Incorporating Scholar's Background Knowledge into Recommender System for Digital Libraries

Bahram Amini, Roliana Ibrahim, Mohd Shahizan Othman, Hamid Rastegari
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia (UTM), Malaysia
avbahram2@live.utm.my , roliana@utm.my , shahizan@utm.my, rhamid2@live.utm.my

***Abstract*- In recent years, recommender systems have received increasing attention in digital libraries since they assist scholars to find the most appropriate articles. However, a major problem of such systems is that they don't subsume user background knowledge into the recommendation process and scholars have to manually sift irrelevant articles obtained in response of queries. Therefore, a great challenging task is how to include scholar's knowledge into personalization process and filter out articles accordingly. To address this problem, a novel cascade recommender framework which incorporates scholar's background knowledge using ontological concepts into the user profiles is proposed. The framework exploits standard ODP structure as ontology modeling as well as lexicographic database (WordNet) for concept disambiguation. The primary experiment over CiteSeerX digital library indicates an increase in user satisfaction.**

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

I. INTRODUCTION

The rapidly growing amount of data in digital libraries is at explosion rate, providing scholars a good source of data for search the interesting information. Unfortunately, traditional search engines returns huge amount of relevant and irrelevant information in response to typical keyword-based searches, rising "information overload" [1]. In other words, non special search engines of digital libraries overwhelm scholars by numerous data that usually doesn't fit their actual needs. To address this problem, personalization approaches have recently been emerged to offer users more relevant result than the traditional search engines by filtering irrelevant or less relevant results and recommend articles closest to the user prior knowledge. In fact, personalized approaches aim to detect user tastes and conforms the result of query to the actual needs of the user by filtering unrelated results and recommending items that deemed user to be interested in.

Personalization for scholars is the process of tailoring and customization of digital library outputs to the individual scholar's need [2]. Misunderstanding the user intention is the source of many recommender system failures. Although educational recommenders share the same properties as, for example, recommenders for electronic news (i.e. assisting users to choose the most relevant news from a large news directory), there are however some striking features that make it impossible to directly apply existing solutions for those applications to digital libraries. For instance, recommendations for scholar should not be guided only by the scholar's preferences but also by particular characteristics such as their research pattern, learning goal, and background knowledge [3].

In scholar domain, background knowledge circumscribes the user expectance and helps recommender to understand the user intension. Misunderstanding of user intension often arises from omitting the user background knowledge. Therefore, one promising step to improvement of recommender system for scholars is to incorporate their prior knowledge into the recommendation process [4]. User background knowledge helps recommender systems effectively to determine the user's information scope and to recommend the most relevant articles close to the user's knowledge [5]. Once the user's prior knowledge specified, the recommender system is able to prune the search result obtained from digital library and fits the results to his knowledge level.

The purpose of this paper is to investigate the impact of user's background knowledge in digital libraries and propose a comprehensive framework to incorporate scholar's background knowledge into recommendation process for digital libraries. The rest of the paper is organized as follows: Section 2 defines the scholar's background knowledge. Section 3 discusses the primitives of modeling background knowledge. Section 4 describes a framework for incorporating background knowledge into recommendation process in cascade style. Section 5 reviews the relevant works and Section 6 deals with implementation and results. Finally, the conclusion and future works has been explained in Section 7.

II. BACKGROUND KNOWLEDGE

The term background knowledge and prior knowledge are generally used interchangeably. McDonald et al. [6] defines background knowledge as what individual already knows about a subject. Biemans et al. [7] also define prior knowledge as all knowledge learners have when entering a learning environment that is potentially relevant for acquiring new knowledge. These definitions are quite similar though use different wordings. Prior knowledge helps to understand the new knowledge by bridging existing concepts to new concepts in a domain. Here, the terms "concepts" and "prior knowledge" are used interchangeably. In educational environments, prior knowledge has great influence on

learner performance because perception of new topics can be done easily based on known concepts [3]. For example, if learner has some knowledge about the Classification, he will better understand reading the book Text Mining. Similarly, recommender system may better understand the meaning of scholar queries if it knows about the user prior knowledge. Therefore, recommender system will filter out the result of search by making correspondence between the topics user already knows and relevant topics.

III. Modeling Background Knowledge

Any personalization system relies on using some forms of user profile [8]. We adopted a simple form of ontology, a hierarchy of terminologies [9], as user profile to model the user prior knowledge as well as user preferences. Thus, ontology contains the primary concepts of user's background knowledge, the collection of concepts which user has already known about a topic.

The required data for development user profile should be collected from the user context either implicitly or implicitly [10]. Initial concepts of the user knowledge should be collected from some resources explicitly, but the updating process is performed in implicit manner when users engage with the system. However, explicit data collection from user feedbacks burden additional tasks on users and the system designer [11].

A. *Resource of Scholar's Background Knowledge*

Scholars study papers to enhance their current knowledge and get into deeper understanding of interested subjects. Relevant articles provide researcher with opportunities to make connections between their prior knowledge and the 'new' material being studied [12]. Therefore, the set of "reading material" is one source of scholar background knowledge which contains key concepts they understood.

Additionally, scholars obtain basic scientific knowledge through formal educations by taking courses at higher education institutes. The content of courses indicates basic knowledge scholar have learned. Scholars also incorporate in various academic activities such as project, teaching courses, etc and then report the contribution at their homepages [13]. Academic homepages are a publicly available rich source of knowledge about academists' proficiency including research interests, CV, and publications [14].

B. *Construction of user profile*

User profile is constructed using a set of weighted concepts which interconnects the concepts through a hierarchical relation rather than simple list of keywords. Each concept defines a particular topic from the domain, and semantic relationships are implicitly declared through the hierarchy structure [15]. Each concept includes several slots, or set of attributes that describes the concepts, depending on the type and complexity of underlying topics.

Manually extraction of concepts from the body of knowledge resources is time consuming, tedious, and error prone task [16]. Therefore, the use of pre-build structures such as domain ontologies as well as automatic concept extraction and ontology development approaches is preferred [17]. Pre-build structures not only save the construction time but also provide appropriate quality for the final artifact.

To integrate concepts extracted from knowledge resources, i.e. reading materials and homepage's information, a reference ontology is employed. Such reference ontology serves as an outline and general framework for modeling user profile. A well-know reference ontology that is widely used in information retrieval is Open Directory Project (ODP[1]) [18]. Figure 1 (first column) depicts a sample of concept hierarchy obtained from ODP, containing two slots which represent the user background knowledge in the field of Computer Science: "Artificial Intelligent" and "Knowledge Management". Moreover, concepts from the knowledge resources as well as concepts in the reference ontology are assigned a weight which represents the strength and intensity of user knowledge in the context.

Topics/ Subtopics	Weights
*** Knowledge_Management**	76
○ Agents	…..
▪ Agent_Technologies	…..
▪ Agent_Communication_Languages	…..
▪ Applications	…..
• Mobile_Commerce	…..
• Education_and_Instruction	…..
○ Belief_Networks	…..
○ Genetic_Programming	…..
▪ Algorithms	…..
……	…..
*** Artificial_Intelligence**	84
○ Agents	35
▪ Agent_Technologies	11
▪ Agent_Communication_Languages	17
▪ Applications	20
• Mobile_Commerce	5
• Education_and_Instruction	12
○ Belief_Networks	22
○ Fuzzy	18
○ Genetic_Programming	30
▪ Algorithms	15
○ Machine_Learning	50
▪ Case-Based_Reasoning	37
▪ Datasets	40

Figure 1: An example of weighted concept hierarchy from ODP representing a user prior knowledge.

For user profiling, we use the top four levels of the ODP hierarchy in "Computer Science" subfield and supervised classification approach [18] to associate the weight of concepts extracted from knowledge sources to the ODP hierarchy. As user rates an article, the topmost important concepts, calculated by term frequencies [19] and

[1] www.dmoz.org

represented by Vector Space Model (VSM) [20], are mapped to the corresponding concepts and the weights are increased.

The proposed user model exploits two distinct weighting schemas: Firstly, each concept in the hierarchal model is associated with a weight which identifies the degree of user knowledge in the given topic area, and secondly, a weight is assigned to the underlying slots in each concept which specifies the intensity of user knowledge in the sub-topics among others. For instance, in Figure 1 (second column) two types of weight are assigned to the concepts and underneath slots. As shown, the respective user has stronger background knowledge in "Artificial Intelligence" category because respected weight is 84 compared to "Knowledge Management" category which its weight is 76. Similarly, the user has richer knowledge in "Machine Learning" (weight 50) than the other four subtopics (slots) with weight 35, 22, 18, 30 in the same topic.

A. *Similarity Measurement*

In the proposed framework, the retrieved articles by digital library are further filtered and re-ranked to match with the use's model. Therefore, a similarity measurement between the concepts in the user's model and concepts of retrieved articles are performed. We use the Cosine Similarity measurement which accepts two concepts presented as vectors by VSM model and calculates the similarities in terms of relatedness [21].

IV. System Framework

Figure 2 represents the conceptual framework of the recommender system. It engages scholar's background knowledge to develop user profile and employs a cascade hybridization approach [22] to filter out and re-rank articles captured from CiteSeerX digital library. Since selecting good articles is a multi step process, cascade approach improves previously filtered articles by the digital library.

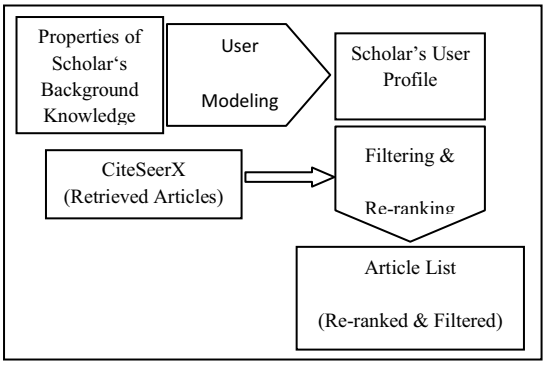

Figure 2: The conceptual framework of the recommender system: A cascade approach using scholar's background knowledge

Figure 3 represents the overall structure of recommender system for scholar including user profiling, similarity computing, and filtering, divided by a dashed line making two parts. The upper part deals with ontology construction which combines three types of information together: 1) a sub part of ODP hierarchy, 2) a terminology list extracted from the set of knowledge sources including individual scholar's articles and homepages, and 3) the WordNet[2] terminologies [23]. This part encompasses the training phase of the framework and is carried out once. As user profile is created, the process is no longer in use. The lower part, running phase, deals with similarity computing between results of search query captured from digital library with the concepts in user profile.

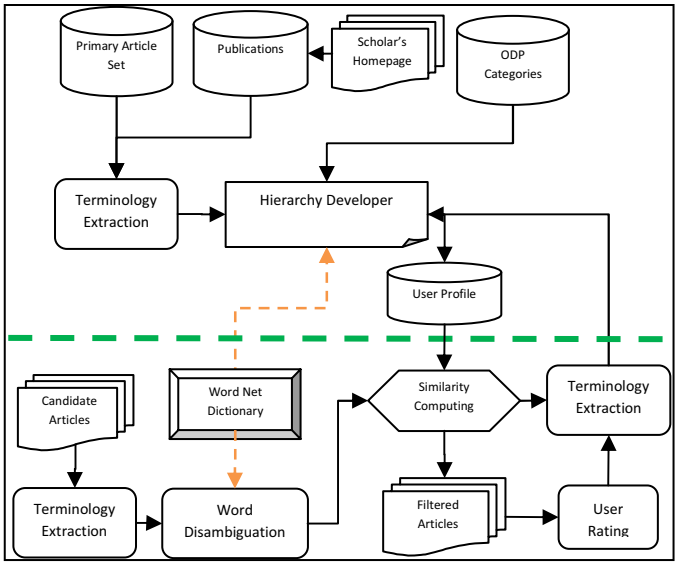

Figure 3: The structure of semantic based user modeling and similarity computing for scholars recommendation using background knowledge

The result of similarity computing which is a set of candidate articles will be filtered out using a predefined threshold and then sorted. The articles which user has chosen to read or rated are further processed and underlying concepts are used to update user. In fact, new important terminologies found in recommended articles will be used to update the user profile by increasing the weight of existing concepts or add to entries of the hierarchy. In the following, the components of the framework are explained in detail.

A. *Terminology Extraction*

Terminology extractor (TE) receives article files from the two information sources; the set of primary articles and the scholar homepage content. The scholar homepage contains links to the publishing articles, research interests, and teaching courses [14]. In cooperation with TE, a web crawler [21] mines the content of each homepage to extract scholar's publications and CV. By analyzing the content of associated text files in the corpuses by IR approaches [18][24], the top N frequent terms that are more important are extracted. Concepts that appeared more frequent in articles represent the richness of user's knowledge. In order to update the user's background knowledge over the time, TE is activated in the same procedure and updates the newly discovered or

[2] http://wordnet.princeton.edu

existing concepts as user study articles and feedback (rate) new articles.

B. *Hierarchy Development*

In order to organize the ontological concepts representing user's prior knowledge, a subset of ODP hierarchy called Canvas Hierarchy (CH) in the field of Computer Science is exploited. Then, the content of hierarchy is updated in 2 steps as follows:

Step 1: The key terms from initial set are extracted by TE and mapped into the hierarchy. This set is an initial content of knowledge sources that user is known. The set of terms extracted from each text file are listed based on the term frequency (tf) and then normalized by using the traditional tf*idf and VSM methods [25], called *tw*.

Step 2: The outlier and noisy terms are removed using a threshold value. Then, remaining terms are used to mark the corresponding terms exist in CH and increase corresponding concept weight by *tw*. The initial weight of each concept is 0 and increased continually by each *tw*.

Figure 4 depicts an example of calculating concept weights and developing CH hierarchy process. As shown, weights of concepts as well as slots are initially set zero. After extracting terms from the corpus articles and computing tw for each term, corresponding weight field in CH has been change with Σ tw.

To update the CH during the running cycle, the similar process (Step 1 and 2) is performed on new text files obtained from knowledge sources which user accomplished. Likewise, the terms are extracted from the body of text files by information extraction approach and the same weighting procedure as above is performed, and for each new concept vector the corresponding tw is computed. New concepts are used to incrementally add to CH terms.

User also can discard some concepts (slots) from the CH by giving negative rating [26]. This type of rating models the cognitive behaviour of 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 specific concept.

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

A. *Term Disambiguation*

Due to the word ambiguity problem [27], traditional relationship between keywords and corresponding articles in scholar domain is not accurate because some words in the text files have different meanings. For example, the terms tree, branch, and leaf have different meanings depending the underlying context, i.e., computer science or agriculture. Consequently, direct mapping between extracted terms and CH terms will not straightforward. In order to improve the accuracy of concept mapping, a semantic enrichment approach based on the WordNet is proposed [27]. Hotho et al. [23] show that incorporating feature of WordNet can produce positive effects and improve text classification process.

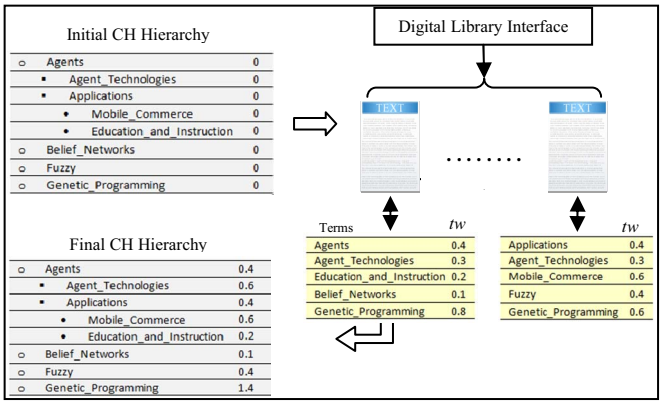

Figure 4: An example of calculating TF*IDF for each article and updating CH hierarchy

WordNet is a large lexical database of English words including nouns, verbs, adjectives and adverbs, grouped into sets of synonym synset. Synsets are interlinked by means of conceptual or semantic relations. For instance, the words "car" and" automobile" belong to the same synset and can be used interchangeably because they are synonyms and refer to the same concept. WordNet provides general definitions and various semantic relations between these synonym sets. Scott et al. [24] suggest the use of word's information from WordNet in the training set by expanding each word in the training set with their synonyms to avoid word ambiguities and misconception.

To perform term disambiguation, we identify and consolidate neighboring nouns which are semantically identical (or closed) to the terms retrieved from the article through TE task [27]. Thus, all semantically relevant synonyms of a reference concept from WordNet are considered as variations of that term and the initial "concept vector" is updated with the terms in the neighborhoods of all the synonyms retrieved from the WordNet [28].

In order to select appropriate neighborhoods (synset) for each concept, a threshold is employed to prune less relevant terms [28] and avoid the noise effect [29]. The terms in WordNet are organized in hierarchy structure. The sub- and super- relation among terms helps to distinguish the granularity of words. When a term matches with a term in level L of the hierarchy, every subsequent (lower levels) term of that term will be included in the user profile and discarded upper levels terms of level L.

B. *Similarity Computing*

The goal of similarity computing is to find the most relevant articles from enormous articles acquired by digital library. For this purpose, the similarity or relatedness of representative concepts (extracted by TE module) with the background knowledge concepts maintained by CH is calculated [30] through the following steps:

1- Using VSM method, the non-zero terms in CH along with associated weights are used to make a concept vector, and the weights are normalized (in range of [0..1]) to be comparable with weights of article's concept vector as mentioned before.

2- For each article in the "Candidate Articles" corpus, a concept vector including more frequent terms and associated weights is computed. The weights are calculated by tf*idf weighting scheme and then normalized in range of [0..1] to be comparable with CH weights in step (1). The value of tf*idf (term-frequency multiply inverse document frequency) is calculated by formula (1) [31]. The tf*idf weight, w(t,d), refers to 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).

$$w(t,d) = \frac{tf * \log\left(\frac{N}{df}\right)}{\sqrt{\sum (tf,d)^2 * \log\left(\frac{N}{df}\right)^2}} \qquad (1)$$

3- For measuring similarities, the Cosine Similarity formula [23] can be used to measure the similarity between CH vector and each article's concept vector, called SM. This measurement is the default computation for information retrieval and can be used as a benchmark for improvement [32]. In this step, classification of articles is performed because articles are assigned to their corresponding concept vectors.

4- The articles are then decreasingly ordered based on SMs (degree of relevancy) and a threshold is employed to filter out less relevant articles based on SMs because an effective ranking of articles will improve the predictive performance of recommender system and provides more accurate recommendations to users [32].

The resulting collection of articles is more relevant to the user's background knowledge and richer than the primary collection because they normally contain more understood concepts. By changing the threshold value in controlled experiments, the degree of appropriateness of articles from the user point of view can be determined. Figure 5 represents a data flow of aforementioned steps. In step 5, TE is performed to extract new concepts from the new rated articles which will be augmented with the current CH to keep the user profile up to date.

V. Related Work

There have been several attempts to incorporate contextual and feature information into recommender system and achieve object filtering accordingly [33]. However, there is no pure evidence which incorporates user's background knowledge into recommender systems, exclusively for scholar domain. Therefore, the works discussed here are possible trends to this approach.

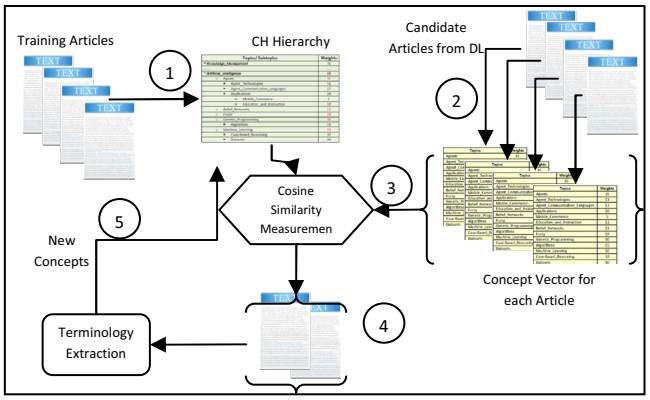

Figure 5: An example of article classification in four steps based on user's background knowledge

Kayed et al. [34] propose a web site ranking algorithm based on the domain ontology. They show how to measure the closeness (relevancy) of retrieved web sites to user queries and use electronic commerce ontology to re-rank them accordingly. The approach employs "concept ontology frequency" as a basis of calculation the relevancy of retrieved web pages. To construct the ontology, they use 10 popular definitions for ecommerce domain and extract 34 most common concepts from them that cover about %90 of domain concepts. The documents (web sites) are extracted by a general search engine such as Google and assign 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 then re-ranks the associated documents according to the number of occurrences.

The work in [35] proposes a hybrid recommender system (content- and collaborative- based) for EachMovie domain and integrates the linguistic knowledge in the process of learning semantic user profiles with a sense-based indexing schema. Semantic profiles are obtained by integrating machine learning algorithms in text categorization and a relevance feedback method with word sense disambiguation strategy based on the lexical knowledge of the WordNet. It exploits the idea behind the word sense disambiguation algorithm to disambiguate each word w by determining the degree of semantic similarity among candidate synsets and those of words in the context, a set of words surrounding w. The semantic similarity between concepts are computed by Leacock-Chodorow approach [36] which is based on the length of the path between each pairs of concepts in IS-A hierarchy.

The research presented in [37] implicitly builds ontology-based user profile based on user's information contents in internet by user's blogs, publications, home

pages, etc. An initial profile from ODP is constructed in hierarchical structure and is further learned by incorporating additional user details collected from the user's documents. It also applied WordNet as a lexicon syntactic pattern for hyponyms to augment the important feature of documents including tf*idf weighting scheme to improve the profile. The ontology-based user profile is further improved in collaborative manner by learning relevant knowledge from similar profiles in the community.

In [38], a personal ontology recommender system for Chinese digital libraries is proposed. It uses traditional cataloging scheme, as reference ontology, for classification. The reference ontology is extracted from the borrowing records of individual users as well as notes keyed by librarian which act as personal user profile on a specific subject. The personal ontology which represents user interests is used to filter out irrelevant or less relevant books based on keyword matching. It is assumed that the collection of keywords of loaned books represent the user interests. Each time user logs into the system, it extracts favorite topics of individual user from different sources and assemble them to builds up the user profile. The system is linked with check in/out information system and collects information in online manner from the user interactions. The cataloging information as well as loan information is used as the source for construction user's ontology.

The work presented in [39] proposes an ontology based information integration and recommendation for scholar domain. It employs five core technologies including: ontology based webpage crawler and classifier, information extractor and recommender, and user integration interface. The system exploits ontology approaches to enrich and disambiguate keyword based queries in search process. The system takes the advantage of pre-built ontology which support webpage crawler for querying the webpage of related scholars and the classifier in classification of webpage.

VI. RESULTS

To test the usability and applicability of the proposed approach for scholar domain, we conducted two case studies. The first one investigated the scholar opinion and tendency on using digital libraries and how they perceive the appropriateness of articles captured by general purpose digital libraries. The second examines the performance of the proposed approach in terms of accuracy and user satisfaction.

A. Case study 1:

To test the usability of incorporating background knowledge into digital libraries, we did interview with 15 international scholars at Department of Computer Science and Information Systems (FSKSM), UTM. They had full experience of using the most popular and powerful digital libraries including SpringerLink, ACM, IEEEXplore, Web of Science, Science Direct, and CiteSeerX. The participants have been questioned with the questions given in Table 1. In the second and third columns, the opinions and measurements of participants to the questions are depicted.

TABLE 1: THE STATEMENTS AND RESULT OF INTERVIEW WITH PARTICIPANTS

Statements	Agree%	Disagree %
Background knowledge is so important to be considered by digital libraries.	100	0
Currently, user's background knowledge is taken into account by digital libraries.	0	100
Users usually receive redundant articles from digital libraries in respond of different search terms.	100	0
The results of search queries from digital libraries are out of user's background knowledge scope.	80	20

As shown, 100 percent of scholars believe that background knowledge is very important to be involved by digital libraries and they confirmed that the most popular digital libraries they engaged do not support such important functionality. Moreover, the lack of background knowledge provides redundant articles as they visit some visited articles many times. Finally, 80 percents of scholars believe that the results of their queries are outside of their prior knowledge. For example, when scholars search for articles about general topics such as "Genetic Algorithms", they received broad articles such as "Evolutionary Algorithms" and "Fuzzy Systems" to very narrow topics such as "Bayesian Network Classification" and "Classification Algorithms". Table 2 shows the results of a typical query against SpringerLink[3] by the phrase "Genetic Algorithms" which returns about 109,000 articles in different sub-fields of "Artificial Intelligent", sorted based on its ranking strategy. Similar results are even worse or better existed on experiencing with other digital libraries.

TABLE 2: THE SUB-FIELDS AND THE CORRESPONDING NUMBER OF PAPERS PROVIDED BY SPRINGERLINK FOR A SPECIFIC QUERY

Rows	Sub-fields/ Subjects	# of Papers
1	Computation by Abstract Devices	18,293
2	Algorithm Analysis and Problem Complexity	18,280
3	Software Engineering	18,139
4	Computer Communication Networks	18,103
5	Database Management	18,020
6	Data Encryption	17,887
7	Engineering	533
8	Software Engineering/Programming	117
9	Computing Methodologies	133

As depicted in table 2, there are 9 different subfields for each main topic, say "Genetic Algorithms". It indicates that digital libraries only search for the terms in title, keyword, and abstract parts given by the user but not based on the user's background knowledge. Therefore, if the digital library knows about the topic or background knowledge of the user, for example "Algorithm Analysis and Problem Complexity" (row 2), the scope of result could be narrowed into 18,280 (about %17) of the original space, and therefore, the performance of navigation would drastically be

[3] http://www.springerlink.com

improved. Of course, the user can select the subfield terms manually that is inflexible and time consuming.

Similarly, as the number of queries increases, the complexity of finding appropriate papers in the topic of discourse is becoming worse and unmanageable, and therefore, the user will have to try different keywords extracted from the body of visited articles and provide more specific terms for his query. On the other hand, enquiry with detailed search keywords may lose the search accuracy because search engine of digital libraries focus on keywords appeared in the body of text but not the concept or meaning of terms since a term may have different meaning in different context.

As a result, a solution to tackle such complexity and inconvenient search is the use of a user profile in the form of taxonomy or terminology to distinguish the area of user interests and background knowledge [30].

B. Case study 2:

We implemented a paper based of the system and invited 7 volunteer scholars in different levels including master and PhD students and graduated researchers from FSKSM. For each participant, a collection of 20 papers which have been recently studied and understood about a specific subject are collected. The collection files are then preprocessed and refined into pure text in which knowledge-less parts including images, formulas, tables, acknowledgments, and references are removed. The term extraction is performed on each file in the corpuses which yielded a set of frequent terms in the respected subjects. Key terms are mapped into CH for individual scholar and weighting is done as key terms repeated in the articles.

To test the feasibility and accuracy of our approach, an offline experiment is carried out. Scholars have been asked to search two independent queries using CiteSeerX digital library on their own given subjects and score the articles based on three rating values as I= Inappropriate, M= Moderate, and A= Appropriate. Ratings are performed separately on articles obtained from CiteSeerX and after re-ranking by the new approach. The ratings "M" and "A" are considered the same and an average is calculated for both. The results of average rating are shown in Table 3. Each entry shows the appropriateness of the retrieved articles from the user point of view.

TABLE 3: RATING OF 7 SCHOLARS ON 4 LISTING USING TWO DIFFERENT SEARCHES

	Users	1	2	3	4	5	6	7
CiteSeerX	X-Top 5	1.3	1.7	2.3	2.1	1.4	3.1	1.9
	X-Top 10	3.5	4.1	3.3	4.0	3.8	4.0	4.3
BK Approach	BK-Top 5	2.2	1.9	3.1	2.6	3.1	3.5	3.3
	BK-Top 10	4.6	4.5	5.3	5.9	5.3	6.0	5.7

Figure 6 represents the average precision of articles in the top-5 and top-10 for three queries (prefix X and BK stand for CiteSeerX and background knowledge methods). As shown, for both topmost ratings, a relatively improvement using the proposed filtering approach based on user's background knowledge is achieved. Figure 7 depicts the percentage of improvement for each scholar query. The minimum and average improvements are %46 and %71 respectively.

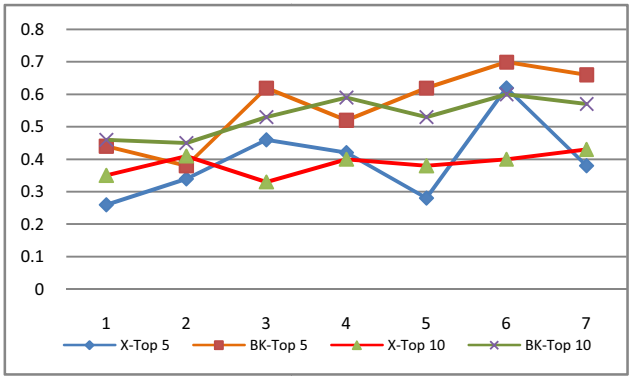

Figure 6: Average precision of top-5 and top-10 rating for three different queries of each scholar

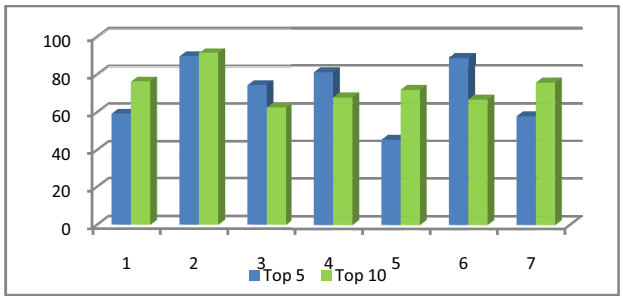

Figure 7: The percentage of improvement for each scholar

VII. CONCLUSION AND FURTHER WORKS

In this paper, a framework for incorporating scholar's background knowledge for digital libraries is proposed. The usability and importance of background knowledge in digital libraries is investigated. We propose the use of technical term frequencies as an ontological model for user's prior knowledge as well as ODP and WordNet for enriching the concepts in concept hierarchy. Employing ready to use sources of knowledge indicates a pleasant method to collect user prior knowledge with no direct user interaction with the system. The hierarchal structure of background knowledge facilitates the ease of concept similarity calculation as well ontology updating. The simple organization of CH assists to keep the ontology current and updated over the time. In order to capture the shift in the user's knowledge over the time, the weights in the hierarchy are reconsidered periodically. An offline experiment using CiteSeerX over a small number of scholars presents reasonable 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 as well as homepages information which contains lecture notes, publications, and project descriptions. Besides, for each part of the input articles, an appropriate weight is being considered in order to reflect the importance degree of each part and qualification of underlying concepts. However, the concepts from different sources of knowledge are not in the same granularity and therefore merging ontologies is a challenging task.

REFERENCES

[1] G. Uchyigit, "Semantically Enhanced Web Personalization," *Web Mining Appl. in E-Commerce & E-Services, SCI 172*, Berlin Heidelberg: Springer-Verlag, 2009, pp. 25-43.

[2] A. Sieg, B. Mobasher, and R. Burke, "Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search," *Work*, vol. 8, 2007, pp. 1-12.

[3] H. Drachsler, H.G.K. Hummel, and R. Koper, "Recommendations for learners are different : Applying memory-based recommender system techniques to lifelong learning," *Proceedings of the 1st Workshop on Social Information Retrieval for Technology-Enhanced Learning & Exchange*, 2007, pp. 18-26.

[4] G. Adomavicius, N. Manouselis, and Y. Kwon, *Multi-Criteria Recommender Systems*, University of Minnesota, MN 55455, USA: 2010.

[5] S.Y. Chen and R. Macredie, "Web-based interaction: A review of three important human factors," *International Journal of Information Management*, vol. 30, Oct. 2010, pp. 379-387.

[6] S. McDonald and R.J. Stevenson, "Effects of Text Structure and Prior Knowledge of the Learner on Navigation in Hypertext," *Human Factors*, 1998, pp. 18-27.

[7] H.J.A. Biemans and R.-J.P. Simons, "Contact-2 : a computer-assisted instructional strategy for promoting conceptual change," *Instructional Science*, vol. 24, 1996, pp. 157-176.

[8] S. Gauch, M. Speretta, and A. Pretschner, "Ontology-Based User Profies for Personalized Search," *Search*, 2005.

[9] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-Based Personalized Search and Browsing," *UMUAI*, Web Intelligence and Agent Systems 1(4), 2004, p. 219–234.

[10] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, "Web Usage Mining as a Tool for Personalization : A Survey," *User Modeling and User-Adapted Interaction*, 2003.

[11] B. Mobasher, "Recommender Systems," *Kunstliche Intelligenz, Special Issue on Web Mining*, BottcherIT Verlag, Bremen, Germany: 2007, pp. 41-43.

[12] N. Strangman and T. Hall, *Background Knowledge*, 2004.

[13] G.R. Lopes and M.A. Martins, "A Personalized Recommender System for Digital Libraries," *ACM Transactions on Information Systems*, 2007, pp. 59-66.

[14] S. Das, C.L. Giles, and P. Mitra, "On Identifying Academic Homepages for Digital Libraries," *JCDL'11*, Ontario, Canada: ACM, 2011.

[15] A. Sieg, B. Mobasher, and R. Burke, "Ontological User Profiles for Personalized Web Search," *American Association for Artificial Intelligence (AAAI)*, 2007, pp. 84-91.

[16] S. Gauch, M. Speretta, and A. Pretschner, "Ontology-Based User Profiles for Personalized Search," *Ontologies, Integrated Series in Information Systems*, Springer US, 2007, pp. 665-694.

[17] L. Zhou, "Ontology Learning: state of the art and open issues," *Springer*, vol. 8, 2007, pp. 241-252.

[18] L. Henderson, "Automated Text Classification in the DMOZ Hierarchy," 2009, pp. 1-25.

[19] P.D. Turney and P. Pantel, "From Frequency to Meaning : Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, vol. 37, 2010, pp. 141-188.

[20] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, 1975, pp. 613-620.

[21] C. D.Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, England: Cambridge University Press, 2009.

[22] B. Amini, R. Ibrahim, and M.S. Othman, "Discovering the Impact of Knowledge in Recommender Systems: A Comparative Study," *Journal of Computer Science*, vol. 2, 2011, pp. 1-14.

[23] A. Hotho, S. Staab, and G. Stumme, "Wordnet improves Text Document Clustering," *In Proc. of the SIGIR 2003 Semantic Web Workshop*, 2003, pp. 541-544.

[24] S. Scott and S. Matwin, "Feature Engineering for Text Classification," *Proceedings of ICML-99, 16th International Conference on Machine Learning*, 1999, pp. 1-13.

[25] G. Wei, M. Bao, and S. Wu, "Research on Ontology-Based Text Representation of Vector Space Model," *IEEE*, 2010.

[26] R. Burke, "Knowledge-based recommender systems," *Library*, 1999, pp. 1-23.

[27] M. Degemmis, P. Lops, and G. Semeraro, "Learning Semantic User Profiles from Text," *Springer Berlin Heidelberg*, vol. LNAI 4093, 2006, pp. 661-672.

[28] D. Tsatsou, F. Menemenis, and P.C. Davis, "A Semantic Framework for Personalized Ad Recommendation based on Advanced Textual Analysis," *ACM RecSys'09*, NY, USA: ACM, 2009, pp. 217-220.

[29] S.S. Anand and B. Mobasher, "Intelligent Techniques for Web Personalization," *Springer Berlin Heidelberg*, vol. 3169, 2005, pp. 1-36.

[30] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User Profiles for Personalized Information Access," *Springer Berlin Heidelberg*, 2007, pp. 54 - 89.

[31] M.J. Pazzani and D. Billsus, "Content-Based Recommendation Systems," *The Adaptive Web 2007, LNCS 4321*, A.W.N. (Eds) P. Brusilovsky, A. Kobsa, ed., Berlin Heidelberg: Springer-Verlag, 2007, pp. 325 - 341.

[32] I.I. Retrieval and T. Mining, "Information Retrieval and Text Mining," *Fundamentals of Predictive Text Mining, Texts in Computer Science*, S.M.Weiss Et Al., eds., London: Springer London, 2010, pp. 75-90.

[33] G. Adomavicius, A. Tuzhilin, and S. Sen, "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Transactions on Information Systems*, vol. 23, 2005, pp. 103-145.

[34] A. Kayed, E. El-Qawasmeh, and Z. Qawaqneh, "Ranking web sites using domain ontology concepts," *Elsevier, infomation Management*, vol. 47, Dec. 2010, pp. 350-355.

[35] M. Degemmis and P. Lops, "A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation," *User Modeling and User-Adapted Interaction*, 2007, pp. 217-255.

[36] A. Budanitsky and G. Hirst, "Semantic distance in WordNet : An experimental , application-oriented evaluation of five measures," *Evaluation*, 1998.

[37] T.H. Duong, M.N. Uddin, D. Li, and G.S. Jo, "A Collaborative Ontology-Based User Profiles System," *System*, 2009, pp. 540-552.

[38] S.-C. Liao, K.-F. Kao, I.-E. Liao, H.-L. Chen, and S.-O. Huang, "PORE: a personal ontology recommender system for digital libraries," *The Electronic Library*, vol. 27, 2009, pp. 496-508.

[39] S.-Y. Yang, "Developing an ontology-supported information integration and recommendation system for scholars," *Expert Systems with Applications*, vol. 37, Oct. 2010, pp. 7065-7079.