ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20160602011232

A PAGERANK-BASED COLLABORATIVE FILTERING RECOMMENDATION APPROACH IN DIGITAL LIBRARIES

Shanshan Guo, Wenyu Zhang, Shuai Zhang

Original scientific paper
In the current era of big data, the explosive growth of digital resources in Digital Libraries (DLs) has led to the serious information overload problem. This trend demands personalized recommendation approaches to provide DL users with digital resources specific to their individual needs. In this paper we present a personalized digital resource recommendation approach, which combines PageRank and Collaborative Filtering (CF) techniques in a unified framework for recommending right digital resources to an active user by generating and analyzing a time-aware network of both user relationships and resource relationships from historical usage data. To address the existing issues in DL deployment, including unstable user profiles, unstable digital resource features, data sparsity and cold start problem, this work adapts the personalized PageRank algorithm to rank the time-aware resource importance for more effective CF, by searching for associative links connecting both active user and his/her initially preferred resources. We further evaluate the performance of the proposed methodology through a case study relative to the traditional CF technique operating on the same historical usage data from a DL.

Keywords: *collaborative filtering; digital library; PageRank algorithm; recommendation approach; social network*

Preporučeni pristup page-rank kolaborativnog filtriranja u digitalnim knjižnicama

Izvorni znanstveni članak
U sadašnje vrijeme opromnog broja podataka, eksplozivni porast digitalnih izvora u Digitalnim Knjižnicama - Digital Libraries (DLs) doveo je do ozbiljnog problema preopterećenja informacijama. Taj trend zahtijeva pristupe personaliziranih preporuka koji bi korisnike DL upoznali s digitalnim izvorima specifičnim za njihove individualne potrebe. U ovom radu predstavljamo personalizirani pristup preporuci digitalnog izvora koji kombinira tehnike PageRank i Collaborative Filtering (CF) u sjedinjenom okviru u svrhu preporuke odgovarajućih digitalnih izvora aktivnom korisniku generirajući i analizirajući mrežu u postojećem vremenu kako odnosa među korisnicima tako i odnosa među izvorima. Kako bi se obradila postojeća pitanja o postavljanju digitalnih knjižnica, uključujući nesigurne profile korisnika, nesigurna obilježja digitalnog izvora, oskudnost podataka i problem hladnog starta, ovaj rad adaptira personalizirani PageRank algoritam kako bi rangirao važnost izvora koji vodi računa o vremenu učinkovitijim CF, tražeći asocijativne linkove koji povezuju i aktivnog korisnika i njegove/njezine početno preferirane izvore. Također ocijenjujemo performansu predložene metodologije kroz analizu slučaja vezanog za tradicionalnu CF tehniku koja koristi iste podatke iz Digitalne knjižnice.

Ključne riječi: *digitalna knjižnica; društvena mreža; kolaborativno filtriranje; PageRank algoritam; pristup koji se preporučuje*

1 Introduction

It has been widely recognized that Digital Libraries (DLs) allow the search, sharing and reuse of vast amount of digital resources in an easy way, but in the current era of big data this fact has led to the serious information overload problem. As personalized recommender systems can provide highly valuable services for large-scaled DLs [19], they have been recently used to aid DL users to find digital resources specific to their individual needs.

Most recommender systems employ Collaborative Filtering (CF) technology [2] as a simulation of word-of-mouth effect to recommend users highly personalized items by aggregating and evaluating the usage data from other similar users or items. Similar users can be identified if they have similar profiles, e.g., they rate similarly or express similar interests on the same items. Similar items can be identified if they are rated similarly by the same users or the same users express similar interests on them. CF has proven widely successful in some famous commercial Web-based systems including Amazon.com, Ebay.com, and Moviefinder.com to make the personalized product recommendation based on the usage data of previous decisions made for similar objectives. Some prototype DL recommender systems also use the CF technique to suggest users the digital resources specific to their individual needs (e.g., [8, 21]).

However, there still exist some issues that impede the deployment of CF technique in large-scaled DL systems:

- *Unstable user profiles*. A DL user's profile may only be stable in a short time. For example, a user may be interested in baby feeding digital resources when she becomes a new mother, but shifts her interests to professional digital resources when she goes to work again.
- *Unstable digital resource features*. A digital resource's feature may only be stable in a short time. For example, a software resource may be upgraded in some time. DL users' ratings or interests on Versions 1.0 and 2.0 are not comparative.
- *Data sparsity*. A DL user is typically reluctant to rate a digital resource after it is downloaded or borrowed, because reading the digital resource, e.g., a professional book, often takes the user too long time for him/her to remember the rating work. Though the DL user's interest can also be acquired by analyzing the user's behaviors, e.g., downloading, commenting or clicking a digital resource, the exponentially increasing digital resources vs the linearly increasing DL users still makes the historical usage data sparser, which are however necessary to identify similar users or digital resources during CF.
- *Cold start*. A lot of new joining DL users or new published digital resources have no historical usage data that can be used to identify similar users or digital resources during CF due to the dynamicity of social network of DL.

The above limitations require an approach that relies not only on the analysis of explicit user ratings or user behaviors, but also on both the consideration of time-aware properties of user profiles and resource features, and on the involvement of the additional source of knowledge (e.g., social network) as the complement to sparse historical usage data. Therefore, in this paper we present a personalized digital resource recommendation approach, which combines PageRank [7] and CF techniques in a unified framework for recommending right digital resources to an active user by generating and analyzing a time-aware network of both user relationships and resource relationships from historical usage data. Our novel idea is to adapt the personalized PageRank algorithm to propagate an influential DL user's time-aware user importance or digital resource's resource importance along the associative links connecting both active user and his/her initially preferred resources, aiming to alleviate the issues of unstable and sparse historical usage data that hinder the usage of traditional CF techniques in DLs.

Although PageRank algorithm was originally invented by the founders of Google as a global ranking model via transitive link analysis to evaluate the numerical importance of each Web page, we believe the social network connecting DL users and digital resources has similar characteristics, enabling us to adapt the personalized PageRank algorithm to it to rank the resource importance for more effective CF.

A case study is conducted to demonstrate how the proposed approach is practical for personalized digital resource recommendation in a DL scenario. The experimental results demonstrate the superiority of our proposed approach over traditional CF methods in providing higher precision rates and more satisfactory precision-recall curves.

2 Related work

CF technique has been widely adopted in personalized recommender systems to deal with information overload problem in various domains including DL area, by suggesting users the products, services or resources specific to their individual needs. It can be classified as user-based or item-based approaches, depending on whether the usage data from similar users or items are aggregated and evaluated.

In a user-based CF system [6, 11, 12], it first computes the similarities between profiles of all user pairs on the basis of their provided ratings or expressed interests to different items, then aggregating and evaluating the target items for the active user from other similar users, and finally recommends the top-N items to the active user based on their aggregated evaluations. In an item-based CF system [5, 9, 13], it first computes the similarities between characteristics of all item pairs on the basis of the ratings or interests provided by different users, then aggregating and evaluating the target items for the active user from other similar items, and finally recommends the top-N items to the active user based on their aggregated evaluations. Hybrid CF system [1, 22, 23] has been motivated by combining both user-based and item-based CF systems and exploiting their respective strengths in order to improve the system performance.

In the DL domain, [21] proposed a quality based recommender system to disseminate information in a university DL. It switches between a user-based CF approach and content-based recommendation approach taking into account the digital resource's quality, to share both the user individual experience and social wisdom [17]. Chen et al. [8] combined CF techniques with a personal ontology model to enhance recommendation performance in DLs by locating like-minded DL users based on ontological similarity computation.

However, CF is vulnerable to data sparsity and cold start problems [18]. When a user or item is new or has no historical usage data, neither user-based nor item-based CF approach can work out a good solution. An available strategy to address such issues is to incorporate social network information such as tag, trust and graph into CF by switching to additional source of knowledge for problem solving. For example, our recent work [22] has combined social tagging data with CF technique for personalized service recommendation by predicting the missing user-dependent Quality-of-Service (QoS) values in a distributed manufacturing service repository. Zou et al. [24] applied a personalized PageRank algorithm to propagate the cold-start users' trust network, from which, the neighbors of a given user are expanded to include others with similar user profiles to his/her original neighbors. Anand and Bharadwaj [3] exploited a graph consisting of various types of entities (users or items), for estimating transitive similarity between entities not directly connected, to bring the entities closer thus alleviating the data sparsity and cold start problems in CF.

In the DL domain, Sun et al. [20] applied a Spreading Activation (SA) [4] algorithm over the graph of domain ontology to propagate the DL users' preferences with user profiles as the initial inputs. When SA converges, more relevant digital resources get higher scores and achieve higher ranking positions. Our recent work [10] has explored Semantic Web technologies for ontology-based modeling of DL service metadata across ubiquitous DLs, enabling to add semantics to DL services to address issues related to representation, cooperation and accessibility of services in or across the communities.

Although the social network information incorporated into CF implies various useful additional sources of knowledge for personalized recommendation (e.g., reputation of users, popularity of items, transitive similarity), the above works only exploit the fixed links connecting both active user and preferred resources with fixed weights, and focus less on the time-aware properties of user profiles, resource features and transitive link analysis, which are very critical to effective CF in dynamic social communities such as DLs. Different from their work, our work adapts the personalized PageRank algorithm to propagate an influential DL user's time-aware user importance or digital resource's resource importance along the dynamic links connecting both active user and preferred resources with adjustable weights, aiming to alleviate the issues of unstable and sparse historical usage data that hinder the usage of traditional CF techniques in DLs.

3 CF recommendation based on PageRank algorithm

In this section, we present a CF recommendation methodology based on PageRank algorithm to explore the propagation and computation of time-aware user importance and resource importance, and rank the resource importance for personalized recommendation with higher importance representing more relevant digital resource.

3.1 Time-aware importance propagation based on PageRank algorithm

Just as the PageRank algorithm can be used in the context of Web graph to evaluate the numerical importance of each Web page via transitive link analysis, it can be adapted to the context of social network of DL to evaluate the numerical importance of each DL node (DL user or digital resource). The social network of DL models the associative links between DL users and digital resources in a similar way as that between Web pages in a Web graph. In general, a node with more ingoing links from the more important nodes with less outgoing links will be ranked with higher importance.

There are two types of associative links between DL nodes, i.e., directive *like* link and bi-directive *resemble* link. The node importance is propagated along the links like that in a Web graph, following the PageRank algorithm in below given rules: a) If a digital resource has been liked (e.g, downloaded, commented or clicked) by an important DL user, then the digital resource will be ranked with higher importance; b) If a DL user/digital resource has been resembled by another important DL user/digital resource, then it will be ranked with higher importance.

Fig. 1 demonstrates an example of such a social network of DL including 4 nodes of DL users (*U*1, *U*2, *U*3 and *U*4), 5 nodes of digital resources (*R*1, *R*2, *R*3, *R*4 and *R*5), and associative links between them. Each link is also weighted differently in a certain time range because both DL users' profiles and digital resources' features may only be stable in a short time.

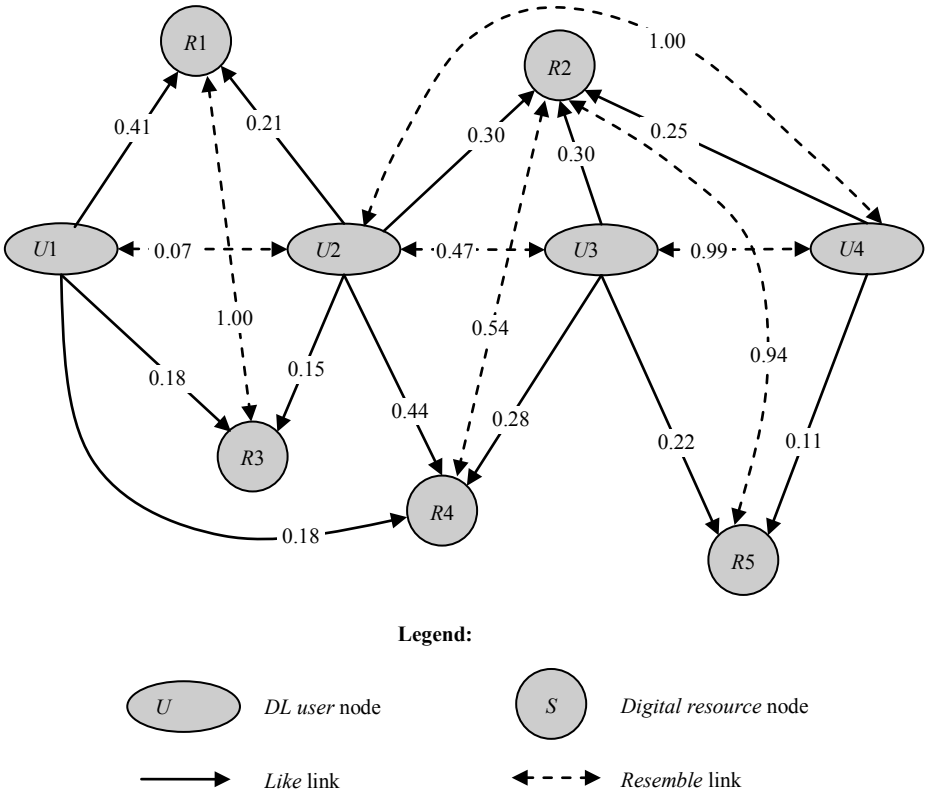

Figure 1 An example of social network of DL

3.2 Time-aware Importance propagation via Like Link

The link weight $w_{i,m}$ on each directive *like* link represents the interest degree $r_{i,m}$ a digital resource *i* has been liked by the DL user *m* in a certain time range. The more frequently and recently the digital resource *i* has been liked (e.g, downloaded, commented or clicked), the more popular and hence important the digital resource *i* is. The link weight $w_{i,m}$ on each *like* link is therefore computed using the below time-aware equation:

$$w_{i,m} = r_{i,m} = \sum_{s=1}^{S} (T - t_s) \times w'_s \tag{1}$$

where *T* denotes the certain time range that is observed for counting the frequency and recency a digital resource *i* has been liked by the DL user *m*; *S* denotes the number of times the digital resource *i* has been downloaded, commented or clicked by the DL user *m* in the time range *T*; t_s denotes the time span the digital resource *i* is downloaded, commented or clicked by the DL user m in the *s*th time with small value of t_s indicating recent

interest and big value of t_s indicating old interest; and w'_s denotes the interest weight assigned to different interests (such as downloading, commenting and clicking) because one kind of interest might express more interest than the other. The link weight $w_{i,m}$ can be normalized within the [0, 1] by dividing it by the maximum *like* link weight corresponding to the most influential *like* link.

3.3 Time-aware Importance propagation via resemble link between two DL users

The link weight $w_{m,n}$ on each bi-directive *resemble* link between two DL users represents the similarity degree the DL user *m* is resembled by another DL user *n* in a certain time range. The more closely the DL user *m* has been resembled, the more relevant and hence important the DL user *m* is. The link weight $w_{m,n}$ on each *resemble* link between two DL users is therefore computed by enhancing the popular linear correlation-based similarity calculation method, i.e., Pearson Correlation Coefficient (PCC) [15] with time-aware properties of user profiles and resource features:

$$w_{m,n} = \frac{\sum_{i=1}^{I}(r_{i,m}-\overline{r}_m)\times(r_{i,n}-\overline{r}_n)\times(T-\left|\overline{t}_{i,m}-\overline{t}_{i,n}\right|)^2}{\sqrt{\sum_{i=1}^{I}(r_{i,m}-\overline{r}_m)^2\times(T-\left|\overline{t}_{i,m}-\overline{t}_{i,n}\right|)^2}\times\sqrt{\sum_{i=1}^{I}(r_{i,n}-\overline{r}_n)^2\times(T-\left|\overline{t}_{i,m}-\overline{t}_{i,n}\right|)^2}} \quad (2)$$

where *T* denotes the certain time range that is observed for counting the frequency and recency various digital resources have been liked by various DL users; *I* denotes the number of digital resources that have been co-liked by both DL users in the time range *T*; $r_{i,m}$ and $r_{i,n}$ denote the time-aware interest degree that the DL users *m* and *n* have liked the digital resource *i* respectively in the time range *T*, and that can be calculated using equation (1); $\overline{r}_m$ and $\overline{r}_n$ denote the average interest degrees that the DL users *m* and *n* have liked the digital resources respectively in the time range *T*; and $\left|\overline{t}_{i,m}-\overline{t}_{i,n}\right|$ denotes the difference of average time span the digital resource *i* is downloaded, commented or clicked by the DL users *m* and *n* in the time range *T*, with smaller value of $\left|\overline{t}_{i,m}-\overline{t}_{i,n}\right|$ indicating that the digital resource *i* has been co-liked by two DL users *m* and *n* in the closer time when the digital resource feature is more stable and hence contribute more significance. $\overline{t}_{i,m}$ represents the average time span the digital resource *i* is downloaded, commented or clicked by the DL user *m*, and is computed using the below given equation:

$$\overline{t}_{i,m} = \frac{1}{S}\sum_{s=1}^{S} t_s \quad (3)$$

where *S* denotes the number of times the digital resource *i* has been downloaded, commented or clicked by the DL user *m* in the time range *T*; and t_s denotes the time span the digital resource *i* is downloaded, commented or clicked by the DL user *m* in the *s*th time with small value of t_s indicating recent interest and big value of t_s indicating old interest. Similarly, $\overline{t}_{i,n}$ represents the average time span the digital resource *i* is downloaded, commented or clicked by the DL user *n*, and is computed using the above equation (3).

Equation (2) shows the link weight $w_{m,n}$ on each bi-directive *resemble* link between two DL users equals the link weight $w_{n,m}$, and is normalized on interval [-1, 1], with a larger value indicating they are more similar. Therefore, only the *resemble* links between two DL users with positive link weight are counted as influential *resemble* links to propagate the user importance.

3.4 Time-aware Importance propagation via resemble link between two digital resources

The link weight $w_{i,j}$ on each bi-directive *resemble* link between two digital resources represents the similarity degree the digital resource *i* is resembled by another digital resource *j* in a certain time range. The more closely the digital resource *i* has been resembled, the more relevant and hence important the digital resource *i* is. The link weight $w_{i,j}$ on each *resemble* link between two digital resources is therefore computed by enhancing the Pearson Correlation Coefficient (PCC) method with time-aware properties of user profiles and resource features:

$$w_{i,j} = \frac{\sum_{m=1}^{M}(r_{i,m}-\overline{r}_i)\times(r_{j,m}-\overline{r}_j)\times(T-\left|\overline{t}_{i,m}-\overline{t}_{j,m}\right|)^2}{\sqrt{\sum_{m=1}^{M}(r_{i,m}-\overline{r}_i)^2\times(T-\left|\overline{t}_{i,m}-\overline{t}_{j,m}\right|)^2}\times\sqrt{\sum_{m=1}^{M}(r_{j,m}-\overline{r}_j)^2\times(T-\left|\overline{t}_{i,m}-\overline{t}_{j,m}\right|)^2}} \quad (4)$$

where *T* denotes the certain time range that is observed for counting the frequency and recency various digital resources have been liked by various DL users; *M* denotes the number of DL users that have co-liked both digital resources in the time range *T*; $r_{i,m}$ and $r_{j,m}$ denote the time-aware interest degree that the DL user *m* has liked the digital resources *i* and *j* respectively in the time range *T*, and that can be calculated using Eq. (1); $\overline{r}_i$ and $\overline{r}_j$ denote the average interest degrees that the DL users have liked the digital resources *i* and *j* respectively in the time range *T*; and $\left|\overline{t}_{i,m}-\overline{t}_{j,m}\right|$ denotes the difference of average time span the digital resources *i* and *j* are downloaded, commented or clicked by the DL user *m* in the time range *T*, with smaller value of $\left|\overline{t}_{i,m}-\overline{t}_{j,m}\right|$ indicating that the digital resources *i* and *j* have been co-liked by the DL user *m* in the closer time when the user profile is more stable

and hence contribute more significance. $\overline{t}_{i,m}$ represents the average time span the digital resource *i* is downloaded, commented or clicked by the DL user *m*, and is computed using Eq. (3). Similarly, $\overline{t}_{j,m}$ represents the average time span the digital resource *j* is downloaded, commented or clicked by the DL user *m*, and is computed using Eq. (3).

Eq. (4) shows the link weight $w_{i,j}$ on each bi-directive *resemble* link between two digital resources equals the link weight $w_{j,i}$, and is normalized on interval [-1, 1], with a larger value indicating they are more similar. Therefore, only the *resemble* links between two digital resources with positive link weight are counted as influential *resemble* links to propagate the resource importance.

3.5 Iterative importance computation based on PageRank algorithm

Given a social network of DL consisting of *V* nodes (including nodes of DL *user* and *digital resource*) and some associative links (including *like* link and *resemble* link), we can simulate a personalized PageRank algorithm [14] by a process of iterative importance computation starting from initial preferences.

The personalized PageRank process starts with picking up a set of initially preferred digital resources by an active DL user, and assigning an initial preference score n_p for both active user and initially preferred digital resources, which is normalized such that $\sum_{p=1} n_p = 1$. Then the $|V|\times 1$ personalized PageRank vector *v* is iteratively computed to rank the importance of both nodes of DL *user* and *digital resource* according to the following equation:

$$v = \alpha M v + (1-\alpha)N \tag{5}$$

where $\boldsymbol{M}$ is a $|V|\times|V|$ transition matrix that is constructed according to the below given equation:

$$m_{i,j} = \frac{w_{i,j}}{\sum_{k\in Out(j)} w_{k,j}} \tag{6}$$

where $w_{i,j}$ denotes the link weight from node *j* to node *i*, *Out* (*j*) denotes the set of nodes *k* that form the corresponding outgoing links from node *j* to node *k*, and $m_{i,j}$ denotes the matrix element in row *i* and column *j* and is the normalized value of $w_{i,j}$ such that $\sum_{k\in Out(j)} m_{k,j} = 1$.

In Eq. (5), α is a damping factor, usually set to 0.85. *N* is the $|V|\times 1$ initial preference vector whose components corresponding to both active user and initially preferred digital resources have initial preference scores n_p assigned by the active user (with $\sum_{p=1} n_p = 1$), while other components have value 0. This biases the importance transition to both active user and initially preferred digital resources over other DL nodes, and thus facilitates the personalized recommendation of digital resources for an active DL user.

As an example, Tab. 1 shows the transition matrix $\boldsymbol{M}$ corresponding to the illustrative social network of DL in Fig. 1. Suppose *U*2 is the active DL user, who picks up *R*3 and *R*5 as initially preferred digital resources, and assigns some initial preference scores in the initial preference vector *N*, i.e., $\{0, 0.5, 0, 0, 0, 0, 0.2, 0, 0.3\}^{\mathrm{T}}$. The initial PageRank vector *v* is a column vector such that all of its components have value 1/9. After iterative importance computation, the converged values of PageRank vector *v* are $\{0.002, 0.095, 0.029, 0.042, 0.134, 0.261, 0.149, 0.098, 0.190\}^{\mathrm{T}}$, whose first four values represent the importance of 4 DL users (*U*1, *U*2, *U3* and *U*4) respectively and whose last five values represent the importance of 5 digital resources (*R*1, *R*2, *R*3, *R*4 and *R*5) respectively. Therefore, the top-*N* (e.g., top-3) digital resources (*R*2, *R*5 and *R*3) with decreasing order of resource importance (0.261, 0.190, 0.149) will be recommended to the active DL user as personalized recommendation results.

Table 1 A Transition Matrix $\boldsymbol{M}$

	*U*1	*U*2	*U*3	*U*4	*R*1	*R*2	*R*3	*R*4	*R*5
*U*1	0	0.07/2.64	0	0	0	0	0	0	0
*U*2	0.07/0.84	0	0.47/2.26	1.00/2.35	0	0	0	0	0
*U*3	0	0.47/2.64	0	0.99/2.35	0	0	0	0	0
*U*4	0	1.00/2.64	0.99/2.26	0	0	0	0	0	0
*R*1	0.41/0.84	0.21/2.64	0	0	0	0	1.00/1.00	0	0
*R*2	0	0.30/2.64	0.30/2.26	0.25/2.35	0	0	0	0.54/0.54	0.94/0.94
*R*3	0.18/0.84	0.15/2.64	0	0	1.00/1.00	0	0	0	0
*R*4	0.18/0.84	0.44/2.64	0.28/2.26	0	0	0.54/1.48	0	0	0
*R*5	0	0	0.22/2.26	0.11/2.35	0	0.94/1.48	0	0	0

4 PageRank-based CF recommender system architecture

Fig. 2 shows the PageRank-based CF recommender system architecture that mainly includes three modules: *keyword matchmaking module*, *importance computation module*, and *resource recommendation module*, and functions according to the following procedures:

Step 1. An active DL user inputs a digital resource request through a set of keywords.

Step 2. The keyword matchmaking module is used to return a set of syntactically similar digital resources by solving a Term Frequency vs. Inverse Document Frequency (TFIDF) [16] weighting problem for dealing with the trade-offs among both resource-keyword matching frequency and overall keyword-related resource frequency.

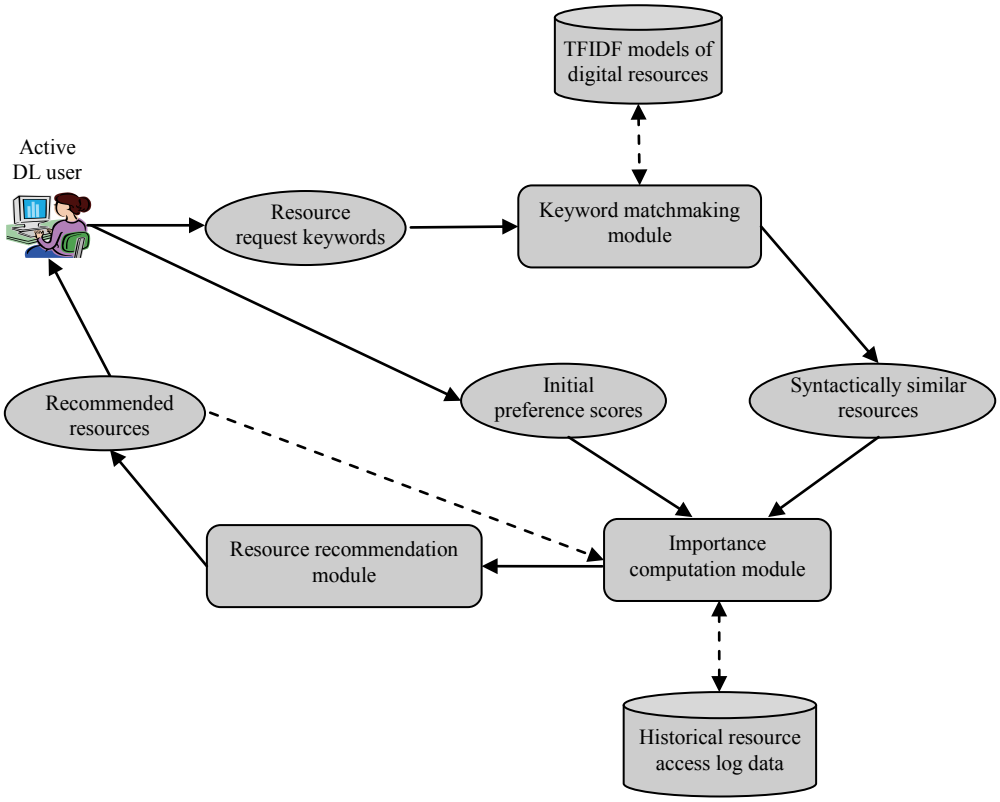

Figure 2 PageRank-based CF recommender system architecture

Step 3. The active DL user picks up a few preferred digital resources from the above syntactically matching digital resources, and assigns initial preference scores for both active user and initially preferred digital resources.

Step 4. The importance computation module is used to search for associative links connecting both active user and his/her initially preferred resources and rank the time-aware resource importance by applying the revised personalized PageRank algorithm on the historical resource access log data.

Step 5. The resource recommendation module is used to recommend the active DL user the digital resources ranked with higher importance. If the active DL user is very interested in few of them that have not been picked up as initially preferred digital resources, s/he may re-pick up a few new preferred digital resources from the first recommendation result, re-assigns new initial preference scores for both active DL user and new initially preferred digital resources, and repeats steps 4 and 5. The repetitive recommendation process helps to facilitate the active DL user to better understand his/her interests and further increases the recommendation accuracy including precision and recall rates.

5 Experimental evaluation and results

This section demonstrates the experimental results from a university DL on the evaluation of the proposed PageRank-based CF recommendation approach compared to traditional CF recommendation in terms of recall and precision rates and precision-recall curves.

A Java-based object-oriented software prototype is implemented. The digital resource repository of the current prototype system under evaluation contains 500 digital resources related with different areas. The historical resource execution log data stores the frequency and recency various digital resources have been downloaded, commented or clicked by different DL users.

5.1 Evaluation metrics

To evaluate the personalized recommendation results, two popular measures, i.e., recall rate, and precision rate are adopted. In the current context, the recall rate refers to the ratio of the number of correctly recommended relevant digital resources to the number of relevant digital resources, and the precision rate refers to the ratio of the number of correctly recommended relevant digital resources to the number of recommended digital resources. The relevant digital resources can be identified as the manual recommendation results generated by a group of domain experts (e.g., library staffs).

To show our proposed method can produce more efficient digital resource recommendation results compared to existing approaches, we select two typical approaches as baseline approaches: a) keyword-based TFIDF-enhanced recommendation and b) keyword-based TFIDF-enhanced CF recommendation.

A series of experiments with different number of top-*N* returned digital resources are carried out in very similar circumstances to make the comparison between the above methods with regard to their recall rates, precision rates, and precision-recall curves. Each one of the registered DL users has the same chances to become the active user in search of digital resources.

5.2 Experimental results

Fig. 3 shows the comparison between the above methods with regard to their average precision rates.

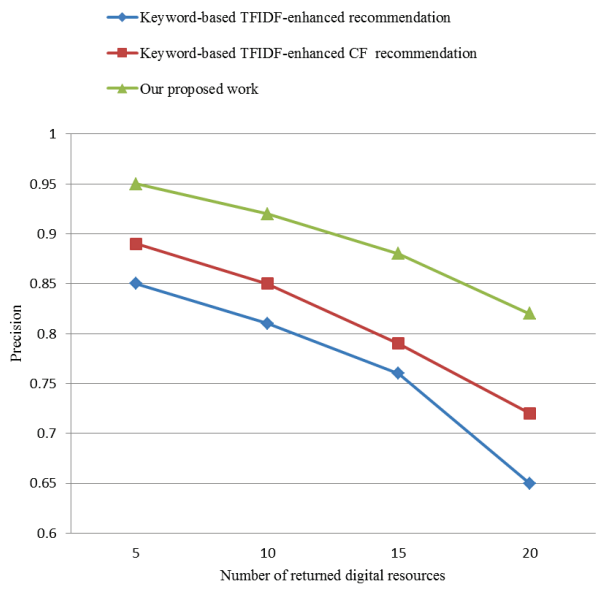

Figure 3 Comparison of the precision rates among three digital resource recommendation methods

The experiments with different number of top-*N* returned digital resources reveal that the precision rates of keyword-based TFIDF-enhanced CF recommendation approach consistently exceed those of keyword-based TFIDF-enhanced recommendation approach. This is because the former adopts the latter's advantages in TFIDF-enhanced keyword matchmaking capability, and complements the latter with CF recommendation capability by aggregating and evaluating the usage data from other similar DL users or digital resources.

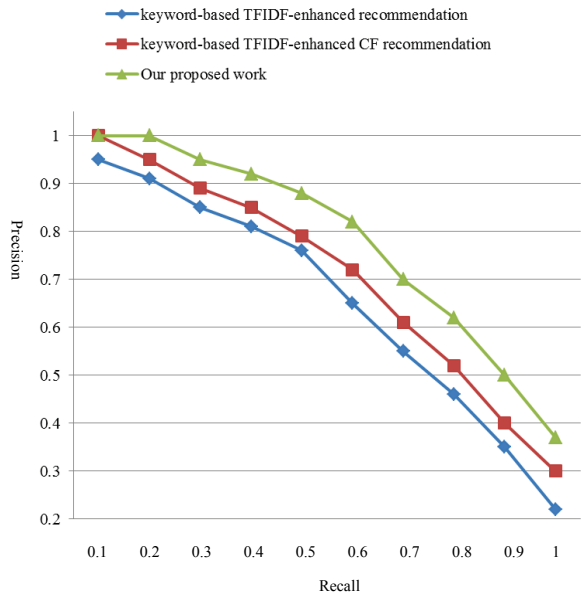

Figure 4 Comparison of the precision-recall curves among three digital resource recommendation methods

The experiments with different number of top-*N* returned digital resources also reveal that the precision rates of our proposed PageRank-based CF recommendation approach consistently exceeds those of keyword-based TFIDF-enhanced CF recommendation approach. This is because the former adopts the latter's advantages in TFIDF-enhanced keyword matchmaking and CF recommendation capabilities, and complements the latter by emphasizing time-aware importance propagation and personalization based on the structural and dynamic properties of social network of DL.

Fig. 4 shows the precision-recall curves for the above methods, illustrating how the average precision rates change when the average recall rates increment from 0.1 to 1. The experiments reveal that under the same recall rates, the precision rates of the proposed work consistently exceed those of keyword-based TFIDF-enhanced CF recommendation approach, which consistently exceed those of keyword-based TFIDF-enhanced recommendation approach.

6 Conclusion

This paper presents a novel PageRank-based CF recommendation approach for personalized digital resource recommendation to overcome the information overload problem in DLs. As opposed to the majority of existing social network-based CF methods for personalized recommendation that focus less on the time-aware properties of user profiles, resource features and transitive link analysis, our proposed approach adapts the personalized PageRank algorithm to rank the time-aware resource importance for more effective CF, by searching for associative links connecting both active user and his/her initially preferred resources. It explores and exploits the structural and dynamic properties of social network of DL, thus alleviating the issues of unstable and sparse historical usage data that hinder the usage of traditional CF techniques in DLs.

Our experimental results show that the proposed approach improves the relevance of recommendation result more significantly than the baseline approaches. Our future work will further validate and improve the effectiveness and performance of proposed approach in dealing with the following two issues resulting from the big data: a) how to self-adapt the weight setting of associative links, for example, by applying genetic algorithm, to optimize the PageRank propagation in a very complicated social network of DL? b) How to leverage the parallel computing technology, for example, MapReduce algorithm, to expedite the PageRank propagation in a large-scaled social network of DL? The proposed approach may also be adapted to the other fields such as e-government, e-science and e-commerce with potential application.

Acknowledgements

The work has been supported by China National Natural Science Foundation (Nos. 51375429, 51475410), Humanities and Social Sciences Research Project of Ministry of Education, China (No. 15YJC870008), Scientific Research Project of Zhejiang Provincial Department of Education, China (No. Y201534507), and Zhejiang Natural Science Foundation of China (No. LY17E050010).

7 References

[1] Abdelwahab, A.; Sekiva, H.; Matsuba, I.; Horiuchi, Y.; Kuroiwa, S. Feature optimization approach for improving the collaborative filtering performance using particle swarm optimization. // Journal of Computational Information Systems. 8, 1(2012), pp. 435-450.

[2] Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. // IEEE Transactions on Knowledge and Data Engineering. 17, 6(2005), pp. 734-749. https://doi.org/10.1109/TKDE.2005.99

[3] Anand, D.; Bharadwaj, K. K. Exploring graph-based global similarity estimates for quality recommendations. // International Journal of Computational Science and Engineering. 9, 3(2014), pp. 188-197. https://doi.org/10.1504/IJCSE.2014.060683

[4] Anderson, J. R. A spreading activation theory of memory. // Journal of Verbal Learning and Verbal Behavior. 22, (1983), pp. 261-295. https://doi.org/10.1016/S0022-5371(83)90201-3

[5] Barragáns-Martínez, A.; Costa-Montenegro, E.; Burguillo, J. C.; Rey-López, M.; Mikic-Fonte, F. A.; Peleteiro, A. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. // Information Sciences. 180, 22(2010), pp. 4290-4311. https://doi.org/10.1016/j.ins.2010.07.024

[6] Bobadilla, J.; Serradilla, F.; Bernal, J. A new collaborative filtering metric that improves the behavior of recommender systems. // Knowledge-Based Systems. 23, 6(2010), pp. 520-528. https://doi.org/10.1016/j.knosys.2010.03.009

[7] Brin, S.; Page, L. The anatomy of a large-scale hypertextual Web search engine. // Computer Networks and ISDN Systems. 30, 1-7(1998), pp. 107-117. https://doi.org/10.1016/s0169-7552(98)00110-x

[8] Chen, L. C.; Kuo, P. J.; Liao, I. E. Ontology-based library recommender system using MapReduce. // Cluster Computing. 18, 1(2015), pp. 113-121. https://doi.org/10.1007/s10586-013-0342-z

[9] Gao, M.; Fu, Y. Q.; Chen, Y. X.; Jiang, F. User-weight model for item-based recommendation systems. // Journal of Software. 7, 9(2012), pp. 2133-2140. https://doi.org/10.4304/jsw.7.9.2133-2140

[10] Guo, S. S.; Zhang, W. Y.; Zhang, S.; Cai, M. Towards building a digital library service metadata model on the Semantic Web. // International Journal of Database Theory and Application. 8, 5(2015), pp. 1-15.

[11] Huang, Y.; Gao, X. D.; Gu, S. J. UARR: a novel similarity measure for collaborative filtering recommendation. // Cybernetics and Information Technologies. 13, (2013), pp. 122-130. https://doi.org/10.2478/cait-2013-0043

[12] Hwang, C. S.; Fong, R. S. A hybrid recommender system based on collaborative filtering and cloud model. // World Academy of Science, Engineering and Technology. 51, (2011), pp. 500-505.

[13] Liu, D. R.; Tsai, P. Y.; Chiu, P. H. Personalized recommendation of popular blog articles for mobile applications. Information Sciences. 181, 9(2011), pp. 1552-1572. https://doi.org/10.1016/j.ins.2011.01.005

[14] Page, L.; Sergey, B.; Rajeev, M.; Terry, W. The PageRank Citation Ranking: Bringing Order to the Web. // Technical Report, Stanford, 1998.

[15] Rodgers, J. L.; Nicewander, W. A. Thirteen ways to look at the correlation coefficient. // The American Statistician. 42, 1(1988), pp. 59-66. https://doi.org/10.2307/2685263

[16] Salton, G.; Buckley, C. Term-weighting approaching in automatic text retrieval. // Information Processing and Management. 24, 5(1988), pp. 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[17] Schenkel, R.; Crecelius, T.; Kacimi, M.; Neumann, T.; Parreira, J.; Spaniol, M.; Weikum, G. Social wisdom for search and recommendation. // Data Engineering Bulletin. 31, 2(2008), pp. 40-49.

[18] Shambour, Q.; Lu, J. A trust-semantic fusion-based recommendation approach for e-business applications. // Decision Support Systems. 54, (2012), pp. 768-780. https://doi.org/10.1016/j.dss.2012.09.005

[19] Smeaton, A.; Callan, J. Joint DELOS-NSF workshop on personalization and recommender systems in digital libraries. // SIGIR Forum. 35, 1(2011), pp. 7-11.

[20] Sun, T.; Zhang, M.; Yan, F.; Deng, Z. H. Personalized search in digital libraries via spreading activation model. // Web Intelligence and Agent Systems. 11, 2(2013), pp. 137-147.

[21] Tejeda-Lorente, A.; Porcel, C.; Peis, E.; Sanz, R.; Herrera-Viedma, E. A quality based recommender system to disseminate information in a university digital library. // Information Sciences. 261, (2014), pp. 52-69. https://doi.org/10.1016/j.ins.2013.10.036

[22] Zhang, W. Y.; Zhang, S.; Chen, Y. G.; Pan, X. W. Combining social network and collaborative filtering for personalised manufacturing service recommendation. // International Journal of Production Research. 51, 22(2013), pp. 6702-6719. https://doi.org/10.1080/00207543.2013.832839

[23] Zheng, X. L.; Ding, W. F.; Xu, J. N.; Chen, D. R. Personalized recommendation based on review topics. // Service Oriented Computing and Applications. 8, 1(2014), pp. 15-31. https://doi.org/10.1007/s11761-013-0140-8

[24] Zou, H. T.; Gong, Z. G.; Zhang, N.; Zhao, W.; Guo, J. Z. TrustRank: a cold-start tolerant recommender system. // Enterprise Information Systems. 9, 2(2015), pp. 117-138. https://doi.org/10.1080/17517575.2013.804587

Authors' addresses

Shanshan Guo
Library, Zhejiang University of Finance and Economics
18 Xueyuan Street, Xiasha, Hangzhou, China 310018
E-mail: guoshanshanc@163.com

Wenyu Zhang, *corresponding author*
School of Information,
Zhejiang University of Finance and Economics
18 Xueyuan Street, Xiasha, Hangzhou, China 310018
E-mail: zhangwyc@zju.edu.cn

Shuai Zhang
School of Information,
Zhejiang University of Finance and Economics
18 Xueyuan Street, Xiasha, Hangzhou, China 310018
E-mail: zs760914@sina.com