2015 IEEE 15th International Conference on Data Mining Workshops

Research Paper Recommendation Based on the Knowledge Gap

Weidong Zhao
Shanghai Key Laboratory of Data Science
School of Software, Fudan University
Shanghai 200433, China
wdzhao@fudan.edu.cn

Ran Wu
Shanghai Key Laboratory of Data Science
School of Software, Fudan University
Shanghai 200433, China
13212010022@fudan.edu.cn

Weihui Dai
School of Management
Fudan University
Shanghai 200433, China
whdai@fudan.edu.cn

Yonghui Dai
School of Information Management and Engineering
Shanghai University of Finance and Economics
Shanghai 200433, China
dyh822@163.com

***Abstract*—The massively growing of literature resource makes it a challenge for researchers to find useful papers. To solve the information overload problem, some researches on personalized paper recommendation have been conducted. However, the knowledge gap between a researcher's background knowledge and research target is seldom concerned. In this paper, we propose a knowledge-gap based literature recommendation method to support researchers to fulfill literature support. Firstly, domain knowledge is modeled with the concept map, based on which background knowledge and target knowledge are analyzed. Then, knowledge gap is defined both graphically and intuitively with the knowledge map. To bridge the knowledge gap, we design a graph-based method to explore some suitable knowledge paths, which can help a researcher to learn the requisite knowledge in accordance with cognition pattern. Finally, experiments are performed to demonstrate the effectiveness of the proposed method.**

Keywords-literature recommendation; knowledge gap; concept map; background knowledge

I. INTRODUCTION

Along with the development of science and technology, great achievements have been made as literatures. However, the increasing number of research papers makes it a challenge for scientific researchers to discover helpful knowledge resources. It is commonly called the information overload problem [1, 2]. To deal with this problem, some information techniques (e.g. information retrieval and recommendation) are adopted by E-libraries and scientific database to support knowledge workers. Research paper recommendation is the method that recommends helpful papers to researchers by exploring their interests with information techniques [3]. By actively providing interesting research materials, literature recommendation can save researchers much valuable time and efforts [4].

Content-based filtering and collaborative filtering are the most widely used recommendation techniques in many contexts, including E-commerce platforms, music website, etc. To recommend papers, content-based filtering builds a user profile based on his/her reading history and recommends new papers that well match the profile. In previous works, user profile is generally built on key words. However, it is insufficient to model the user's preference. To refine the preference model, several models have been proposed, such as label-enriched model [5], ontology-expansion method [6]. Collaborative filtering method first discovers like-minded researchers and recommends papers based on their interests. The key issue in collaborative filtering is to measure user similarity. Several methods have been proposed to analyze the relationship between researchers from various perspectives, such as e-mail communication, co-author relationship, reference relationship, project cooperation and interaction on social media [7, 10].

Distinguished from usual information resources (e.g. news and tweet), academic literature is an important knowledge resource for a researcher. In other words, research papers are crucial learning materials. Researchers gain their background knowledge from previous reading of literature. Before navigating an academic database, a researcher usually has a clear knowledge goal, which is commonly documented as research proposal, requirement specification or planning statement. The disparity between the knowledge goal and his/her background knowledge is called "knowledge gap". To narrow the gap, a researcher retrieves literature databases and discovers papers he/she need. By learning and digesting these papers, he/she transforms the embedded knowledge into his/her own. However, the knowledge gap issue is seldom concerned in previous works. Since a user's historical preference cannot precisely reveal his/her current knowledge requirements, previous methods hardly narrowed the knowledge gap. As a result, the excessive number of literatures makes it difficult for researches to meet their requirements even they have clear knowledge retrieval purposes. Therefore, continuous

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.40

efforts are needed to support knowledge workers by bridging the knowledge gap.

In this paper, we present an approach called knowledge gap based recommendation (KGR) to bridge the knowledge gap between a researcher's background knowledge and research target. The method presents domain knowledge in the form of concept maps. First, a set of central concepts are extracted from domain corpus. Then, the strategy builds links between these concepts according to their association relation. A researcher's reading records are analyzed to model his/her background knowledge; and target knowledge is extracted from their research proposals. In the domain concept map, the knowledge gap is defined as the shortest paths that connect these two kinds of knowledge. Finally, concepts in those paths are utilized to discover well-matched papers, which can help bridge the user's knowledge gap.

The paper is structured as follows. Section 2 summarizes recent works related to literature recommendation. In Section 3, the method to build the domain concept map is introduced. Section 4 describes the method to define the knowledge gap, and illustrates the recommendation algorithm based on the gap. Section 5 gives a brief case study and Section 6 evaluates the proposed method. Section 7 draws conclusions.

II. RELATED WORKS

Many researchers realize that user preference cannot be the only guidance when recommending resources. Therefore, they consider other factors, such as domain knowledge, user background, learning targets and cognitive patterns. The context of literature reading is not well-formed, making it difficult to determine the user's learning targets and cognitive patterns. As a result, previous researches can only work in standardized teaching situations, like e-learning. For example, Zhang et al. organized disciplines and curriculum information as a knowledge tree, from which association rules between resources and courses were mined to recommend teaching resources [12]. Tang and McCalla proposed a paper recommendation method focusing on teaching characteristics, including the user's knowledge level and knowledge goals [16]. Based on these characteristics, a set of ordered papers are recommended. This paper focuses on filling the knowledge gap between the user's background knowledge and research targets.

In some researches, domain knowledge was modeled in the form of a domain taxonomy, ontology or concept network. Liang et al. established a semantic network according to visited documents. The semantic network is composed of several connected semantic trees, and connections in each semantic tree reflect the inheritance relationship between concepts. Then spreading activation was used to semantically expand the user's interest [9]. Spreading activation can achieve beneficial knowledge expansion, enriching original knowledge model with closely related knowledge. Xu et al. combined concept networks with social networks, and recommended experts to users with comprehensive utilization of semantic relations between concepts, social cooperation between experts and professional relationship between concepts and experts [10]. Cantador and Castells established a semantic network composed of domain concepts, and user interests took the form of concepts and user interest levels. The algorithm expanded user interests by spreading activation, then clustered concepts in semantic networks according to user preferences. Users with similar concept clusters are considered to have similar interests [11].

Some approaches extracted and utilized the user's background to build the personalized user profile, which reflected unique background and requirements of each user. Chen et al. built an adaptive ontology for each user according to the user's browsing and reading behavior. User-concept, user-user patterns were mined from the ontology, and resources were recommended to the user according to similar patterns in pattern library [32]. Hawalah and Fasli set up a personalized interest ontology based on user interest and view. Spreading activation was applied to expand user interest, aiming to find relevant concepts user might be interested in [13]. Some recommendation methods planned a learning path for the user on the basis of dependent relationships between resources, so as to help the user achieve a certain goal. Yu et al. created three ontologies representing the learner's background knowledge, learning resources and domain knowledge respectively. Similar learning resources were found after computing similarity between user background and learning resource, then these resources were organized as a learning path by their prior relationships [14]. Durand et al. thought the learner's knowledge base needed to satisfy the prerequisite when learning resources. Only in this way could the learner learn the resource and gain the knowledge it offered. On that basis, the strategy introduced in [15] established a directed graph according to competencies required and offered by learning objects, and recommended a sequence of learning objects in a well-defined order to the learner. Ordered learning paths could help users to reach the goal under the condition of their competencies, so that users can follow some learning paths from the initial set to the targeted set of competencies.

This paper establishes a domain knowledge model--concept map, and the user's knowledge gap is determined utilizing the domain background knowledge. To determine a user's knowledge gap, background knowledge is extracted from his/her reading records and the target knowledge is extracted from a research proposal. Generally, a researcher will write a proposal that indicates the future and goal of further research before conducting research. Herein, the basic premise is that research proposals can be provided.

III. DOMAIN CONCEPT MAP CONSTRUCTION

There are a variety of methods to build a structured and formalized abstraction of domain knowledge. The usual way is to extract key words with high frequency as knowledge factors and analyze their relationships. Hierarchical relations, which reflect generalization relation between concepts, are utilized to build models like taxonomy and hierarchical structure [22,23,24]. However, word frequency cannot precisely indicate the semantics of a paper.

In this paper, we take semantics of documents into account. The method takes domain corpus as input and

outputs the concept map which represents the core knowledge structure of a domain. The concept map is a graphical representation of both knowledge and thinking visualization [18,19]. From the concept map, we can get the logical relationship between knowledge nodes [20].

The concept map is represented as an undirected weighted graph $G=\{V,E,CW\}$. Each node $v \in V$ is a knowledge factor. Each edge $e \in E \subseteq \{V \times V\}$ represents the relation between two factors. $CW \in [0,1]$ is the weight of edges, which reflects the strength of the relationship. The stronger the relation is, the greater the weight is.

The basic assumption of traditional feature extraction methods, such as TF-IDF, is that the most significant words are those of high frequency. Since it does not consider the semantic information in a document, the problem of synonym/polysemy cannot be solved. LDA (Latent Dirichlet Allocation), a popular probabilistic topic model, assumes that all documents in the corpus share a number of implicit themes[21]. Every document is modeled as a mixture of words of implicit themes with a certain probability. We employ LDA to extract implicit topics that can reflect the themes in a domain. It can help solve the synonym/polysemy problem. Each topic can be represented as some related words. A knowledge node in the concept map represents a topic in the domain, and edges are built to reflect the correlation between topics.

A. Model Training

In LDA, each document is seen as a probability distribution of some topics, and each topic is a probability distribution of some words. By learning "document-topic" distribution θ and "topic-word" distribution φ, we can know the proportion of topics in each paper. After segmentation, removing stop-words and stemming, the domain corpus D will be represented as a set of unique words $S=\{k_1, k_2,\cdots, k_K\}$. Assuming D contains TN topics $T=\{t_1, t_2,\cdots, t_{TN}\}$, the parameters θ, φ obtained after training is shown as :

$$\theta_{d,t} = \frac{n_d^{(t)} + \alpha_t}{\sum_{t=1}^{TN} n_d^{(t)} + \alpha_t} \tag{1}$$

$$\varphi_{t,k} = \frac{n_t^{(k)} + \beta_k}{\sum_{k=1}^{KN} n_t^{(k)} + \beta_k} \tag{2}$$

where $n_d^{(t)}$ is the effective number of words belonging to the topic t in the paper d, and $n_t^{(k)}$ is the number of words which belong to t. α_t and β_k are constant.

Since an academic paper may contain various topics with different probability. We say that a document belongs to a topic only when the probability is large enough. The membership relation is defined as follows:

Definition 1 Define d_i belongs to t_j only when $\theta_{d_i,k_j} \geq \varphi_\alpha$, where θ_{d_i,k_j} is the probability that d_i corresponds to t_j, and φ_α is the threshold. For a new paper d that is not in the corpus, trained parameter φ can be used to calculate the probability that d belongs to t as (3).

$$w(d,t) = \frac{\sum_{k \in d} \varphi_{t,k}}{\sum_{t \in T} \sum_{k \in d} \varphi_{t,k}} \tag{3}$$

.

B. Construction of Concept Map

Each node in the concept map represents a topic in domain paper set. An edge is defined as correlation relationship, which implies links in certain knowledge dimension. Bose et al. proposed that concepts are associated in certain dimension when they appeared in the same document[25]. If two different topics appear in the same paper, it means that they have certain relevance in some dimensions. If two different topics appear in the same paper for many times, they are considered to have correlations.

Definition 2 If a paper d_i belongs to the topic t_i and t_j at the same time, then t_i and t_j co-occur for one time. The weight CW is defined according to times of co-occurrence.

Definition 3 The weight of the edge between t_i and t_j is defined as (4).

$$CW(t_i,t_j) = \frac{|d^{t_i} \cap d^{t_j}|}{|d^{t_i}| \times |d^{t_j}|} \tag{4}$$

Where d^{t_i} is the paper set that belongs to t_i, and $|d^{t_i}|$ is the corresponding size. $d^{t_i} \cap d^{t_j}$ represents the paper set that belongs to t_i and t_j simultaneously.

The main steps to construct a concept map is as follows:

1) The method creates a node in the concept map for each topic t_i in $T=\{t_1,t_2,\ldots,t_T\}$;
2) For each topic $t_i \in T$, if the co-occurrence times of t_i and t_j is greater than threshold ω, then an edge is created between them.
3) The weight CW of each edge is calculated with (4).

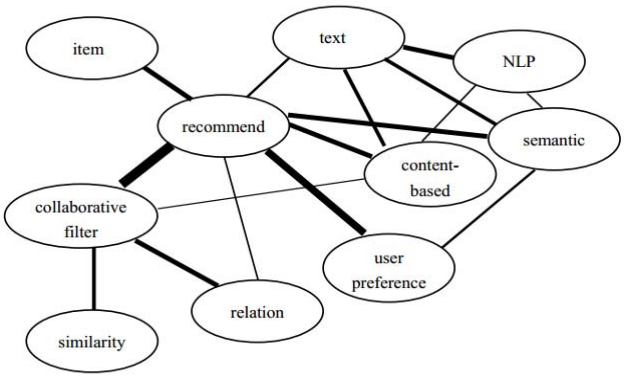

Figure 1. An example of the concept map

Fig. 1 gives an example of the concept map. An ellipse corresponds to a topic in domain corpus, and a line between

two ellipses represents the correlation of two corresponding topics. The width of a line stands for the *CW* weight of this correlation. The larger the *CW* weight of a correlation is, the wider the corresponding line is.

IV. KNOWLEDGE GAP ANALYSIS

In a formal learning scenario such as e-learning, learning is a highly structured process, and there exist specific logical links among learning materials. A researcher's learning goal may be clearly defined, and a study path to achieve this goal can be found according to logical links among those materials. For example, Rosson and Carroll introduced a scenario-based method [27] to define user requirements. Zhang et al used pattern mining to predict the user's goal[12]. Different from formal learning, reading and attending conferences are all informal learning process for researchers. It is hard to give a clear definition of the researcher's learning goal and knowledge needs. Therefore, this paper analyzes the user's background knowledge, research goals and knowledge gap on the basis of domain concept map.

Knowledge is obtained by observing and cognizing new things through existing concepts according to Ausubel's Meaningful Learning Theory. Learning is a process to set up a concept network and constantly add new concepts to it[28]. A learner need relate new knowledge to his/her own background knowledge while learning, only in this way can he/she achieve meaningful learning results. When we want to solve a problem, we need to constantly activate interrelated knowledge in our background knowledge to explore solutions [23]. Ferrari and Gnesi built a concept map to represent the user's background knowledge according to papers read by the user, and simulated the process of understanding natural language by finding least-cost paths on the map [26]. Relevance of knowledge is an important factor to guarantee the effectiveness of study. Some algorithms organized an ordered learning path to recommend learning objects, which considers the characteristics of learning objects and the user's background knowledge [14,15].

In this paper, knowledge gap is determined by relevance between knowledge. According to Ausubel's theory, meaningful learning is achieved only when new learned knowledge has connections with the learner's background knowledge. In order to achieve the research goal, the user needs to add new knowledge to his/her background knowledge base constantly, and to build the connection between new added knowledge and background knowledge. After all research-related knowledge becomes a part of background knowledge, the user would have necessary knowledge to complete his/her research. If the user cannot include necessary research goal knowledge into his/her background knowledge, there will exist a knowledge gap between background knowledge and research goal knowledge. The user cannot have comprehensive understanding of research goal with the knowledge gap, making it hard to finish the target research. In this section, the user's reading records and research proposal are analyzed to depict background knowledge and research goal knowledge. In order to help the user supplement all necessary knowledge into background knowledge, research goal knowledge extracted is extended using spreading activation on the basis of domain concept map. The extended research goal knowledge will contain knowledge closely related to research goal. Then knowledge paths are searched that link two kinds of knowledge, so that the user can learn all research knowledge along these paths. These knowledge paths represent the user's knowledge gap. Finally, research papers that contain knowledge the user lacks are recommended according to the knowledge paths.

A. Knowledge Model

The knowledge model of a user consists of his/her background knowledge and research goal knowledge. In this section, user's reading records are utilized to model his/her background knowledge, and some research proposals are used to extract the research target knowledge.

1) Background knowledge

The assumption here is that a user can gain knowledge by reading and digesting literatures. By referring to some related literature, a user will improve his/her knowledge level on certain knowledge topics. To model a user's background knowledge, his/her historical reading records are analyzed by the LDA algorithm. In this way, a user's knowledge level on each knowledge topic is properly measured.

Define expertise as a user's knowledge level on a certain topic. A user's expertise will improve as he/she read more related papers**.**

Definition 4 Let the literature set that the user *u* has read be $UD=\{ud_1, ud_2,\ldots, ud_m\}$. Define the expertise level that *u* has on a topic *t* as (5)

$$Expertise(u,t) = \sum_{i=1}^{m} w(ud_i,t) \tag{5}$$

where *m* is the number of papers the user has read, and $w(d_i,t)$ is calculated by (3). That is, expertise is the accumulation of the user's reading history. With the threshold φ_b, we extract the topic set $UT = \{t \mid Expertise(u,t) > \phi_t, t \in T\}$ as the background knowledge of *u*, denoted as $UT=\{ut_1, ut_2,\ldots, ut_n\}$.

2) Research Target Knowledge

The proposal a user provides introduces the background, direction and outline of his/her research. However, the description about future research in this single proposal is fairly general, thus it cannot reflect knowledge demand of the researcher. LDA is utilized to extract topics in the proposal, the topic set $GT=\{\ importance(u,t) > \varphi_i\ ,\ t \in T\}$ stands for the target knowledge. Then the target knowledge is extended to determine comprehensive knowledge demand. Spreading activation is utilized to extend research goal knowledge to other closely related knowledge. Research goal knowledge is extended along links in domain concept map. A link in domain concept map indicates the relevance between two knowledge factors in some dimension. This multi-dimensional knowledge expansion improves comprehensiveness of research goal knowledge.

The spreading activation model is an extension of content-based filtering, and it can reach the nodes that are highly associated with the initial nodes through network links [9]. As an iterative process, spreading activation starts from some original nodes in the network and activates other nodes that link to original nodes. The activated nodes begin to activate other connected ones in the same manner. This process will not stop until some conditions are satisfied. The activation value, threshold and the maximum spreading distance are the essential parameters in spreading activation model. The activation value reflects working condition of activated nodes. Spreading distance is the number of links that are spread. The threshold and the maximum spreading distance are the values for the activation process to stop. That is, if the activation value of a node is lower than the threshold, or the spreading distance of this node has reached the maximum diffusion distance, then the activation process will stop. The spreading activation process is executed as follows.

Definition 5 The Importance of t to u is defined as (6).

$$importance(u,t)=w(rd,t) \quad (6)$$

That is the weight of t in the research proposal rd. The importance weight of all topics in research goal knowledge are defined as $TW=\{tw_1, tw_2,\ldots, tw_q\}$, where $tw_i=importance(u, t_i)$.

The spreading activation process is executed as follows.

1) *Initialization*. The topic set in *GT* is chosen as initial activation nodes, with weight *TW* as initial activation value. Activation values of other nodes are set 0. The threshold λ and maximum spreading distance *sd* are initialized.

2) *Spreading activation process*. Nodes are activated according to the connections in domain concept map starting from initial activation nodes. When a node is activated via a connection by another node, the activation value is product of the activation value and *CW* weight of the connection. For example, if a node whose activation value is 4 activates a node through a connection with *CW* weight 0.3, then the activation value of newly activated node is 1.2. If a node is activated by many nodes, then the activation value is the sum of all activation values from all activating nodes.

3) *Termination condition*. Every time activation is completed, the termination condition will be checked. If the activation value of a node is below the threshold or spreading distance reaches the maximum spreading distance, the spreading activation will stop. The extended topic set is *EGT*.

Node selection. For each topic t in *EGT*, delete the topics of which $importance(u,t)<\varphi_t$. The document set obtained $CRT=\{crt_1,crt_2,\cdots,crt_p\}$ represents target knowledge of the user.

B. Analysis of Knowledge Gap

In this section some knowledge paths that connect research goal and background knowledge will be searched in domain concept map. Topic nodes in these paths can help the user learn related knowledge starting with their own knowledge until they add knowledge about research goal into their own background knowledge. The paths represent the knowledge that the user needs before grasping research target knowledge, namely knowledge gap of the user.

Usually a researcher's research goal is not a part of his/her background knowledge, which means that research target knowledge does not coincide with the user's background knowledge completely and there exists no connection between these two kinds of knowledge. New knowledge is needed so that concept paths can be found between the two kinds of knowledge that can help achieve the research goal. From the perspective of the concept map, the user needs concept paths that link the background knowledge and target knowledge, which are concept links. There exist lots of concept paths between a user's background knowledge and his/her research goal, and target knowledge can be learned through any path. For example, a user has t_2 as background knowledge and t_7 as research target knowledge, and there exist concept paths $\langle t_2,t_3,t_7\rangle$,$\langle t_2,t_1,t_3,t_7\rangle$,$\langle t_2,t_1,t_4,t_5,t_7\rangle$ in G ($t\in$ T). Different concept paths contain different concepts, thus the number of concepts that the user needs to learn is also different. In order to help improve learning efficiency, the shortest paths are selected to determine the knowledge gap.

The concept set *UT* and *CRT* are extracted from the literature the user has read and research plan to represent background knowledge and target knowledge. There will be overlapping concept nodes between the user's background knowledge and target knowledge as the research is conducted. And real research target knowledge now is $TT=CRT-UT\cap CRT$.

Let the background knowledge be $UT=\{ut_1,ut_2,\cdots,ut_n\}$ ($t\in T$)and research target knowledge be $TT=\{tt_1,tt_2,\cdots,tt_k\}$ ($t\in T$) Starting from research target knowledge, the shortest path to *UT* is searched using the Dijkstra algorithm for each concept in TT. The distance between concept nodes is defined as reciprocal of the weight of correlation relationship $CW(t_a,t_b)$, which is $DL(t_a,t_b)=1/CW(t_a,t_b)$. This means that the closer two concepts are, the shorter the distance between them is. After k shortest paths are found, all the paths represent the knowledge gap of the user.

Definition 6 The priority weight *GW* measures the importance of t in the path p

$$GW(t, path) = \frac{Position(t, path)}{\sum_{ti\in path} Position(ti, path)} \quad (7)$$

Where $Position(t, path)$ is the position of t in the *path*. The closer a topic node is to background knowledge in a knowledge path, the more contribution it makes to learning new knowledge. On the contrary, the topics closer to target knowledge are more important for bridging knowledge gap. For example, there is a knowledge path $p=\langle k_1,k_2,k_3,k_4\rangle$ with k_4 being target knowledge of the user. The position of each node on the path is (1,2,3,4)and the priority weight is (1/10,2/10,3/10,4/10) respectively. As we can see, nodes that are closer to the target knowledge have higher weights.

Suppose a user u has $UT=\{ut_1,ut_2,ut_3\}$ as background knowledge and $TT=\{tt_1,tt_2,tt_3,tt_4\}$ as target knowledge. According to the knowledge gap analysis method, paths linking background knowledge and research target knowledge are $p_1=\langle tt_1,t_i,t_k,ut_3\rangle$, $p_2=\langle tt_2,t_j,t_k,ut_2\rangle$, $p_3=\langle tt_3,tt_1,t_l,ut_2\rangle$, $p_4=\langle tt_4,t_k,t_j,t_1,ut_3\rangle$, thus the knowledge gap of u is $KG=\{p_1,p_2,p_3,p_4\}$.

C. Recommendation Based on Knowledge Gap

Since a concept may have different meanings in different phrases, the combination of concepts can reflect knowledge accurately. For example, network could be social network, neural network or road network. Thus we take a path of concepts as a user's knowledge gap rather than keywords. According to a user's knowledge gap, papers are recommended to narrow the gap by fulfilling knowledge the user lacks. After determining the knowledge gap KG, concept paths in KG are utilized to as keywords to find papers that match concept paths. We define a matching score MS to measure the matching degree between a paper d and a concept path in (8).

$$MS(d,path)=\sum_{t\in path} w(d,t)\times GW(t,path) \quad (8)$$

Where $w(d,t)$ is the probability that d belongs to topic t, and $GW(t,path)$ is the priority weight of t in $path$. The matching score will increase if the topic distribution in a paper can fulfill the knowledge gap(especially target knowledge).

In order to study all target knowledge, all knowledge paths in KG should be matched. The benefit of a paper is defined as (9).

$$Benefit(d,KG)=\sum_{p\in KG} MS(d,p)\times importance(u,tp) \quad (9)$$

Where tp is target knowledge in p, $importance(u,tp)$ is the importance of tp for the user. During the recommendation process, (9) is used to calculate each document's utility value, and k papers with the highest utility are recommended to the user.

Algorithm 1 shows the process of knowledge recommendation based on gap analysis. Domain knowledge graph G, research proposal T, paper set B the user read and domain document set D as input. The algorithm is divided into three steps: background knowledge analysis, target extraction and gap knowledge recommendation. At the stage of background knowledge analysis, the papers the user has read are preprocessed first. Then, according to definition 4 topics whose expertise is higher than the threshold are chosen as user background knowledge. During target knowledge extraction, target knowledge and user' demand for each topic is determined by (3), then target knowledge nodes are spread in the concept map G. After removing nodes with lower importance, the topic set TT represents the target knowledge. Knowledge recommendation phase finds the shortest paths from target knowledge to background knowledge using the Dijkstra algorithm. And then papers that can fulfill knowledge gap are chosen, k papers with the highest matching score are recommended.

Algorithm 1. Recommending papers

Input: concept map G, research proposal P, referenced documents B, domain document set D
Output: literature list R
Procedure:

```
Background Knowledge UT, Target Knowledge GT, Knowledge Gap KG;Topics T;
  // background knowledge evaluation
  preprocess(B);
  for topic in T:
      UT(topic)=Expertise(u,topic)
  remove(UT, φ_t )
 //remove less expertised knowledge topic
 // target knowledge extraction
 preprocess(P);
 for topic in T:
     GT(topic)=importance(u,topic)
 TT=SpreadingActivation(G,GT)
 remove(GT, φ_i )
 //remove less important knowledge topic
 // gap knowledge recommendation
 for ts in TT:
         path(ts)=Min{Dijkstra(G,tb,ts) | tb∈UT}
     // explore concept learning path
     KG.add(path(ts))
 for document in D:
     Match(document)=Benefit(d,KG);
 return top k document
```

V. EXPERIMENTS

In this section, experiments are conducted to compare the KGR with other recommendation methods in different indicators.

40 postgraduate students taking the "recommender systems" course from Software School in a University are recruited as subjects. The data set is preprocessed first. The title, abstract, key words and body of papers are chosen as experiment data. Gibbs sampling algorithm is utilized to sample topic-word probability on the basis of topics given by domain experts.

40 subjects search and read some papers related to "recommend" and "personalized recommendation" freely, and they need to record and submit the papers they have read. After each subject have read the same number of papers (20), they are given the same research proposal about recommendation and asked to read it. Based on their reading records and research proposals, three recommendation methods recommend 5, 10, 15, 20 papers to each subject respectively including keyword matching (KMR, for short), semantic expansion based(SER, for short) and knowledge gap based(KGR, for short). The subjects are required to read all recommended papers. After that all subjects need to mark each recommended paper according to the measure "whether recommended papers can help achieve research target in the proposal" (if yes marked as true, else false). Fig. 2 shows the result of three methods with 5, 10, 15, 20 recommended papers respectively.

It can be seen that with increase of the number of recommended papers, precision of three methods all go down. However, KGR holds the highest accuracy with more

stable trend, which indicates it outperforms other methods. Precision of SER is a little better than KMR when the number of recommended papers is smaller. But as the number increases, this advantage is gradually narrowing until surpassed. This is because SER could produce concepts closely related to background knowledge when recommending less paper. Thus, it can gain higher precision. Yet, as the number increases, it would include some less related concepts, which cause worse precision. As you can see, KMR is more stable compared with SER.

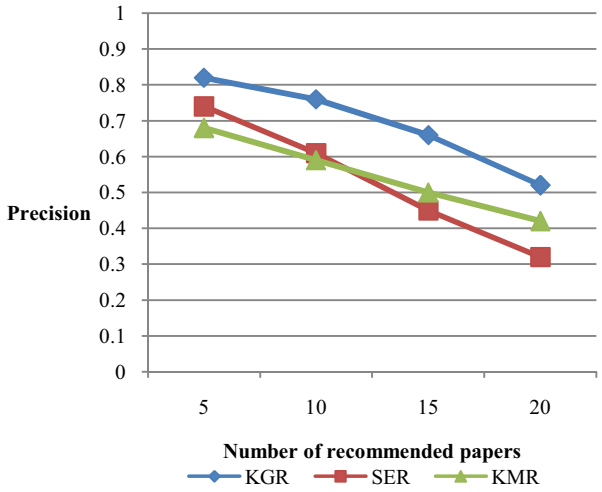

Figure 2. Comparison of precision

ROC curve (receiver operating characteristic curve), which can be used to evaluate recommendation methods, consists of true positive rate (TPR) and false positive rate (FPR). Diagonal from (0, 0) to (1, 1) represents the random prediction, the point above the diagonal indicates better classification result (better than random prediction), and the point below indicates worse result (worse than random prediction).

Fig.3 is ROC of three methods when the number recommended papers is 15. As it can be seen from Fig.3, KGR has larger AUC compared to KMR and SER, which means KG has the most accurate prediction compared to the other methods. KGR has the highest true positive rate at the same false positive rate, which indicates that papers recommended by KGR have higher evaluations from subjects compared to the other methods.

All experiment subjects are asked to search and read papers related to "recommend" and "personalization" freely. This work is divided into five stages according to the number of papers subjects have read. After each stage ends, three recommendation methods recommend 10 papers to each subject respectively. Then all subjects read all recommended papers and give satisfaction scores in range of 1 to 5 to every paper according to evaluation standards in TABLE 1. The Standards reflect the subject's satisfaction level to a paper, and higher score indicates that the paper can provide more valuable knowledge for research.

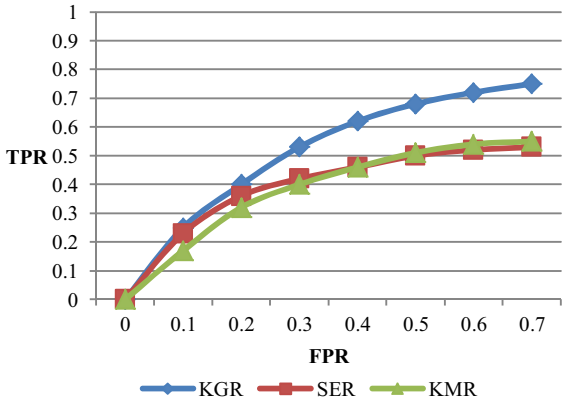

Figure 3. Comparison of ROC

TABLE 1. EVALUATION STANDARDS

No.	Standards	Score
1	Does the paper provide the knowledge you don't have but help for research?	1-5
2	Does the paper help deepen or expand your understanding of research?	1-5
3	Does the paper help yield new insights about existing knowledge, like introducing new features, usage and meaning?	1-5
4	Does the paper help connect multiple knowledge concepts?	1-5
5	Does the paper help inspire new ideas for research?	1-5

Average scores are calculated after all subjects grade recommended papers. Fig.4 displays user satisfaction as the number of papers they read changes. It is clearly observed that satisfaction of three methods all declines as the number of papers increase. The reason may be that it is more difficult to find a valuable paper to recommend as a subject read more and more papers, and have more comprehensive background knowledge. SER and KMR show significant declines, whereas KGR remains relatively stable.

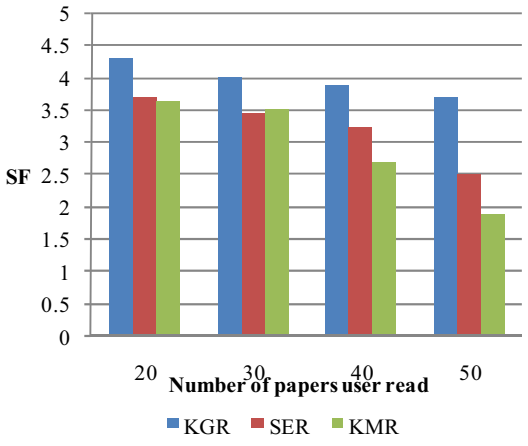

Figure 4. Comparison of user satisfaction.

VI. CONCLUSIONS

In this paper, we propose a method to recommend papers for researchers based on knowledge gap. Distinguished from previous methods, user preference is no longer the only factor for recommendation. In KGR, the user's background knowledge and his research targets are used to determine his knowledge gap. KGR aims to recommend papers that can help bridge knowledge gap. The method firstly builds a domain concept map based on domain corpus. Then the user's knowledge gap is thoroughly analyzed. The contributions of this paper are as follows: topics in a domain and their correlations are organized in the form of concept map, which can reflect domain background knowledge; research target and background knowledge of the user are used to determine the user's knowledge gap; recommendation aims to fulfill the knowledge gap and help achieve research goal.

Since the method utilizes research proposal and reading history to define the user's knowledge gap, major concern is how to use more data to find knowledge the user really lacks.

ACKNOWLEDGMENT

This paper was supported by Shanghai Pujiang Program no.14PJC017, National Nature and Science Foundation of China under Grants no.71071038 and no.91324010, and National High-tech R&D Program (863 Program) of China under Grants no.2012AA02A612. Many thanks to Mrs. Hongzhi Hu for her assistant work to Ran Wu and Weihui Dai who are the joint corresponding authors of this paper.

REFERENCES

[1] H. Drachsler, H. G. Hummel, and R. Koper, "Identifying the goal, user model and conditions of recommender systems for formal and informal learning," Journal of Digital Information, vol.10, no.2, 2009.

[2] M. Salehi and I. N. Kamalabadi, "Hybrid recommendation approach for learning material based on sequential pattern of the accessed material and the learner's preference tree," Knowledge-Based Systems, vol.48, pp.57-69, 2013.

[3] C. Basu, H. Hirsh, W. W. Cohen, and C. Nevill-Manning, "Technical paper recommendation: A study in combining multiple information sources," Journal of Artificial Intelligence Research, vol.1, pp.231-252, 2001.

[4] C. Pan and W. Li, "Research paper recommendation with topic analysis," Proceedings of Computer Design and Applications, V4-264, 2010.

[5] Z. Guan, C. Wang, and J. Bu et al, "Document recommendation in social tagging services," Proceedings of World Wide Web, pp.391-400, 2010.

[6] Z. Yu, Y. Nakamura, S. Jang, S. Kajita, and K. Mase, "Ontology-based semantic recommendation for context-aware e-learning," Proceedings of Ubiquitous Intelligence and Computing, pp.898-907, 2007.

[7] E. Davoodi, M. Afsharchi, and K. A. Kianmehr, "Social network-based approach to expert recommendation system," Proceedings of Hybrid Artificial Intelligent Systems, pp.91-102, 2012.

[8] T. Y. Tang and G. A. McCalla, "Multidimensional paper recommender: Experiments and evaluations," Internet Computing, IEEE, vol.13, no.4, pp.34-41, 2009.

[9] T. P. Liang, Y. F. Yang, D. N. Chen, and Y. C. Ku, "A semantic-expansion approach to personalized knowledge recommendation," Decision Support Systems, vol.45, no.3, pp. 401-412, 2008.

[10] Y. Xu, X. Guo, J. Hao, J. Ma, R. Y. Lau, and W. Xu, "Combining social network and semantic concept analysis for personalized academic researcher recommendation," Decision Support Systems, vol.54, no.1, pp.564-573, 2012.

[11] I. Cantador and P. Castells, "Multilayered semantic social network modeling by ontology-based user profiles clustering: Application to collaborative filtering," Proceedings of Managing Knowledge in a World of Networks, pp.334-349, 2006.

[12] H. Zhang, W. Ni, M. Zhao, Y. Liu, and Y. A. Yang, "A hybrid recommendation approach for network teaching resources based on knowledge-tree," Proceedings of Control Conference, pp.3450-3455, 2014.

[13] A. Hawalah and M. Fasli, "Using user personalized ontological profile to infer semantic knowledge for personalized recommendation," Proceedings of E-Commerce and Web Technologies, pp.282-295, 2011.

[14] Z. Yu, Y. Nakamura, S. Jang, S. Kajita, and K. Mase, "Ontology-based semantic recommendation for context-aware e-learning," Proceedings of Ubiquitous Intelligence and Computing, pp. 898-907, 2007.

[15] G. Durand, N. Belacel, and F. LaPlante, "Graph theory based model for learning path recommendation," Information Sciences, vol.251, pp.10-21, 2013

[16] T. Tang and G. McCalla, "Beyond learners' interest: personalized paper recommendation based on their pedagogical features for an e-learning system," Proceedings of Trends in Artificial Intelligence, pp.301-310, 2004

[17] D. P. Ausubel, The Psychology of Meaningful Verbal Learning, New York: Grune & Stratton, 1963.

[18] M. Willerman and R. A. Mac Harg, "The concept map as an advance organizer," Journal of Research in Science Teaching, vol.28, no.8, pp.705-711, 1991.

[19] J. D. Novak, Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations. Routledge. 2010.

[20] F. Lehmann, Semantic Networks in Artificial Intelligence. Elsevier Science Inc., 1992.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," The Journal of Machine Learning Research, vol.3, pp.993-1022, 2003

[22] M. Sanderson and B. Croft, "Deriving concept hierarchies from text," Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval,pp. 206-213, 1999.

[23] Eric Tsui, W.M. Wang, C.F. Cheung, et al., "A concept–relationship acquisition and inference approach for hierarchical taxonomy construction from tags," Information Processing & Management, vol.46, pp.44-57, 2010.

[24] J. De Knijff, F. Frasincar, and F. Hogenboom, "Domain taxonomy learning from text: The subsumption method versus hierarchical clustering," Data & Knowledge Engineering, vol.83, pp.54-69, 2013.

[25] A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, "Incorporating concept hierarchies into usage mining based recommendations," Proceedings of Advances in Web Mining and Web Usage Analysis, pp.110-126, 2007.

[26] A. Ferrari and S. Gnesi, "Using collective intelligence to detect pragmatic ambiguities," Proceedings of Requirements Engineering, pp.191-200, 2012.

[27] O. C. Santos and J. G. Boticario, "Practical guidelines for designing and evaluating educationally oriented recommendations," Computers & Education, vol.81, pp.354-374, 2015.

[28] Y. J. Chen, H. C. Chu, Y. M. Chen, and C. Y. Chao, "Adapting domain ontology for personalized knowledge search and recommendation," Information & Management, vol.50, no.6, pp. 285-303, 2013.