describes a different extent of a topic, i.e., general/wider scope of user's interests. However, in recent years, a hard question is how to discover and characterize user intent. One existing method is quantification, i.e. using relevance weight of a pattern [20], and then re-ranking ,which is a document-based technique.

Here, we propose a hierarchical concept level-finding method, which is a knowledge-based technique as an alternative solution. A novel ontology-based approach is introduced to discover user search intent. Library of Congress Subject Headings (LCSH), which has a widespread coverage in various knowledge domains, is applied as a world knowledge base for learning personalized ontologies. According to these

Proceedings of the 16th Australasian Document Computing Symposium, Canberra, Australia, 2 December 2011.

comprehensive ontologies, diverse information is allocated in a number of concept levels. These levels might be linked if relationships exist. We assume that an actual user search intent can be found from one of the levels. The higher a level, the broader extent it has. Conversely, extents are more specific while towards lower levels. The idea is similar to a zooming navigation. To define a certain search intent, an iterative mining algorithm is developed. The attempt-to-propose approach is evaluated experimentally by 100 topics in Reuters Corpus Volume 1 (RCV1). The results indicate the top precision performance is improved remarkably. This paper will help to design a search strategy to avoid the ambiguous troubles caused by typical search techniques and owns the potential value to construct an innovative zooming navigation.

The rest of the paper is organized as follows: Section 2, some significant related work is discussed; Section 3, a world knowledge ontology and its definitions are presented; Section 4, primary components of the proposed approach are fully described; Section 5, a number of scientific experiments are conduced, and the major experimental results are outlined and discussed for evaluation; Section 6, a conclusion is brought to summarize this work and specify potential work in future.

2 Related Work

This section is classified into two categories, one is about user information needs, the other is focusing on ontology-based techniques.

User Information Needs

For user information need acquisition, many efforts have been involved to improve the accuracy and effectiveness. Closely related to our work, user ontology consisting of both concepts and semantic relations is presented by Jiang and Tan [5]. Their goal is to represent and capture users' interests in target domain. Subsequently, a method, they called Spreading Activation Theory (SAT), is employed for providing personalized services. Li and Zhong [9] develop a term-based ontology leaning method for acquiring user information needs. More recently, Tao et al. [14] propose an ontology-based knowledge retrieval framework to capture user information needs by considering user knowledge background and user's local instance repository (user profile) with association roles and data mining techniques. Other work also realizes the importance of user information need, they treat user interest as implicit feedback and store in user profile. Trajkova and Gauch [16], and Liu et al. [10] learn a user's profile from her/his browsing history, whereas Sieg et al. [13] utilize ontological user profile on the basis of the user's interaction with a concept hierarchy which captures the domain knowledge, and Tao et al. [15] [14] require the user to specify a profile manually. In short, these work aim to enhance search performance through asking users explicit feedback, such as preferences, or collected implicit feedback, which are normally either expensive in extraction or inaccurate in description.

Ontology-Based Techniques

Ontology is a collection of concepts and their interrelationships, which provide an abstract view of an application domain. It is an explicit specification of a conceptualization. Over the recent years, people who are mentioned below have often held the hypothesis that ontology-based approaches should perform better than traditional ones on IR, since ontologies are more discriminative and arguably carry more "semantics". As a result, many research concentrate on how to use ontology techniques. Zhong [19] proposes a learning approach for task (or domain-specific) ontology, which employs various mining techniques and natural-language understanding methods. Li and Zhong [9] present an automatic ontology learning method, in which a class is called a compound concept, assembled by primitive classes that are the smallest concepts and cannot be divided any further. Liu and Singh [11] develop ConceptNet ontology and attempt to specify common sense knowledge. However, ConceptNet does not count expert knowledge. Navigli et al. [12] build an ontology called OntoLearn to mine the semantic relations among the concepts from Web documents. Gauch et al. [4] use ontology references based on the categorisation of online portals and propose to learn personalised ontology for users. Developed by King et al. [6], IntelliOnto is built based on the DDC (Dewey Decimal Classification) system and attempt to describe the background knowledge.

Unfortunately, the previous work on ontology learning covers only a small size of concepts, where mainly uses "Is-A" (super-class, or sub-class) relation in the knowledge backbone. They don't consider to mine and characterize knowledge in a concept level rather than domains. To extend these methods, the backbone of personalized ontologies is been determined to build a real hierarchical structure by applying information in a world knowledge repository.

3 LCSH: World Knowledge Base

World knowledge is the common-sense knowledge acquired by people based on their experience and education. It can be considered as an exhaustive repository to maintain the known knowledge by human being [18]. LCSH is an ideal world knowledge representation because of a rich vocabulary is used to cover all subject areas. Meanwhile, wealthy semantic relations among terms are good at reveal precise relationships of subjects. In the LCSH, subject headings are basic semantic units for conveying domain knowledge and concepts, they have three main types of references: *Broader Term* (BT), *Narrower Term* (NT) and *Related Term* (RT). BT means a hypernym, is a more general term, e.g. "Pressure" is a generalization of "Blood Pressure"; NT means a hyponym, is a more specific term, e.g. "Economic Crisis" is a specialization of "Crisis". These two references are used in our model to indicate the $is-a$ relations among subjects in the world knowledge base. To facilitate in-levels ontology construction later in this paper. The references are firstly redefined to $ancestor$ and $descendant$ lexical relations in our approach respectively. $ancestor$ refers to the concept of BT, and $descendant$ refers to NT, more information can be found in Table 1. All the subjects are formalized as:

Type	Paraphrase	Example
Ancestor	is the general term for	"profession is the general term for scientist" $\Longrightarrow Ancestor(profession, scientist)$
Descendant	is-a	"scientist is a profession" $\Longrightarrow Descendant(scientist, profession)$

Table 1: Examples for redefined relations

Definition 1 (Subjects): Let $\mathbb{S}$ denote a set of subject headings in LCSH, a subject $s \in \mathbb{S}$ is formalized a triple $(label, ancestor, descendant)$, where

- $label$ is the heading of s in LCSH thesaurus;
- $accestor$ is a function regarding the subjects that are more general and located a higher level than s in the world knowledge base;
- $descendant$ is a function regarding the subjects that are more specific and located a lower level than s in the world knowledge base.

At this stage, there is only one relation $r = (is - a)$ considered by our approach. Thus, the world knowledge base can be formalized as:

Definition 2 (World Knowledge Base): A world knowledge base ontology is a directed acyclic graph structure defined as a pair $\Theta := (\mathbb{S}, r)$, consisting of

- $\mathbb{S}$ is a set of subjects in LCSH $\mathbb{S} = \{s_1, s_2, ..., s_n\}$;
- r is the semantic relation $r = (is - a)$ existing among the subjects in $\mathbb{S}$;

4 Proposed Approach

A quick overview of the proposed approach is illustrated in Figure 1. The paper first holds a hypothesis that a user search intent should exist somewhere in an ontology. It is treated as a user information need and represented by a range of concept extent. The intent could be general or specific. In order to minimum user burdens, a query is the only input for the proposed approach, it likes a real search activity. To cope with a user query, a subject-based search model is developed in order to retrieve matching results from a LCSH database. The function is similar as a keyword-based search except the type of returned results is a list of subject headings. Both "AND" and "OR" operators are employed at the same time. This process might increase information redundancy, however, it can extend a scope to cover potential user intents with restricted user information. Semantic relation extractions are conducted for all the matching subjects for learning personalized ontologies. After that, all terms appearing in the subjects are used to do a query expansion, and then find semantically similar matches rather than lexically dissimilar by taking the extracted relations into account. The related methods of learning personalized ontologies and in-levels ontology mining are mainly explained in the next two subsections.

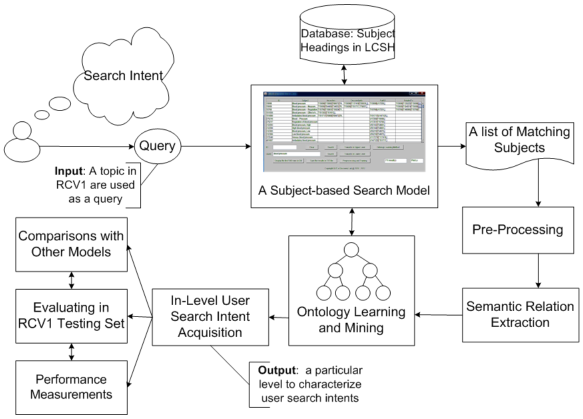
Database: Subject Headings in LCSH
Search Intent
Query
Input: A topic in RCV1 are used as a query
A Subject-based Search Model
A list of Matching Subjects
Pre-Processing
Semantic Relation Extraction
Ontology Learning and Mining
In-Level User Search Intent Acquisition
Output: a particular level to characterize user search intents
Comparisons with Other Models
Evaluating in RCV1 Testing Set
Performance Measurements

Figure 1: The general architecture design

4.1 Personalized Ontology Learning

Concept hierarchy is an essential subtask of ontology learning. In theory, it is a prerequisite hierarchy where a mount of nodes represent concepts in a domain, and related links are served as prerequisite relationships. For this paper, we create a special hierarchy format to satisfy our research purposes. The hierarchical backbone is drawn as Figure 2. One of the objectives is to make use of this hierarchy to allocate information into a well structure, which facilitate users to access information items. Another objective is to infer implicit knowledge by tracking internal relationships among subjects. The gathered implicit knowledge will be used to estimate whether a user search intent is characterized in a certain level.

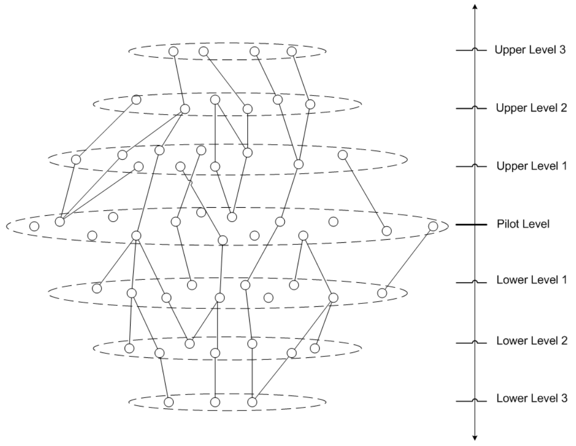

Figure 2: The backbone of in-levels hierarchy

After implementing the subject-based search model, a list of matching subjects can be obtained. Each of them is realized as a concept here and represented by a single node in Figure 2. To learn a personalized ontology, a concept's domain is confirmed by taking these subjects into account because all the subjects have their domains, which are considered as correlative information regarding a topic. While the subjects are all in the same abstract level, where is originally defined as "Pilot Level". Initially, a formalized definition for its domain is provided as:

Definition 3 (Domain for a Level): Let C_i denote a set of subjects $C_i := \{s_1, s_2, ..., s_h\}$ in a level l_i. We define its domain by $dom(l)_i := dom(s_1) \cup dom(s_2) \cup ... \cup dom(s_h)$, where $dom(s_h)$ contains all the terms in the label of subject s_h.

Dashed circulars in Figure 2 are utilized to indicate the domains of different levels. With respect to the ontology learning, we also formalize:

Definition 4 (Personalized Ontology): the personalized ontology for a topic in a 4-tuple $\Theta^p := (\mathbb{C}, L, DOM, e)$, where

- $\mathbb{C}$ is a super set of C including all subjects in levels $\mathbb{C} = \{s_1, s_2, ...s_h\}$;
- L is a set of levels consisting of a domain and subjects $L = \{l_1, l_2, ..., l_i\}$;
- $DOM := (dom(l_1), dom(l_2), ..., dom(l_e))$;
- e is the number of levels.

The semantic relations specified in the world knowledge base benefit our approach to acquire a set of new subjects in other levels. All the levels are classified into two directions, one is upper, the other is lower. The pilot level is selected as a benchmark in the hierarchy. Some subjects might just occur one time in the pilot because of no semantic relations. The major in-levels concept is similar to a knowledge generalization process. Indeed, a subject in a upper level covers a more general knowledge than the lower one. In other words, the knowledge in a level can be summarized by the knowledge in the next upper level. Eventually, all the knowledge in the world knowledge base will be summarized in philosophy. This is a main reason why the domains of upper levels are getting smaller when moving towards to the peak of the backbone in Figure 2, where looks like a shape of cone. However, why this happens as the same in lower levels? From a perspective of IR, the subject-based search model uses to return specific subjects based on keywords. The majority of matching subjects are usually lack of semantic relations in the pilot level to extend more knowledge. As a result, the number of subjects in lower levels are decreasing as well as their domains. Therefore, the shape becomes a inverse cone. Note that the backbone structure is not a formal tree, a node can has more than one parent or child.

4.2 In-Levels Ontology Mining Method

To prove the hypothesis mentioned earlier, an iterative ontology mining method is proposed in this section. It starts from the pilot level, and then builds a personalized ontology (the backbone of in-levels hierarchy) in order to find a suitable level for a search intent. The building process simply employs the $is-a$ relation to find all parents in an upper level or get all children from a lower level. For understanding precisely, an entire study is separated into two phases to explain the method in details. Each phrase involves several steps.

Phase 1: Represent feature in levels
There are two main objectives: 1) to decide subjects and their weights for the pilot level l_ρ; and 2) to represent the pilot level l_ρ as a feature vector F_ρ. Firstly, retrieve a number of matching subjects from the pilot level l_ρ after implementing the subject-based search model. Then, calculate a weight for all subjects $s \in C_\rho$ by using the following equation:

$$w(s) = \frac{|q \cap s|}{|s|} \qquad (1)$$

where $|q \cap s|$ denotes the number of terms appeared in both query q and subject s, $|s|$ denotes the total number of terms in subjects. Therefore, a set of subject weight pairs are obtained as $S(w) = \{< s_1, w_1 >, < s_2, w_2 >, ..., < s_n, w_n >\}$.

Secondly, expand the query to a set of terms by union all terms from the submitted query and matching subjects,and let $Q_\rho = \{t_1, t_2, ..., t_m\}$. For example, the submitted query is $query = \{t_1, t_2\}$, other subjects are $s_1 = \{t_1, t_2\}$, $s_2 = \{t_1, t_2, t_6\}$, and $s_3 =$

$\{t_1, t_2, t_5, t_8\}$. After that, $Q_\rho = query \cup s_1 \cup s_2 \cup s_3 = \{t_1, t_2, t_5, t_6, t_8\}$.

Thirdly, Calculate weights for all terms $t \in Q_\rho$ via using the following equation:

$$weight(t) = \sum_{t \in s, s \in C_\rho} \frac{w(s)}{|s|} \qquad (2)$$

Then, we receive a set of term weight pairs as a feature vector $F_\rho = \{< t_1, w_1 >, < t_2, w_2 >, ..., < t_m, w_m > \}$ to represent this level, where $w_m = weight(t_m)$.
Phase 2: Determine the best level for user search intents
The goal is to determine the appropriate level for characterizing the user search intent according to a training set. Let D_t stand for a set of documents in the training set, which has $D_t = {D_t}^+ \cup {D_t}^-$. ${D_t}^+$ is a set of positive documents, and ${D_t}^-$ stands for negative ones, where t denotes a certain topic. All these documents have been initialized a value of either 0 or 1 by linguists. We calculate a weight for each document in the training set by using the feature F_ρ. Thus, rank D_t by using Ranking Algorithm provided as follows:

$$rank(d) = \sum_{t \in Q_\rho}^{n} weight(t) \qquad (3)$$

Based on the ranked documents, a top-K precision $precision(l_\rho)$ can be calculated for the pilot level l_ρ by apply the equation below:

$$precision(l_\rho) = \frac{\sum_{i=1}^{K} f(d_i)}{K} \qquad (4)$$

where $f(d_i) = 1$ if d_i is relevant, otherwise $f(d_i) = 0$.

Algorithm 1 discovering search intent (e.g.upper levels only)

```
Input:
    $D^t$ in the training set of RCV1; $F_\rho$; Parameter $\mu$.
Output:
    A suitable level to match a user intent
 1: Let $j = \rho, i = j$;
 2: Let $i = i + 1$; //$\rho = \rho + 1$, shift to the upper level;
 3: Get $Q_i$ and $F_i$;                //refer to Phase 1;
 4: Use $F_i$ to rank $D_t$;            //see Eq.(3)
 5: Get $precision(l_i)$;               //see Eq.(4)
 6: if $precision(l_i) < precision(l_j)$ then
 7:     return $l_j$;
 8: else
 9:     if $i - j > \mu$ then
10:         return $l_i$
11:     end if
12: end if
13: $j = i$;
14: Go to 2;
```

The pilot level is possibly not the expected level resulting in shift to the upper level $l_{\rho+1}$ or lower levels in the hierarchy. Thereby, a new set of subjects in $l_{\rho+1}$ are returned by getting all subjects s that have a $is - a$ relationship with any subjects in the pilot level l_ρ. Repeat the above steps to rank the documents D_t by using the feature vector $F_{\rho+1}$ for level $l_{\rho+1}$, and then calculate the top-K precision $precision(l_{\rho+1})$. If $precision(l_\rho) > precision(l_{\rho+1})$, return l_ρ as the appropriate one as user search intent. Otherwise, go to step two and implement the same procedure in $l_{\rho+2}$ again. Algorithm 1 describes the idea that is looping for upper levels until meeting the most satisfactory level based on precision performance, where parameter μ is used to control the distance between the selected level l_i with the pilot level l_ρ. If a level is too far away with the pilot level, we assume that it is less significant to search intents. To save space, the paper omits the explanation for lower levels because its algorithm is quite similar as Algorithm 1. According to two phases above, we are able to gain a level with the best top-K precision among all the hierarchical levels. This level is considered as the output of user search intents from our proposed approach.

According to two phases above, we enable to gain a level with the best top-K precision among all the hierarchical levels. This level is considered as the output to match a user search intent.

5 Evaluation

In this session, it first states the data collections used for the subject-based search model and our experiments. Thus, a description is provided to explain our experimental design. Evaluated results are also presented after examining by diverse performance measurements in the concept hierarchy.

5.1 Data Collections

LCSH was chosen as the database for the subject-based search model development. The size is approximately 719 mega bytes stored in Microsoft Office Access. Initially, 20 tables were created to save different data. The data of topical, corporate, and geographic subjects (491,250 subjects in total) were extracted for building the ontology of world knowledge base. Meanwhile, there are five different references linking all these subjects in LCSH. The paper only adopted the references of BT and NT, and encoded as a semantic relation of $is - a$.

As a well-known evaluation methodology founded by IR research community, the Text Retrieval Conference (TREC)[1] Filtering Track is widely applied to evaluate the effectiveness of search applications. RCV1 was applied in our experiments because it's a crucial component of TREC-11 2002 Filtering Track, which

[1] http://trec.nist.gov/

		$pr@20$		$MeanAve.Pre.$		$F_1 - Measure$	
	# Subjects	Value	% Improve	Value	% Improve	Value	% Improve
Upper Lv.7	**25.96**	**0.204**	**21.19**	**0.228**	**0.07**	**0.281**	**-1.025**
Upper Lv.6	37.76	0.199	18.42	0.224	-1.43	0.279	-1.66
Upper Lv.5	54.04	0.193	14.39	0.225	-1.01	0.281	-1.02
Upper Lv.4	75.96	0.18	6.49	0.221	-2.69	0.278	-2.1
Upper Lv.3	114.16	0.183	8.53	0.223	-1.9	0.28	-1.35
Upper Lv.2	178.8	0.188	11.79	0.229	0.55	0.284	0.08
Upper Lv.1	365.16	0.18	7.03	0.231	1.5	0.287	1.15
Pilot Lv.	**2132.04**	**0.168**		**0.228**		**0.284**	
Lower Lv.1	370.04	0.17	1.19	0.228	0.39	0.284	0.21
Lower Lv.2	103.84	0.19	11.31	0.23	0.88	0.285	0.5
Lower Lv.3	32.52	0.174	3.54	0.222	-2.41	0.2798	-1.63

Table 2: First 50 topics performance measurement results

is a corpus that contains totally 806,791 documents. These documents were produced by Reuter's journalists between August 20, 1996 and August 19, 1997. All of them were marked in Extensible Markup Language (XML). They distributed into training and testing sets. Before adopting these XML files for our experiments, they have been tokenized into plain texts. Entirely, RCV1 covers 100 topics by two types: 1) the first 50 topics have been developed by assessors from the National Institute of Standards and Technology, and the assessors have made the relevant judgements manually; 2) the last 50 topics have been built automatically by machine learning instead of by human being for intersection topics. To minimize bias in experiments, the paper conducted a pre-processing for all the queries, subjects in LCSH, and XML files in RCV1 corpus. The pre-processing includes stop-words removal and stemming by applying Porter Stemmer algorithm. The development of the subject-based search model and our experiments were encoded by JAVA.

5.2 Measures & Baseline Model

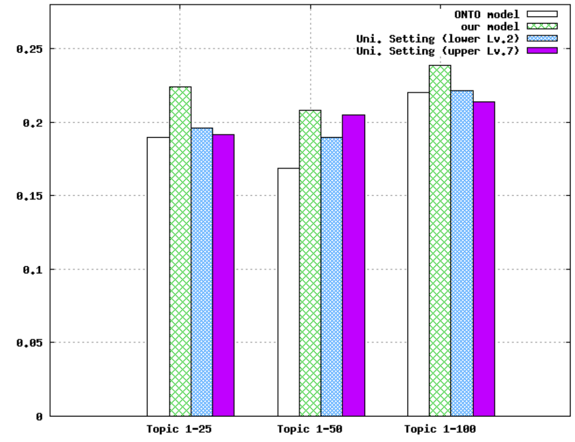

Figure 3: Top 20 precision results

In order to prove the accuracy and feasibility of our approach, the paper estimated all the levels by applying four state-of-the art measuring methods. There are top 20 precision based on the relevance judgement in RCV1 ($pr@20$), the precision averages at 11 standard recall levels ($11-points$), the *Mean Average Precision* (MAP), and the $F_1 - Measure$.

Top 20 precision is considered as the most important standard in the evaluation since a web searcher is mostly going to look at the top 20 documents [7]. In the domain or IR, precision is the percentage of retrieved documents that are relevant. Each document in RCV1 has already been judged the relevance by 0 and 1. Compared to these judgements, the top 20 precision can be computed.

$$pr@20 = \frac{|\{first20ranked\ docs\} \cap \{relevant\ docs\}|}{20}$$

MAP is correlated with Average Precision ($Ave(p)$). $Ave(p)$ is the average of precision at each relevant document retrieved in the ranked sequence. Consisting of the $Ave(p)$, the equation of MAP is formed as:

$$MAP = \frac{1}{|Q|} \sum_{s=1}^{|Q|} \text{Ave}\,(\text{p})$$

where Q stands for the number of queries. *F1-measure* was first introduced by C. J. van Rijsbergen [17]. It combines recall and precision with an equal weight in the following form:

$$F_1 - Measure = \frac{2 \times precision \times recall}{(precision + recall)}$$

$11-points$ measure is also used to estimate the performance of retrieval models by averaging precisions at 11 standard recall levels (i.e. $recall = 0, 0.1, 0.2, ..., 1$).

A ontology model named ONTO model from Tao et al in 2010 [14] was selected as one baseline for evaluating. The ONTO model has already evaluated its outperformed capability after comparing with a number of other models. Hence, the related comparisons are meaningful. Furthermore, two uniform level settings(parameter $\mu = 2\, and\, 7$) were selected as baseline models, which are upper level 7 and lower level 2 respectively.

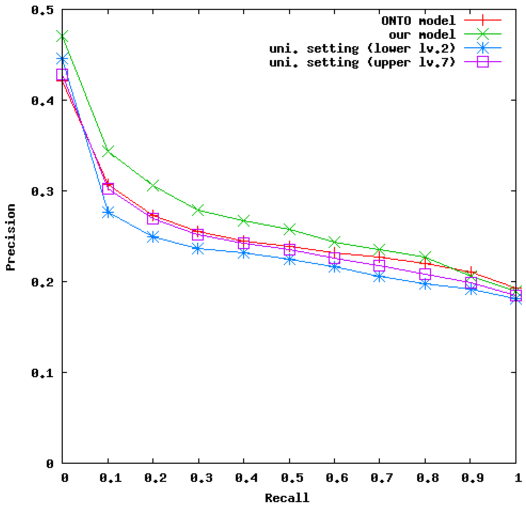

Figure 4: First 1-25 11-points performance

5.3 Results and Discussion

The performance of the experimental models was measured by the mentioned measurements for different levels. Table 2 includes all statistical results computed after implementing the first 50 topics in RCV1. To indicate the influences affected, $the\ percentage\ change\ in\ performance$ was used to compute the difference in Top 20 precision, MAP, and $F_1 - Measure$ results among the levels. It is formulated as:

$$\%Improve = \frac{Result_{level} - Result_{pilot}}{Result_{pilot}} \times 100$$

The lager $\%Improve$ value the more significant improvement achieved. We noticed that the number of subjects are decreasing while towards upper or lower levels. This is an evidence to prove that the proposed hierarchy is reasonable. Refer to subsection 4.1, we can picture a shape for upper levels as a cone. In contract, an inverse cone is for lower levels.

As shown in Table 2, the results of MAP and $F_1 - Measure$ are not improving dramatically. The main reason is that they both considered about the same recall computing over the RCV1 data sets. However, if only considering about precision, the upper level 7 has the best result on pr@20, which is 21.19% better than the baseline model. As a result, for the first 50 topics, the upper level 7 was determined as the best level to characterize user search intents. According to Figure 3, the pr@20 results are always leading the others. The results from the uniform settings are also superior to the ONTO model on first 25 and 50 topics. The results of $11 - points$ on first 25 and 50 topics are illustrated in Figure 4 and 5.

Limitations: Three main limitations exist in this work. The first one is: our investigation is mainly focusing on the usage of $is - a$ relations in LCSH. The other relations, including $used - for$ and $related - to$ are regardless within our approach. As a result, the maximum number of depth detected based on the constructed concept hierarchy is 28 but not 37 as specified in the LCSH specification [14]. Some of useful implicit knowledge might be not entirely discovered from world knowledge representation. The second limitation is caused naturally from the LCSH. In reality, user interests are usually changing all the time. The choice and form of headings are not necessarily current because the LCSH terms have evolved over time, but they can never be totally up to date. This might lead to misinterpretation of user search intents. The last one is about the dataset that applied for evaluation. It is a textual collection of news, but database used for searching is a subject collection of library headings. This might possibly influence experimental results.

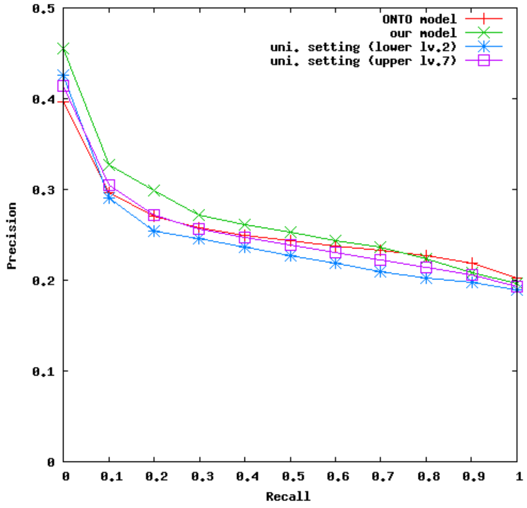

Figure 5: First 1-50 11-points performance

In sum, the proposed ontology-based approach is been proven successfully based on the experiments. The improvements are consistent and significant on the top 20 precision measure. The related results indicate the overall performance are better than the baseline model.

6 Conclusions

A novel ontology-based approach is introduced for user search intents discovery. It utilizes a subject-based search model to filer out irrelevant information, and then allocates matching results into a world knowledge base - LCSH. A concept-based hierarchy is built by applying semantic relations from the world knowledge in order to characterize accurate user intents in an actual level. A huge test bed was utilized for a number of experiments. The experimental results demonstrate that our proposed approach is working and promising. It can enhance the search effectiveness in top precision. The major contribution of this paper is describing an alternative ontology-based method to discover user search intents except pattern mining. This research will significantly influence the development of personalized Web search services, and the related deliverables have potentials to contribute intelligent search navigation.

In future, we plan to investigate how to use the rest semantic relations in LCSH, including *equivalence* ($used - for$) and *associative* ($related - to$). These semantic relations can not only assist to revise subjects appeared in a level, but also navigate to user search intents more precisely. The implicit knowledge including search intents can be obtained effectively from a world knowledge base. Recent studies report that patten mining methods are effective strategies for relevance information acquisition in a short number of terms rather than content-based methods [8]. Since now our approach achieved the task to discover user search intents in a contain level. Another attempt is to discover a certain subject in the level.

References

[1] T. Berners-Lee, J. Hendler, O. Lassila et al. The semantic web. *Scientific american*, Volume 284, Number 5, pages 28–37, 2001.

[2] P.A. Chirita, C.S. Firan and W. Nejdl. Personalized query expansion for the web. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 7–14. ACM, 2007.

[3] Z. Dou, R. Song, J.R. Wen and X. Yuan. Evaluating the Effectiveness of Personalized Web Search. *IEEE Transactions on Knowledge and Data Engineering*, pages 1178–1190, 2008.

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

[5] X. Jiang and A.H. Tan. Learning and inferencing in user ontology for personalized Semantic Web search. *Information Sciences*, Volume 179, Number 16, pages 2794–2808, 2009.

[6] J.D. King, Y. Li, X. Tao and R. Nayak. Mining world knowledge for analysis of search engine content. *Web Intelligence and Agent Systems*, Volume 5, Number 3, pages 233–253, 2007.

[7] H.V. Leighton and J. Srivastava. First 20 precision among world wide web search services(search engines). *Journal of the American Society for Information Science*, Volume 50, Number 10, pages 870–881, 1999.

[8] Y. Li, A. Algarni and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 753–762. ACM, 2010.

[9] Y. Li and N. Zhong. Mining ontology for automatically acquiring web user information needs. *IEEE Transactions on Knowledge and Data Engineering*, pages 554–568, 2006.

[10] F. Liu, C. Yu and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on knowledge and data engineering*, Volume 16, Number 1, pages 28–40, 2004.

[11] H. Liu and P. Singh. Commonsense reasoning in and over natural language. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 293–306. Springer, 2004.

[12] R. Navigli, P. Velardi and A. Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, Volume 18, Number 1, pages 22–31, 2003.

[13] A. Sieg, B. Mobasher and R. Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525–534. ACM, 2007.

[14] X. Tao, Y. Li and N. Zhong. A personalized ontology model for web information gathering. *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[15] X. Tao, Y. Li, N. Zhong and R. Nayak. Ontology mining for personalizedweb information gathering. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 351–358. IEEE Computer Society, 2007.

[16] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proceedings of RIAO*, Volume 4, pages 380–389. Citeseer, 2004.

[17] CJ Van Rijsbergen. Information retrieval, chapter 7. *Butterworths, London*, Volume 2, pages 111–143, 1979.

[18] L.A. Zadeh. Web intelligence and world knowledge-the concept of web iq (wiq). In *Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the*, Volume 1, pages 1–3. IEEE.

[19] N. Zhong. Representation and construction of ontologies for Web intelligence. *International Journal of Foundations of Computer Science*, Volume 13, Number 4, pages 555–570, 2002.

[20] X. Zhou, S.T. Wu, Y. Li, Y. Xu, R.Y.K. Lau and P.D. Bruza. Utilizing search intent in topic ontology-based user profile for web mining. In *IEEE/WIC/ACM International Conference on Web Intelligence, 2006. WI 2006*, pages 558–564, 2006.