Exploiting Scholar’s Background Knowledge to Improve Recommender System for Digital Libraries

Bahram Amini, Roliana Ibrahim, Mohd Shahizan Othman

Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia (UTM), Malaysia
avbahram2@live.utm.my, roliana@utm.my, shahizan@utm.my

Abstract

Recommender systems for digital libraries have received increasing attention since they assist scholars to find the most suitable articles for research purposes. Many research studies have recently conducted to model the user interests in order to suggest scientific articles based on the scholar’s preferences. However, a major problem of such systems is that they do not undertake user’s background knowledge into the recommendation process and scholars typically have to sift manually irrelevant articles retrieved from digital libraries. Therefore, a challenging task is how to collect and exploit sufficient scholar’s academic knowledge into the personalization process in order to improve the recommendation accuracy. To address this problem, a recommender framework that consolidates scholar’s background knowledge based on the ontological modeling is proposed. The framework exploits Wikipedia as a lexicographic database for concept disambiguation and semantic concept mapping. The practical evaluation by a group of scholars over CiteSeerX digital library indicates an improvement in accuracy of recommendation task.

Keywords: Recommender System, User Profile, Background Knowledge, Ontology, Digital Library

1. Introduction

The rapidly growing amount of data in digital libraries is at explosion rate, providing scholars a good source of data for research tasks. Unfortunately, traditional search engines return a lot of relevant and irrelevant information in response of typical queries that overwhelms the user and produces “information overload” [1]. In fact, non special digital libraries overwhelm scholars by numerous articles that usually do not fit their actual needs. To address this problem, personalization approaches have recently emerged to recommend users more relevant results than the traditional search engines by filtering irrelevant or less relevant results and suggesting scientific articles as close as possible to the scholar’s prior knowledge.

By definition, personalization for scholars is the process of tailoring the outputs of digital libraries to the individual scholar’s informative needs [2]. Misunderstanding the user intention is the source of many recommender system failures. Although educational based recommender systems are similar to electronic news recommender, which assist users to find the most relevant newsletter from a large news collection, there are some striking features that make it impossible to apply directly existing solutions for those applications to digital libraries. In fact, recommendation for scholars should be guided not only by the scholar’s preferences but also by the scholar’s characteristics such as learning goal and background knowledge [3].

In scholar domain, background knowledge circumscribes the user expectance and helps recommender to understand the user intention. Misunderstanding of user intension often arises from omitting the user’s background knowledge in a topic of interest. Therefore, one promising step to improve recommender system for scholars is to incorporate their prior knowledge into the recommendation process [4]. In fact, the user’s background knowledge identifies the user information scope and enables the recommender system to recommend the most relevant articles within the area of scholar’s knowledge [5]. Once the user’s prior knowledge is specified, the recommender system will be able to prune the search results retrieved from a particular digital library, and re-rank the results to match their knowledge level accordingly.

This paper which extends our previous work [6] aims to investigate the impact of user’s background knowledge in scholar domain and proposes a framework to interlock scholar’s background knowledge into the recommendation process for digital libraries. Here, the terms user and researcher as well as
scholar are used interchangeably which all refer to an expert user who interacts with digital libraries. Also, the terms “background knowledge” and “prior knowledge” equally refer to the same meaning. The rest of the paper is organized as follows: Section 2 describes the basics of background knowledge for scholars and reviews the similar works. Section 3 explains the methodology for modeling background knowledge into recommender system. Section 4 describes a framework for incorporating background knowledge in the recommendation process. Section 5 discusses the framework implementation and practical results. Finally, Section 6 explains the research conclusion and further improvement in detail.

2. Backgrounds

2.1. Backgrounds Knowledge

This section discusses the concept of “background knowledge” in the scholar domain. McDonald et al. [7] define background knowledge as what the individual already knows about a subject or topic. Biemans et al. [8] also define prior knowledge as the whole knowledge which learners have known when starting a new learning subject that is the primary knowledge for acquiring new knowledge. Consequently, prior knowledge helps researchers to understand “new” knowledge by bridging known concepts to the new learning concepts in a given subject. Here, “concepts” form building blocks of “prior knowledge”.

In educational settings including scholar domain, prior knowledge has significant influence on researcher performance because perception of new topics can be easily achieved based on the previously known concepts [3]. For example, if a learner has some knowledge about the Classification concept of Artificial Intelligence, he will better understand the book’s section focusing on Text Mining. Thus, the recommender system will better understand the meaning of scholar’s queries which contains Text Mining if it knows about the researcher’s prior knowledge and the query terms containing Classification word. As a result, the recommender system can better filter out the irrelevant result of search by making a correspondence between the known and new topics.

2.2. Related Work

Kayed et al. [9] propose a web site ranking algorithm using the domain ontology. They show how to measure the closeness (relatedness) of retrieved web sites to the user queries, and use the electronic commerce ontology to re-rank them accordingly. This approach employs “concept ontology frequency” as a basis of calculating the relevancy of retrieved web pages. To construct the ontology, they use 10 popular definitions for e-Commerce domain and then extract 34 most common concepts from them, which cover about %90 of the domain concepts. The documents (web page content) are extracted by a general search engine such as Google and assigned a “rank” to each document based on the “distance” from the ontology concepts. The algorithm then computes the occurrences of words in the ontology, and afterward, re-ranks the associated documents according to their occurrence frequencies.

The approach presented in [10] implicitly builds ontology-based user profile based on the user’s information content from the internet including user’s blogs, publications, home pages, etc. An initial profile from the Open Directory Project (ODP)\(^1\) is constructed in a hierarchical structure and is further learned by incorporating additional user details collected from the user’s documents. It also applies WordNet as a lexicon syntactic pattern for hyponyms to augment the salient feature of documents measured by the tf*idf weighting scheme for improving the profile. The ontology-based user profile is further improved in a collaborative manner by learning relevant knowledge from similar profiles in the community.

The study in [11] discusses a personal ontology recommender system for Chinese digital libraries. It uses traditional cataloging scheme as the reference ontology for classification. The reference ontology is extracted from the “borrowing records” of the individual user as well as “notes keyed” by librarian which act as a personal user profile. The personal ontology, which represents the user interests, is engaged to filter out irrelevant or less relevant books based on the keyword matching method. It

\(^1\) http://www.dmoz.org
assumes that the collection of keywords of loaned books represents the user interests. When users log into the system, it extracts favorite topics of individual users from different sources and assemble them to build up the user profile. The system is linked with library’s check in/out information system and collects online information from the user interactions. Thus, the cataloging information and the loan information are both used as knowledge sources for construction user’s ontology.

The work in [12] proposes an ontology-based information integration and recommendation for the scholar’s domain. The system exploits ontology to enrich and disambiguate keyword-based queries in the search process. The system also takes the advantages of a classifier and a prebuilt ontology to support web page crawler for querying and classifying web pages of respected scholars.

Jomsri et al. in [13] proposes a personalized re-ranking technique for searching academic papers which engages the paper’s social tags, title and abstract to develop the indices and individual researcher’s profiles. The “tags” are scholar known keywords, which are assigned to the interested articles by means of a social bookmarking system such as CiteULike.

However, the significant feature of our approach is that it benefits from richer reference ontology by merging two relevant domain categories which increases the concept coverage and, in turn, recommendation accuracy. Indeed, the new reference ontology is able to capture broader and deeper topic areas of Computer Science, and consequently, represents comprehensive scholar’s background knowledge. Moreover, to disambiguate the key terms, Wikipedia is employed which provides live and broader source of key terms in the scholar’s domain. Finally, the proposed approach catches richer knowledge resources compared to the aforementioned approaches.

3. Methodology

We adopted a basic form of ontology, a hierarchy of terminologies [14], as a framework for user profiles to model the user prior knowledge as well as user preferences. The ontology should contain the fundamental concepts of user’s knowledge, the collection of concepts which users have known about some topics. Moreover, the required data for development of user profile should be collected from the user context, either implicitly or explicitly [15]. The initial concepts of user knowledge should be also collected from some textual resources explicitly, but the updating process can be performed in an implicit manner when users engage the system.

3.1. Resource of Scholar’s Knowledge

Scholars study papers to enhance their current knowledge and get into a deeper understanding of interested subjects. Relevant articles provide a researcher with opportunities to make connections between their prior knowledge and the “new” material being studied [16]. Therefore, the set of “reading material” is one source of scholar background knowledge which contains key concepts they know. Additionally, scholars obtain basic scientific knowledge through “formal educations” by taking courses at higher education institutes. In fact, the content of courses is the primary knowledge which scholars had previously learned before starting research. Besides, scholars incorporate in various academic activities such as research projects, teaching, etc., and ultimately publish their contributions at their homepages [17]. Fortunately, academic homepages are a publicly available rich source of knowledge about the academists including research interests, CV, publications, and the similar [18].

Figure 1 depicts the “textual collection” of scholar’s academic knowledge. We emphasis on textual information since it is the most prominent resource of published knowledge in the community.
Moreover, some resources such as blogs, though are textual, do not contain professional knowledge. Thus, they are excluded from the collection. Having identified knowledge resources, we apply statistical text analysis and information retrieval approaches [19] on the knowledge resources to logically represent them in a declarative formalism such as ontology structure.

3.2. Construction of User Profile

In our approach, the user profile is constructed using a set of weighted concepts which interconnects the concepts through a hierarchical relation instead of a simple list of keywords or Bag Of Words (BOW) [20]. Each concept contributes into a particular topic in the domain, and semantic relationships are explicitly declared through the hierarchy by “is-a” and “part-of” structure [21]. Non-leaf concepts contain several sub-concepts that attribute the respected super concepts. In general, this structure mimics the real model of scholar’s knowledge in mind where each new concept is grounded on some known concepts with a degree of dependency.

In order to link various concepts extracted from knowledge resources, the use of prebuilt structure such as domain ontologies using ontology engineering approaches is proposed [22]. The prebuilt structures not only save the construction time but also provide the desired quality for the final artifact. To integrate concepts extracted from the knowledge resources, e.g., reading materials and homepage’s information, an extended reference ontology called X-ODP, which is built up by merging ODP with VLIB2 (The World Wide Web Virtual Library) for subfields of Computer Science, is employed. Such reference ontology serves as a general framework for representing user model with sufficient knowledge coverage.

Figure 2 represents the structure of the X-ODP development. As shown, ontological concepts from the ODP and VLIB are merged and refined (aligned) together to develop a more complete domain ontology. The idea of developing an enhanced reference ontology is that ODP is not sufficient for capturing the whole areas of scholar’s knowledge in Computer Science. For example, there are no subjects for representing “text mining” or “back propagation algorithm” concepts. Besides, concepts which represent scholar’s knowledge are sometime multi-disciplinary.

Table 1 represents a sample derived from X-ODP hierarchy, containing two super concepts which represent the user’s background knowledge in the field of Computer Science: “Artificial Intelligent” and “Knowledge Management”. Moreover, concepts from the knowledge resources are assigned a weight which represents the strength and intensity of user knowledge in the field of study. For instance, the user has stronger background knowledge in “Artificial Intelligence” topic because the topic (concept) weight is 58.4 compared to “Knowledge Management” with weight 47.6. Similarly, the user has richer knowledge in “Machine Learning” (weight 35.3) than the other four subtopics with weight 13.5, 22.7, 18.0, 23.2 at the first level (red numbers).

For user profiling, we use the top four levels of the X-ODP hierarchy and employ popular statistical text analysis approach to identify the important concepts based on the their weights, i.e., the higher weight a concept has, the more important it is. The most important concepts are identified by the \( \text{tf} \times \text{idf} \) weighting scheme [23], given in Formula (1). The weight \( w(t,d) \) calculates the weight term \( t \) in a document \( d \), and it is a function of the frequency of \( t \) in the document, \( (tf,d) \), which is the number of documents that contain the term \( t \) (df) and the total number of documents in the given collection (N).

The important concepts are then disambiguated and mapped to the corresponding concepts within the reference ontology. Wikipedia is used to disambiguate the uncertain and vague terms [24]. The mapping uses Wikipedia as an indexing mechanism which assigns similar concepts (grammatical variant, synonymy, or semantically relatedness) with the known meaningful concepts in the domain.

![Figure 2. The overall structure of reference ontology development using two pre-built ontologies](http://vlib.org/Computing)
Moreover, the proposed weighting schema exposes two important properties: Firstly, each concept in the hierarchical is associated with a weight which identifies the intensive degree of user’s knowledge in each concept, and secondly the weight of super concept shares from underlying sub concepts. This weighting schema supports hierarchical order the concepts where updating a weight will spread to all of its super concepts thru the root concept.

4. The System Framework

Figure 3 represents the framework and components of the recommender system including user profiling, similarity computing, and filtering. It is divided by a dashed line in two parts: the upper part deals with ontology construction which combines three types of information: 1) the new reference ontology, X-ODP 2) terminology extraction which extracts concepts from knowledge sources including scholars’ articles and homepage’s content, and 3) the Wikipedia for disambiguation process [25]. This part encompasses the training phase of the framework and is carried out once. The lower part deals with concept extraction from candidate articles queried from the CiteSeerX digital library, concept disambiguation, and similarity computing between the user’s profile and the VSMs of candidate articles. Also, the Wikification process disambiguates the concepts of candidate articles.

The output of similarity computing is a set of candidate articles that are further filtered out using a threshold and then re-ranked. In order to update the user profiles, the articles which have been selected by the users for studying will be processed and their concepts are extracted and disambiguated. In fact, the new key terminologies found within the recommended articles will be used to update the user profile by either increasing the weight of existing concepts or add entries to the hierarchy. In the following subsections, the components of the framework are explained in detail.

<table>
<thead>
<tr>
<th>Topics/ Subtopics</th>
<th>tf*idf</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Artificial Intelligence</td>
<td>58.4</td>
</tr>
<tr>
<td>o Agents</td>
<td>13.5</td>
</tr>
<tr>
<td>o Agent Technologies</td>
<td>8.60</td>
</tr>
<tr>
<td>o Agent_Communication_Languages</td>
<td>2.17</td>
</tr>
<tr>
<td>o Applications</td>
<td>12.0</td>
</tr>
<tr>
<td>o Mobile_Commerce</td>
<td>3.50</td>
</tr>
<tr>
<td>o Belief_Networks</td>
<td>22.7</td>
</tr>
<tr>
<td>o Fuzzy</td>
<td>18.0</td>
</tr>
<tr>
<td>o Genetic Programming</td>
<td>23.2</td>
</tr>
<tr>
<td>o Algorithms</td>
<td>15.2</td>
</tr>
<tr>
<td>o Machine_Learning</td>
<td>35.3</td>
</tr>
<tr>
<td>o Case-Based_Reasoning</td>
<td>37.1</td>
</tr>
<tr>
<td>o Datasets</td>
<td>40.8</td>
</tr>
<tr>
<td>* Knowledge_Management</td>
<td>47.6</td>
</tr>
<tr>
<td>o Agents</td>
<td></td>
</tr>
<tr>
<td>o Agent Technologies</td>
<td>......</td>
</tr>
<tr>
<td>o Agent_Communication_Languages</td>
<td>......</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

Table 1. An example of weighted concept hierarchy from X-ODP which represents a typical user’s knowledge

![Figure 3. The structure of semantic based user modeling and similarity computing for scholars recommendation using background knowledge](image-url)
4.1 Concept Extraction

In our proposal, the component Concept Extractor (CE) extracts concepts from two information sources; the set of reading articles and the scholar homepage which involves publications and research interests. For each scholar, a corpus of text files containing a number of these two sets is collected. Having analyzed the content of corpuses by the Information Retrieval approaches [19], the top N terms which are important are extracted. We rename the terms as concepts if they are multi-words, discriminant, and having higher tf*idf in the corpuses (using a threshold) so that they represent the user’s knowledge effectively. In order to update the user’s background knowledge over the time, CE extracts the new concepts from the rated articles and updates the respected weights accordingly.

4.2 User Profile Development

The user profile developed at training phase should be updated at running phase. In order to organize the ontological concepts in which representing the user’s prior knowledge, the reference ontology or X-ODP is engaged as a template hierarchy (TH) at the training phase. The initial weight of each concept in TH is set to zero and increased gradually by each concept’s weight. In fact, the concepts from an initial set of training data are extracted by CE and disambiguated using Wikipedia, and then mapped into TH. Formerly, the outliers and noisy concepts are removed using a threshold value.

To update the TH at running phase, the same process is performed on rated articles. Likewise, the concepts are extracted from the new rated articles by CE and the same weighting procedure as training phase is performed. The weight of new concepts is aggregated to existing concept’s weight to update the TH. However, new concepts are inserted to TH additively.

The user also can discard some concepts from TH by giving negative rating [26]. This rating models the cognitive behavior of the user when the content of recommendation does not match with his background knowledge. Another approach is to allow user manually decrease or increase the concept weights in order to update his knowledge level for desired topics.

Finally, to keep concept weights in TH reasonably small, a weight-reduction procedure is periodically run to decrement all non-zero weights by the minimum value in TH. For example, if the maximum weight value exceeds from a predefined MAX value, then all non-zero values in TH are subtracted by the minimum value in TH.

4.3 Wikification

There is an ambiguity problem [27] among extracted concepts since syntactically different concepts may refer to the same concept and vice versa. Thus, direct mapping between extracted concepts to TH concepts is not always accurate. For example, the terms “Fuzzy System” and “Fuzzy Logic” are semantically similar as they imply the same knowledge. In order to improve the accuracy of concept mapping, a semantic enrichment approach based on the Wikipedia is employed [28].

Wikipedia captures a wide variety of concepts that different people with different knowledge refer to the same concept [29] and contains detailed subjects where other directories such as WordNet [30] lacks sufficient descriptive knowledge. Each Wiki page contains a full description of a single concept and respected topics that describe the synonym concepts (subjects) [25].

To perform term disambiguation, we identify and consolidate neighboring terms which are semantically similar to the terms retrieved from the article by TE [27]. Thus, all semantically relevant synonyms of a reference concept from Wikipedia are considered as variations of that term and the initial “concept vector” is updated with the terms in the neighborhoods of all the similar concepts retrieved from the Wikipedia. In order to select appropriate neighborhoods for each concept, a threshold is employed to prune less relevant terms.

4.4 Similarity Computing

The goal of similarity computing is to find the most relevant articles among numerous articles retrieved by a digital library. In order to filter out and re-rank the retrieved articles from the digital
library, it is required to compare the similarity between each article and the user’s profile. Thus, the concepts from the articles are extracted and represented by means of Vector Space Model [31]. Since the concepts of both sides are ambiguous, a disambiguation process or so-called Wikification [28] against the concepts is performed. In fact, Wikipedia bridge concepts of both sides to unified concepts.

The similarity or relatedness of representative concepts with the user profile is performed through the following steps [32]:

- The concepts with non-zero weight in TH are used to build a temporary VSM vector, and the weights are normalized in the range of [0..1] to be comparable with the article’s VSM weights.
- For each candidate article retrieved from the digital library, a VSM vector containing important concepts and associated weights is developed. The weights are calculated by tf*idf scheme and then normalized by the same manner.
- The Cosine Similarity formula [33] is used to measure the similarity between the n-dimensional VSM vectors. Articles with highest similarity (using a threshold) are classified as “good” and then re-ranked to be recommended to the user.

The cosine similarity (Formula 2) measures the similarity between two n-dimensional vectors by finding the cosine angle or \( \cos(\theta) \) of them. For two given vectors \( V \) and \( S \), the cosine similarity is represented using a dot product operator.

\[
\text{Similarity} (v, s) = \cos(\theta) = \frac{\vec{v} \cdot \vec{s}}{||\vec{v}|| \cdot ||\vec{s}||}
\]

5. Implementation

The objective of implementation is to examine the applicability and the performance of the proposed approach in terms of accuracy. We implemented a prototype of the framework with assistance of fifteen volunteer scholars at different levels including masters, PhD students, and graduated researchers from the Faculty of Computer Science and Information Systems (FSKSM) at UTM. For each participant, a collection of 20 papers in line of their current research which have been recently studied and understood are collected. Other information such as homepage address and their publications are investigated through the web non-intrusively.

5.1 Development of Reference Ontology

Initially, we developed the X-ODP reference ontology by merging ODP and VLIB hierarchies. We first engaged ODP as the backbone of subject hierarchy and extracted all relevant subjects to the Computer Science domain. Currently, it contains 786,225 subjects varying from 3 to 5 abstraction levels. The fragment of Computer Science topics contains 8,471 entries. We applied heuristic filters iteratively in five stages to sift irrelevant topics from ODP, i.e., those topics that belong to the domain but are not relevant to the research domain. Some heuristic information is also used to enhance the filters, such as semantic relevance of topics to research topics using several “conference themes” from the web. For example, “Development Kits” is a trivial topic in the domain but entirely irrelevant to a “research subject” since not only no research is conducted about it but also it does not convey substantial scientific knowledge.

In the next stages, subjects and their relationships from VLIB directory is engaged to enrich the subject hierarchy of ODP. In this stage, cross-checking of subjects and inclusion of missing subjects for the three topmost levels are applied. The resulting hierarchy contains five class levels, 183 disjoint classes in the tree.

5.2 Experiments

This experiment constructs scholars profile and examines the framework in a real scenario by filtering and re-ranking articles retrieved from the CiteSeerX digital library. First, the articles in each collection (corpus) are transformed into plain text and knowledge-free parts such as images, formulas, tables, acknowledgments, and references are pruned. The term extraction is then performed on each text file in the corpuses which produced a set of more frequent terms based on the tf*idf in their
particular research topic. Next, the key terms are disambiguated using Wikipedia and mapped into TH for each participant researcher.

To test the accuracy of recommender framework, an offline experiment is carried out. The participant scholars have been asked to search three independent queries on their own subjects using CiteSeerX and score the articles based on three point rating as follows:

I= Inappropriate, M= Moderate, and A= Appropriate

Besides, the ratings are performed independently on two cases: 1) articles queried against CiteSeerX, and 2) articles that are filtered and re-ranked using the background knowledge profiles. The ratings “M” and “A” are considered to be the same value, and an average is calculated for both cases. The average values of rating for different scholar are shown in Table 2. As shown, scholar #1, for instance, was interested in 1.3 of the top-5 CiteSeerX articles (X-Top 5) while 2.2 of articles on average after filtering and re-ranking using his background knowledge (BK-Top 5). Similar results appeared when the user was interested in 3.5 of the top-10 articles before, and 4.6, after applying our approach.

Table 2. The average rating of 15 scholars on 4 cases using three different search queries

<table>
<thead>
<tr>
<th>Users</th>
<th>X-Top 5</th>
<th>X-Top 10</th>
<th>BK-Top 5</th>
<th>BK-Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3</td>
<td>3.5</td>
<td>2.2</td>
<td>4.6</td>
</tr>
<tr>
<td>2</td>
<td>1.7</td>
<td>4.1</td>
<td>1.9</td>
<td>4.5</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>3.3</td>
<td>3.1</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>2.1</td>
<td>4.0</td>
<td>2.6</td>
<td>5.3</td>
</tr>
<tr>
<td>5</td>
<td>1.4</td>
<td>3.8</td>
<td>3.1</td>
<td>6.0</td>
</tr>
<tr>
<td>6</td>
<td>3.1</td>
<td>4.3</td>
<td>3.5</td>
<td>5.7</td>
</tr>
<tr>
<td>7</td>
<td>1.9</td>
<td>5.5</td>
<td>4.6</td>
<td>6.3</td>
</tr>
<tr>
<td>8</td>
<td>1.9</td>
<td>4.1</td>
<td>3.7</td>
<td>6.3</td>
</tr>
<tr>
<td>9</td>
<td>2.4</td>
<td>4.4</td>
<td>3.7</td>
<td>5.0</td>
</tr>
<tr>
<td>10</td>
<td>3.1</td>
<td>4.1</td>
<td>2.2</td>
<td>4.6</td>
</tr>
<tr>
<td>11</td>
<td>3.5</td>
<td>4.5</td>
<td>3.7</td>
<td>4.7</td>
</tr>
<tr>
<td>12</td>
<td>3.1</td>
<td>4.0</td>
<td>2.4</td>
<td>4.1</td>
</tr>
<tr>
<td>13</td>
<td>3.5</td>
<td>4.5</td>
<td>3.3</td>
<td>4.6</td>
</tr>
<tr>
<td>14</td>
<td>3.1</td>
<td>4.0</td>
<td>3.3</td>
<td>4.6</td>
</tr>
<tr>
<td>15</td>
<td>3.7</td>
<td>4.5</td>
<td>4.0</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Figure 4 also represents a pictorial view of the precision (accuracy) obtained for Top-5 and Top-10 ratings in both cases; without and with the presence of background knowledge in each case. As shown, for both ratings, a significant improvement in precision by incorporating the scholar’s background knowledge in filtering and re-ranking is achieved, i.e., for scholars numbered 1 to 11, the improvement is significant. It should be noted that the improvement in precisions for scholars number 12 thru 15 are venial (low) because they study in Information Technology and Management fields which have very few common topics with the Computer Science domain. We included these scholars to the experiment to act as negative examples.

However, the tolerance in improvement percentage for both cases is a function of several factors including the overall cognition of respected researcher in the assessment task, the accommodation degree of supplied articles by the researchers with their current research state, and the lack of academic homepage for some fresh researchers.

Figure 4. The comparison of precision for 15 scholars on three different queries

6. Conclusion and Further Works

In this paper, a framework for exploiting scholar’s background knowledge into a recommender system for digital libraries is presented. The significance of using background knowledge in a
recommender system in terms of usability and performance in the Computer Science domain is investigated. Due to lack of concept coverage by ODP, we enriched the reference ontology thru augmenting VLIB directory to ODP which provides richer ontological concepts for user modeling. We also employed both statistical Information Retrieval approaches and Vector Space Model for representing scholar’s knowledge, cosine similarity for measuring closeness of articles to the user profiles, and Wikipedia for concept disambiguation and unification. The hierarchal structure of scholar’s background knowledge facilitates the ease of concept similarity calculation as well as profile updating. The simple organization of Template Hierarchy assists to keep the ontology current and updated over time. In order to capture the shift in the user’s knowledge over time, the weights in the hierarchy are reconsidered periodically. An offline experiment over CiteSeerX by fifteen researchers presents a significant improvement in terms of precision using scholar’s background knowledge.

For further improvements, we aim to accommodate complementary sources of knowledge such as scholar’s formal educations, other homepage information such as lecture notes and project descriptions, and mediated profiles provided by digital libraries. However, the concepts from different knowledge resources are not in the same granularity, and therefore, merging ontologies will be a challenging task. Potential improvement may be also achieved with an enhanced similarity computing such as thematic term set similarity proposed in [34] for English language.

Acknowledgement

This work is supported by the Ministry of Higher Education (MOHE) under the Exploratory Research Grant Scheme (Vote No. R.J130000.7828.4L051). The authors would like to thank the management of Universiti Teknologi Malaysia (UTM) and the Research Management Center (RMC) for the support in making this study a success.

7. References


