Int. J. Production Economics 165 (2015) 185–193

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework

Jianshan Sun [a], Wei Xu [b,*], Jian Ma [c], Jiasen Sun [d]

[a] School of Management, HeFei University of Technology, Hefei 230009, PR China
[b] School of Information, Renmin University of China, Beijing 100872, PR China
[c] Department of Information Systems, City University of Hong Kong, Hong Kong, China
[d] Dongwu Business School, Soochow University, Suzhou 215006, PR China

ARTICLE INFO

Article history:
Received 31 December 2013
Accepted 27 December 2014
Available online 6 January 2015

Keywords:
Social network services
Expert finding
Recommendation
Research analytics framework
Big data analytics

ABSTRACT

With the rapid proliferation of information technology, the increasing amount of information available has posted significant challenges on relevant information discovery for users. An alternative way is to find an expert with specific expertise. Expert recommendation is important in variety of contexts ranging from industry to academia. Information retrieval methods or graph-based methods have been proposed to approach this problem in previous research while some important contextual factors are ignored. In this paper, considering the factors of topic relevance, expert quality, and researcher connectivity, we propose a novel researcher modeling approach to recommend experts in scientific communities. The proposed recommendation method is well evaluated and compared with some commonly used recommendation models. Furthermore, the proposed method has been implemented in ScholarMate (www.scholarmate.com), an online research social network platform. The experimental results exhibit that the proposed method is more effective than baseline methods, and it is a potential recommendation method to find domain experts on research social network services.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid proliferation of information technology, the increasing amount of information available has posted significant challenges on relevant information discovery for users. An alternative way is to find an expert with specific expertise to help with a task or address a problem. The discovery of expertise is important in variety of contexts ranging from industry to academia (Deng et al., 2008). For example, in organization settings, an employee may require a highly trained specialist to consult with a specific problem when he has no experience to a specific project, so finding experts may help him reduce time cost and facilitate a better solution. In research communities, recommending external experts for research project selection (Sun et al., 2008) and assigning reviewers to paper manuscripts automatically for peer-review (Karimzadehgan et al., 2008) are two essential tasks. Furthermore, when researchers need guidance on a subject matter or need to find collaborators working together in related areas, they can also refer to expert finding systems. As discussed above, finding suitable domain experts with appropriate skills and knowledge is critical important in research communities (Sun et al. 2014). Since manually approaches to construct expert databases are labor-intensive and time-consuming, various automated approaches have been proposed to build researcher profile and find potential experts (Afzal and Maurer, 2011).

With the advent of expert search track in the text retrieval conference (TREC) Enterprise Track (Craswell et al., 2005), amount of attention has been paid to expert finding tasks. Information retrieval methods (Balog et al., 2009, 2007; Craswell et al., 2005; Fang and Zhai, 2007; Macdonald and Ounis, 2008; Serdyukov et al., 2007) and graph-based methods (Cao et al., 2005; Deng et al., 2012; Dom et al., 2003; Zhang et al., 2007; Zhou et al., 2007) as two main research streams have been proposed to find experts for given query topics. Information retrieval methods, either profile-based methods or document-based methods, treat expert finding task similar to traditional ad hoc retrieval tasks while other contextual factors (such as expert quality and social connections) affecting retrieval performance were ignored. Keyword semantics problem which often caused keyword mismatch problem was also ignored in expert finding research. Graph-based methods referred to PageRank algorithm principles in constructed networks. It has also achieved comparative performance in specific domains.

* Correspondence to:School of Information, Renmin University of China No. 59, Zhongguancun Road, Haidian District, Beijing, 100872, China Tel.: +86 10 82500904; fax: +86 10 62511184.
E-mail address: weixu@ruc.edu.cn (W. Xu).

http://dx.doi.org/10.1016/j.ijpe.2014.12.038
0925-5273/© 2015 Elsevier B.V. All rights reserved.

However, with the rapid development of online social networks, these methods suffer from run-time efficiency problem and it is difficult to implement them in large-scale context (Meng et al. 2014). Taking ScholarMate[1] for example, there are hundreds of thousands of online researchers, and millions of research materials, so finding experts in similar research fields, and ranking them in terms of some criterion are indeed time-consuming. To overcome the shortages of current studies, since topic relevance, academic connectivity and expertise level are all valuable for expert finding as identified in existing literatures, we leverage research analytics framework to profile researchers by incorporating multi-source data and semantic queries, and then implement MapReduce platform to speed up the computation process. Thus, our proposed method aims to provide a better solution for expert finding by using big data analytics.

In this research, we propose a novel big data analytics approach for researcher modeling combining three-dimensional characteristics to recommend relevant experts in research communities. The semantic content of query and researcher profile is analyzed to solve keyword mismatch problem. Meanwhile, expertise levels of experts are modeled and accumulated expertise from other linked experts is aggregated to improve recommendation performance. Then, a two-stage recommendation process is conducted to provide more relevant and authoritative experts using MapReduce platform. The prototype system for expert finding has been implemented as an application service in ScholarMate. The proposed approach is evaluated through designed real experiment in ScholarMate. The results show that the proposed approach outperforms other baseline methods.

This paper presents three main contributions. Firstly, we construct researcher profiles by combining multi-dimensional characteristics using research analytics framework (RAF). Topic relevance, authoritative quality and enriched expertise are deeply mined and then aggregated to improve retrieval accuracy. Secondly, we alleviate keyword mismatch problem in traditional methods and provide sufficient experimental evaluations and investigate how proposed research analytics framework approach works in terms of improving the expert finding performance. Thirdly, the designed recommender system by MapReduce is especially fit for research communities. We have incorporated it into online research social network website to facilitate content sharing and potential researcher collaboration.

The rest of the paper is organized as follows. Section 2 reviews related literatures and identifies the research gap. Section 3 describes the novel big data analytics method for expert finding, and elaborates a researcher modeling process using research analytics framework and expert finding algorithm based on MapReduce platform. Section 4 shows the implemented system interfaces and Section 5 presents the experiment results. We give the summarization of our research and point out the future work in Section 6.

2. Literature review

Expert finding task addresses the problem of providing a ranking list of people who are knowledgeable and authoritative about a given query topic (Wang et al., 2013). This task has attracted much attention in recent years after its first inclusion in the TREC Enterprise Track (Craswell et al., 2005). The expert finding task have been studied in various contexts such as enterprise corpora (Balog et al., 2009), sparse data university environments (Balog et al., 2007), online knowledge communities (Wang et al., 2013) and digital libraries (Gollapalli et al., 2011). Generally, two research streams approached expert finding problem: one is to employ information retrieval techniques from the relevance perspective; the other is to apply graph-based methods to a network (social network or heterogeneous information network) from the connectivity perspective.

The methods in relevance stream could be further classified into two categories: profile-based methods and document-based methods (Fang and Zhai, 2007). Profile-based methods directly built the expert candidate profile based on associated documents and then generated the ranking score according to the profile in response to a user query. The model 1 in Balog et al. (2009) first built a term-based expertise profile (virtual document vector) for each candidate, and ranked the candidate experts based on the relevance scores of their profiles for a given topic by using traditional ad hoc retrieval models. On the other hand, document-based methods first ranked documents in the corpus given a query topic. Then they found the associated candidates from the subset of retrieved documents. The model 2 in Balog et al. (2009) employed language model to find experts based on ranked documents and the experimental results showed it outperformed the model 1. Furthermore, Cao et al. (2005) used a probabilistic approach to rank experts by combining relevance model and co-occurrence model. The topic-based model and the hybrid model were exploited by Deng et al. (2008) to achieve better performance than basic language models. Macdonald and Ounis (2008) presented yet another approach based on a voting model for expert search. Profile-based methods operated efficiently due to smaller documents in size for modeling while they performed less effectively than other approaches because they could not measure each document individually. Document-based methods allowed the application of advanced text modeling techniques in ranking individual documents while they showed inconvenient data management. Moreover, the proposed techniques above ignored keywords semantics, so synonym and polysemy problems were appeared, which further caused mismatch problem (Sun et al., 2014). In this paper, we construct keyword correlation matrix for semantic query expansion to improve expert finding performance.

The methods in connectivity stream benefited from the success of PageRank (Page et al., 1999) and hyperlink-induced topic search (HITS) (Kleinberg, 1999) algorithms in search engines. In particular, connectivity propagation models for expert search were studied in e-mail collections (Dom et al., 2003) and enterprise collections (Serdyukov et al., 2008). PageRank algorithm was employed in Wang et al. (2013) to compute expert's authority and then relevance and authority were combined to rank experts in online communities. Zhou et al. (2007) proposed a coupled random walk model between authorship networks and citation networks for ranking authors and documents together. Recently, heterogeneous bibliographic networks were modeled and exploited (Deng et al., 2012) for expertise ranking. Although graph-based techniques were advanced and sophisticated for expert ranking, their success is dependent on the accuracy of whole network construction and the computation cost is often very high. In this research, we employ the AuthorRank algorithm on the researchers' collaboration network and leverage MapReduce framework to speed up the computation process.

Meanwhile, Some scholars have argued that it is not enough to find experts by only looking at the queries' without taking the users into consideration (Hofmann et al., 2010). They claimed that several contextual factors (e.g. proximity and expert quality) may have effects on the decision concerning which experts to recommend. Users were more likely to select expertise search results that included social network information (Shami et al., 2008). Hofmann et al. (2010) argued that many of these factors could be considered in the modeling process, and claimed that integrating them with retrieval models could improve retrieval performance. In addition, existing studies show that topic relevance, academic connectivity and expertise level are all valuable information sources for expert finding (Shami et al., 2008). However, very limited research combines them to give a better solution to the problem. In this research, we consider the expertise level of researchers by measuring the quality of their publications and

[1] www.scholarmate.com.

research projects, and proposes research analytics framework to integrate relevance, quality, and connectivity for researcher profiling.

Furthermore, since a number of researchers emerge in Scholarmate, in which some complex relationships exist among researchers, such as friendship and researchers in one research/teaching groups, previous research was lack of run-time efficiency to find experts for large-scare research communities. With the rapid development of cloud computing (Armbrust et al., 2010; Brian et al., 2008) and big data research (Chen et al., 2012; Cheng et al., 2012), some useful big data analytics tools such as MapReduce can speed up similarity computation (Elsayed et al., 2008) and social network analysis (Shi et al., 2013), so, in this research, we offer MapReduce as an effective computation tool to improve the efficiency of expert finding process in research community.

To cover the research gap discussed above, we propose a novel big data analytics to find knowledgeable experts for research social network services. We leverage research analytics framework to profile researchers in research community, and offer MapReduce platform to compute researcher profiling and discover domain experts.

3. Methodology

In Silva et al. (2013), we defined research analytics framework (RAF) and successfully applied it into the context of research project selection. We consider research analytics as the application of methods and theories (including scientometrics, business intelligence and social network analysis) to transform research related data into relevant information in research management. In this study, we leverage the RAF in the context of large scale expert finder application for research social network services. The proposed expert finder system employs relevance analysis, quality analysis, and connectivity analysis modules to build a more comprehensive researcher model for expert finding task. The process and main components of the proposed system are shown in Fig. 1.

3.1. Profiling

In general, profiling is the process of identifying and determining relevant information and attributes that can be used to characterize a given object. In the expert finding context, we focus on how to collect necessary data to construct more comprehensive researcher profile. As stated in Park and Chang (2009), researcher profiles can be constructed from two ways: declaration and inference based on observation of research activities. Declared profiles are reflected by subjective information which often contains self-claimed interests, expertise and skills. The subjective information is often represented by structured keywords. However, to obtain this information needs imposing extra work on the researcher so it is often incomplete and hard to update. Observation and interpretation of research activities has the potential to build more accurate profiles and these profiles can be constructed automatically and objectively. In research communities, submitting research manuscripts and applying scientific grants are two main activities of researchers. Thus, publications and projects can be considered as two important evidences to assess researchers' expertise. As illustrated in Fig. 1 (Profiling part), publications and projects contain several interrelated entities. Researchers are linked through co-authoring journal articles or collaborating scientific projects. Articles are often located in journals whose discipline ranks (i.e. A, B, C) reflect authors' expertise level. Projects can be classified into different levels (i.e. Nation, Ministry, Province, and City) which reflect authors' expertise level. Furthermore, there are several keywords contained in publications and projects and they reflect the authors' research areas.

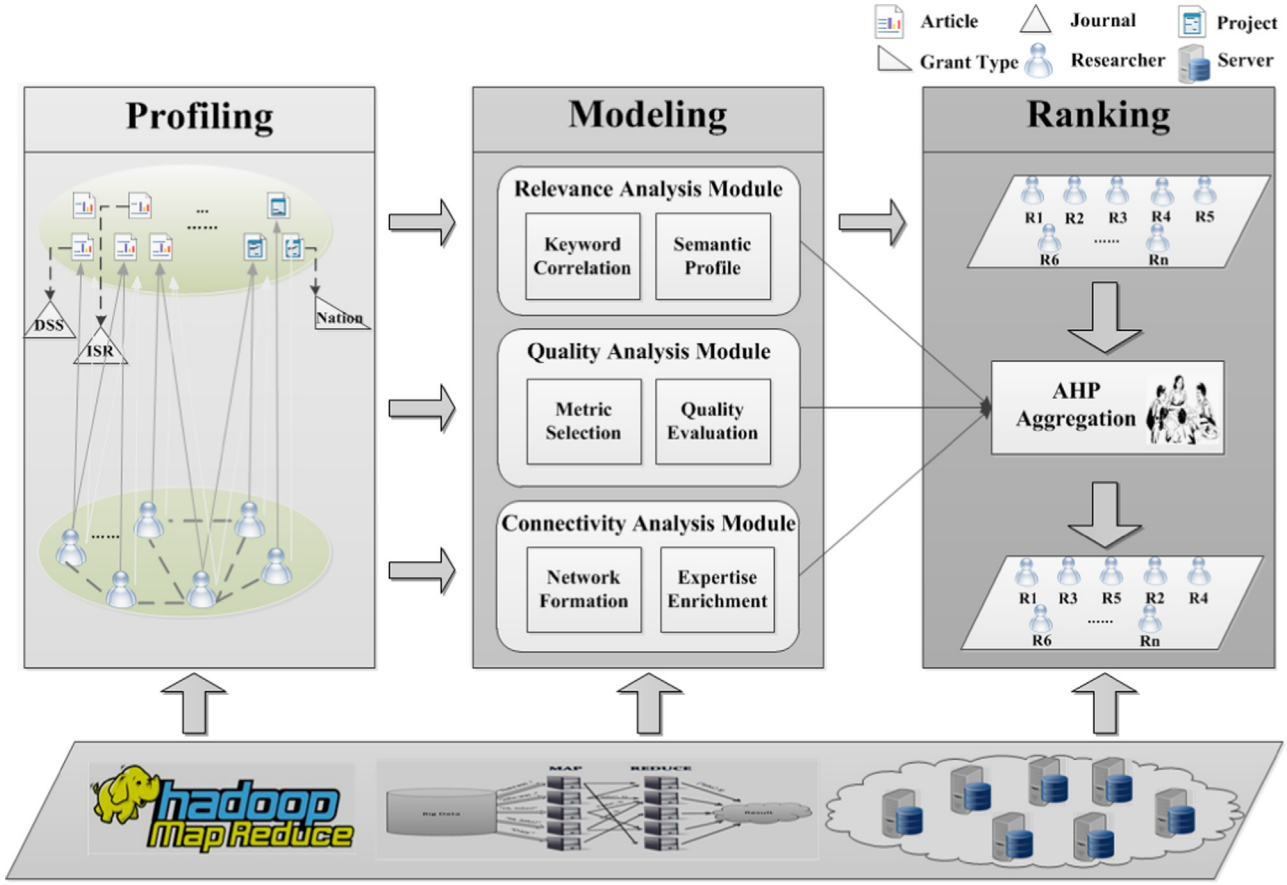

Fig. 1. The architecture of the expert finder system.

In this research, we have collected all the scientific information about researchers and generated correlation matrices for researcher modeling. The Keyword–Document matrix is used for relevance analysis module. The Article–Journal matrix and Project–Type matrix are used for quality analysis module. The researcher–researcher matrix is used for connectivity analysis module.

3.2. Modeling

3.2.1. Relevance analysis module

This module proposes the semantic keyword weighting method to determine the content relevance of candidate experts. As proposed in previous work (Sun et al., 2014), this method can overcome keyword mismatch problem of traditional keyword weighting methods (such as TFIDF, BM25). Keyword–Document (KD) matrix generated in the previous section is used to derive keyword similarity. The KD matrix is a $n_k \times n_d$ matrix which is used to denote the association between keywords and documents, where $n_k$is the number of keywords andn_d is the number of documents. The elements in the KD matrix denote weighted frequency scores (FS) of a keyword in the document. FS can be calculated as follows:

$$FS_{kd} = \alpha f_{tit} + \beta f_{key} + \gamma f_{abs} \tag{1}$$

where f_{tit} is the frequency of the keyword in the document title, f_{key} is the frequency of the keyword in author-assigned keywords list and f_{abs} is the frequency of the keyword in the document abstract. α, β and γ, subject to the equation $\alpha+\beta+\gamma=1$, are weights of f_{tit}, f_{key} and f_{abs} respectively. The detail settings refer to Sun et al. (2014).

3.2.1.1. Keyword correlation calculation. A novel keyword similarity method relying on mutual reinforcement principle (Wu et al., 2013) is employed in this work. The method uses an iterative approach to compute similarities whereby the similarity between any two objects (keywords or documents) is computed based on the similarities already computed in the previous iteration. In detail, the similarity computation is performed as follows:

Initial step

$$sk^0(k_m, k_n) = \theta_{mn}, \quad sd^0(d_m, d_d) = \theta_{mn} \tag{2}$$

In pth step

$$sk^p(k_m,\ k_n) = \frac{SK^p(k_m,\ k_n)}{\sqrt{SK^p(k_m,\ k_m)} \cdot \sqrt{SK^p(k_n,\ k_n)}} \tag{3}$$

$$sd^p(d_m,\ d_n) = \frac{SD^p(d_m,\ d_n)}{\sqrt{SD^p(d_m,\ d_m)} \cdot \sqrt{SD^p(d_n,\ d_n)}} \tag{4}$$

where

$$SK^p(k_m,\ k_n) = \sum_{i,j=1}^{n_d} FS_{mi} \cdot \varphi_{ij} \cdot sd^{p-1}(d_i,\ d_j) \cdot FS_{nj} \tag{5}$$

$$SD^p(d_m,\ d_n) = \sum_{i,j=1}^{n_k} FS_{im} \cdot \varphi_{ij} \cdot sk^{p-1}(k_i,\ k_j) \cdot FS_{jn} \tag{6}$$

In the initial step, keyword similarity $sk^0(k_m, k_n)$ and the document similarity $sd^0(a_m, d_n)$ are defined. Each keyword (resp., document) is similar only to itself and it is dissimilar to all other keywords (resp., document). At the $p$thstep, let $sk^p(k_m,\ k_n)$ (resp., $sd^p(d_m,\ d_n)$) be the keyword (resp., document) similarity betweenk_m and k_n (resp., $d_m,\ d_n$). In Eqs. (5–6), φ_{ij} is equal to 1 if $i=j$, otherwise it is equal toφ where φ is mutual reinforcement factor and $\varphi\epsilon[0, 1]$. In this way, the keyword correlation matrix can be constructed and it is used to compute matching degree between semantic query and researcher profile as presented in the next section.

3.2.1.2. Semantic profile matching. In this research, we have applied a filtering strategy to efficiently recommend selected list of experts from the researcher database. As exact keyword matching could generate inadequate results, we employ semantic keyword matching to filter out irrelevant researchers. Previous research has demonstrated the high performance of the semantic keyword matching method in document retrieval (Quattrone et al., 2011) and article recommendation (Sun et al., 2014). The user query is extended by adding similar keywords. To make it less complicated, we add five more keywords to the query. Similar keywords are identified based on the pre-computed keyword correlation matrix. Then, the enriched query is used to match with potential expert profiles. The relevance score by matching the enriched query with researcher profile is calculated as follows:

$$RS(q, r) = \sum_{i=1}^{n_k} FS_{kr}(i) \cdot sim(a, i) \tag{7}$$

where $RS(q, r)$ denotes keyword matching degree of the query and researcher profile; n_k is the number of distinct keywords in enriched query and a is the keyword of query itself; $FS_{kr}(i)$ represents the frequency score of keyword i in the researcher profile; $sim(a,i)$ indicates keyword similarity, where $sim(a,i)=1$, if $i=a$ and $sim(a,i)$ equals similarity value in the keyword correlation matrix otherwise.

3.2.2. Quality analysis module

This module proposes the query-sensitive quality analysis to assess expertise level of candidate experts. We firstly select out researcher's relevant documents (publications and projects) in terms of enriched query profile. Then, we measure the expertise level of a potential expert in terms of the quantity of documents and quality of the documents. Finally, we adopt a weighted scheme, proposed by Sun et al. (2008), to generate an expertise quality measure of overall contribution of a researcher to the field.

To measure the expertise level of a potential expert, his publications and projects are considered (Sun et al., 2008). For publications, some metrics have been offered from three aspects: the quantity of papers and the journal ranking of published papers, total citations, and social metrics (Piwowar, 2013) including like, comments, share, and collection to references. The quality score of expert for the query, $QS^{pub}(q, r)$, is expressed as:

$$\begin{aligned} QS^{pub}(q, r) = \alpha \sum_{i=1}^{4} \lambda_i^{pub} q_i^{pub} + \beta C^{pub} \\ + \gamma(L^{pub} + (CP^{pub} - CN^{pub}) \\ + S^{pub} + R^{pub}) \end{aligned} \tag{8}$$

where λ_i^{pub} is the weight of journal rank i (i.e. A, B, C and D) and q_i^{pub} is researcher's total number of publications in journal rank i. C^{pub} is the total number of paper citations. L^{pub}, CP^{pub}, CN^{pub}, S^{pub} and R^{pub} are social metrics. L^{pub} is the total number of like among published papers; CP^{pub} and CN^{pub} represent the total number of positive comments and negative comments respectively; S^{pub} is the total number of share among published papers; R is the total number of papers collected to references.

For projects, some metrics have also been suggested from three aspects: the quantity of projects and the project weight, the number of members joining in project group, and social metrics including like, comments, and share. The quality score of expert for the query, $QS^{pro}(q, r)$, is expressed as:

$$QS^{pro}(q, r) = \alpha \sum_{i=1}^{4} \lambda_i^{pro} q_i^{pro} + \beta G^{pro} + \gamma(L^{pro} + (CP^{pro} - CN^{pro}) + S^{pro}) \tag{9}$$

where λ_j^{pro} the weight of project type j (i.e. Nation, Ministry, Province, and City) and q_i^{pro} is researcher's total number of projects

in project type j. G^{pro} is the total number of members in project groups. L^{pro}, CP^{pro}, CN^{pro}, and S^{pro} are social metrics. L^{pro} is the total number of like among fund projects; CP^{pro} and CN^{pro} represent the total number of positive comments and negative comments respectively; S^{pro} is the total number of share among funded projects.

In terms of the quality of publications and projects of expert, the quality score of expert for the query, $QS(q,r)$, is expressed as

$$QS(q,r)=\mu QS^{pub}(q,r)+(1-\mu)QS^{pro}(q,r) \tag{10}$$

3.2.3. Connectivity analysis module

PageRank algorithm is the heart of the Google (Page et al., 1999) and this ranking mechanism has demonstrated its super performance in search engine applications. PageRank was originally designed to rank retrieval results based on the hyperlink structure of the web, which is a directed but binary graph in nature. However, the researcher collaboration network is often constructed by a weighted graph since collaboration frequency and author number should be reflected in the collaboration network. We therefore rank expert in the collaboration network by use of AuthorRank algorithm (Liu et al., 2005), which is a modification of PageRank considering link weight.

This module proposes connectivity analysis to rank expert candidates by the assumption that researchers accumulate expertise from collaborating with other experts. Firstly, relevant documents (publications and projects) are retrieved with respect to user queries. The article or project which contains one or more keyword from the user query profile or enriched query profile can be considered relevant document. Secondly, we can construct researcher collaboration network based on article co-authorships and project cooperation relationships. Therefore, the corresponding connectivity matrix C is generated. The edge weight in the connectivity matrix is defined according to Deng et al. (2012),

$$C_{ij}=\sum_{k=1}^{N}\frac{\delta_i^k\delta_j^k}{n_{d_k}-1} \tag{11}$$

where $\delta_i^k=1$ if researcher i is one of the collaborators in documentd_k, $\delta_i^k=0$ otherwise, and N is the number of documents and n_{d_k} is the number of authors in document d_k. The link weights express how strongly related two nodes, and these weights can therefore be used to determine the amount of AuthorRank that should be transferred from one node to the nodes it connects to. Finally, the AuthorRank value of an author i is then given as follows:

$$AR(i)=(1-d)+d\sum_{j=0}^{n}AR(j)\times c_{ji} \tag{12}$$

where d is the damping factor, $AR(j)$ corresponds to the AuthorRank of the backlinking node, and c_{ji} corresponds to the edge weight between node j and i. The AuthorRank can be calculated with the same iterative algorithm used by PageRank. AuthorRank can be considered as a generalization of PageRank by substituting c_{ji} with$1/D(j)$ in PageRank, in which $D(j)$ is defined as the number of links going out of node j. In the Section 3.4, we will implement the AuthorRank algorithm by the MapReduce framework. Therefore, the connectivity score of expert for the query, $CS(q,r)$, is expressed as follows:

$$CS(q,r)=AR(r) \tag{13}$$

3.3. Ranking

For each query asked by the users in research communities, the system outputs a list of recommended experts. To provide expert recommendation effectively and efficiently, two-stage recommendation strategy is employed as shown in Fig. 1. In the first stage, initial results are output by matching researcher profile with semantic query profile and lots of irrelevant researchers can be filtered out. In the second stage, the relevance score, quality score, and connectivity score, which are derived from the former analysis modules, need to be further aggregated with appropriate weighting strategy. After aggregation, the final expert ranking list can be more appropriate and accurate. Among aggregation models, analytic hierarchy process (AHP) is one well known method to solve multi-criteria decision-making (MCDM) problems, which determinate the relative importance or weight of criteria by mathematical pair-wise comparison (Dyer and Forman, 1992). It has been applied extensively in many research fields, such as social network analysis and recommendation systems. In this research, we choose AHP approach as our aggregation model.

With the collected information, expert's performances of relevance, quality and connectivity can be measured by AHP. Each expert's overall performance integrated relevance, quality and connectivity can be achieved by weighted geometric average method. So the task is to determine the relative importance of these three criteria by decision makers (DM). A panel of DMs has been organized to give pair-wise comparison matrices for the criteria. The matrices for three indicators are shown as follows:

$$\begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix} \tag{14}$$

where v_{ij} is the proportion of two criteria. According to the algorithm of AHP, the weights on each criterion (v_1, v_2, v_3) can be obtained, and the performance of expert can be computed as follows:

$$S(q,r)=v_1RS(q,r)+v_2QS(q,r)+v_3ES(q,r) \tag{15}$$

According to the ranking of $S(q,r)$, the appropriate experts can be selected in research community.

3.4. Computing by MapReduce

In this work, we implement the MapReduce framework on Hadoop, a cloud computing infrastructure, to do the large-scale computation tasks involved in expert finder system. MapReduce is a popular framework for data-intensive parallel computation in shared-nothing clusters of machines. It has been successfully applied in many applications such as indexing crawled documents, analyzing web access logs, machine learning (Cheng et al., 2012; Elsayed et al., 2008). MapReduce is a distributed computing paradigm based on two higher-order functions: map and reduce. The map function applies a user-defined function to each key–value pair in the input and generates a list of intermediate key–value pairs. These generated pairs are then sorted and grouped by the key and are further passed as inputs to the reduce function. The reduce function applies a second user-defined function to every intermediate key and all its associated values, and produces the final result. The signatures of the functions that compose the phases of the MapReduce computation are as follows:

$$\text{Map}: (k_1,\ v_1)\rightarrow[(k_2,\ v_2)] \tag{16}$$

$$\text{Reduce}: (k_2,\ [v_2])\rightarrow[(k_3,\ v_3)] \tag{17}$$

Where $(k_1,\ v_1)$ are the original key and value of a record, which are transformed to $(k_2,\ v_2)$by the Mappers. Finally, these intermediate key–value pairs are aggregated together by the Reducers on the same key k_2, further resulting in final outputs $(k_3,\ v_3)$. The illustration of MapReduce framework is shown in Fig. 2.

In this research, MapReduce is suggested to improve the efficiency of expert finding process in research community. Now we employ MapReduce framework to calculate document similarity as an example. In order to compute $sim(x,y)$ for all the document pairs in a batch mode, we first build an inverted index for all keywords in the

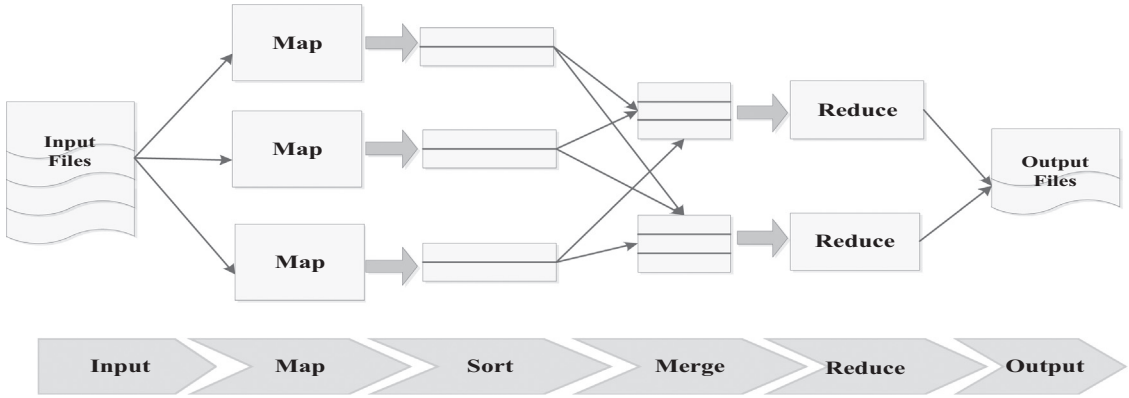

Fig. 2. The MapReduce framework.

vocabulary. For each keyword k, there is a corresponding posting in the inverted index Ind:

$$< (d_1, c_{k,d_1}), (d_2, c_{k,d_2}), \ldots, (d_i, c_{k,d_i}) \ldots > \quad (18)$$

where $d_i$is a document's profile and $c_{k,d_i}$is the weight. Then, we generate a mapper for each pair of documents in each inverted index posting. Finally, all of the intermediate results calculated by these mappers are aggregated by the reducers. We summarize these steps in the following algorithm, and shown in Table 1.

It is easy to see that the same algorithm can be employed for computing similarities for keywords. Furthermore, researcher AuthorRank value can be computed via MapReduce. The computation process mainly has three steps: calculating author rank, calculating outlinks and sorting the results.

Table 1
Document similarity computation algorithm via MapReduce.

Input: Inverted index *Ind*
Process:
Initialize *sim(x,y):*
$sim(x,y) := 0, \forall x, y \in D$
For all $k \in V$ Do
For all$x,\ y \in p(w)$ Do
Map:$map(key := x:y,\ v = x,y) \rightarrow < key := x:y,\ v' = c_{k,x} \cdot c_{k,y} >$
For all $x, y \in D$ Do
Reduce: $sim(x,y) := \sum_{key = x:y} v_{key}'$
Output: $sim(x,y)$ for all $x, y \in D$

4. System implementation

ScholarMate is an online professional social network community platform, particularly developed for academic researchers by the authors' team. It aims to foster a knowledge sharing cyberspace for researchers to collect and share different kinds of resources (e.g., publications, research progress reports). Different from other online researcher communities which require researchers to input research outputs manually, it can automatically collect a particular researcher's outputs from various sources, like CNKI[2], ISI[3] and Scopus.[4] On the ScholarMate, researchers can add other researchers into their contact list as friends. Besides, researchers with similar interests can collaborate via self-organized special interest group (SIG) functions. They can share their professional works in terms of publications, projects, papers with other community members and their friends, and receive comments and suggestions. The proposed approach is implemented as one of the application services in ScholarMate. The system provides main functions to collect and extract researcher related data including publications, projects and academic relations. Fig.3 presents the interfaces of researcher homepage and expert finder application.

5. Experimental evaluation

5.1. Data and methodology

In this section, we empirically evaluate the accuracy and scalability of our proposed approach. For the accuracy comparison, we compared our method with two representative expert finder methods in the literature. They are listed as follows:

(1) Model 2 Method (abbreviated M2): this is the state-of-the-art expert finder method, which was proposed in Balog et al. (2009). M2 employs language model to retrieved relevant documents and returns candidates associated with relevant documents as possible experts. It has been empirically demonstrated its better performance than other methods (Balog et al., 2009).
(2) Semantic Model 2 Method (abbreviated SM2): this is the enhanced method of M2 by using our pre-computed keyword correlation matrix. SM2 considers keyword semantics in order to find more relevant experts.
(3) Research Analytic Framework Method (abbreviated RAF): this is our proposed method. It leverages relevance, quality and connectivity to find domain experts.

Technically, three recommendation methods are implemented in the same experimental condition and our experiments are conducted in a Hadoop platform. Since it is difficult to objectively evaluate the relevance of experts for the given query, we conducted a subjective user study to validate the effectiveness of three expert finder methods according to human perception of relatedness. We used 30 representative topic queries as the source queries (i.e., the selected queries which are supported to expert finder systems and used to return relevant experts). In order to compare the performance of our proposed method with two baseline methods, for each query, we returned three different expert lists with each containing 10 experts. However, as different expert finding schemes my return identical expert, there are quite a few duplicates. We invited 15 evaluators including 5 Ph.D. students and 10 postgraduate students to assess the expert finder performance. For each query in the selected 30 source queries, the subject was asked to first read the source query and get familiar with it. Then, the subject was provided with the computed experts returned by three schemes (in a mixed manner) in a random

[2] http://www.cnki.net/.
[3] http://portal.isiknowledge.com.
[4] http://www.scopus.com.

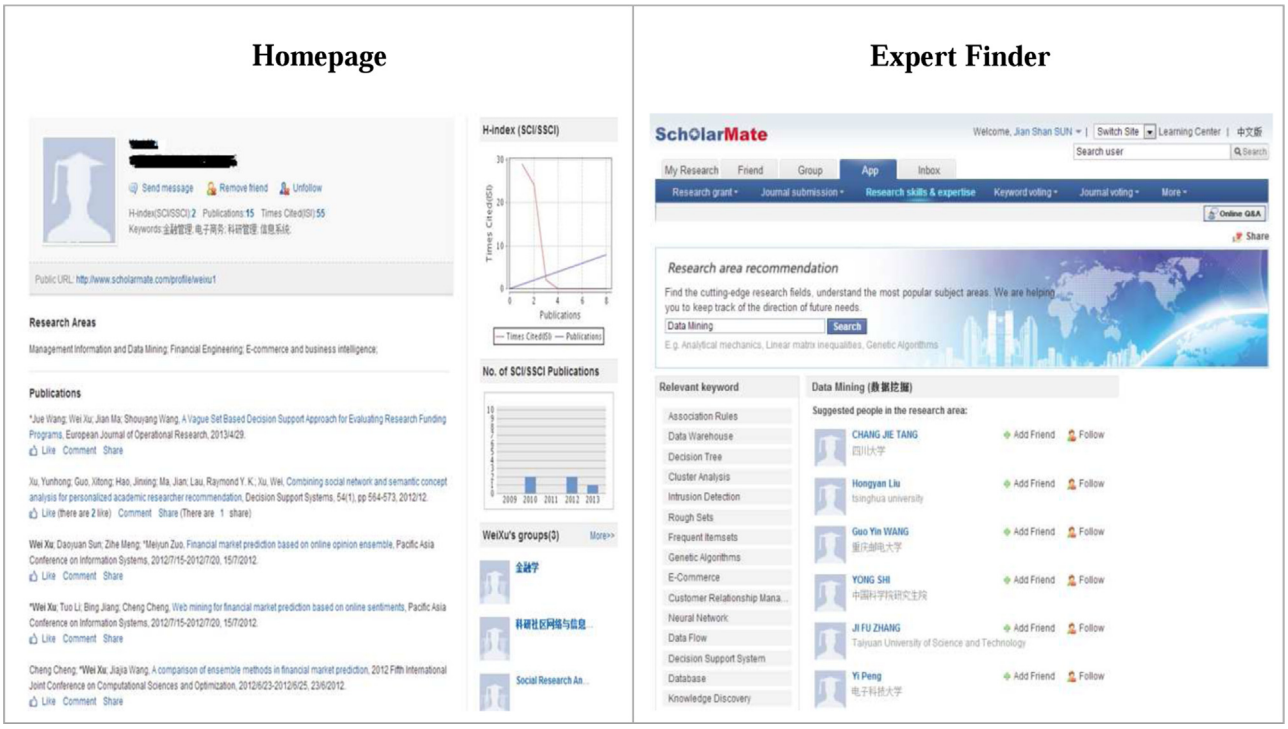

Fig. 3. Researcher homepage and expert finder interface in ScholarMate.

order. The subjects did not know by which scheme the current expert was returned ensuring that the rank of a recommendation did not influence subjects' perception. For each query, the subject rated each returned expert on a 5-point Likert (Likert, 1932) scale ranging from 1 to 5. Notice that the one mark means that the subject is not interested at all in the recommended opportunity: this recommendation is not relevant. On the contrary, a recommendation is all the more relevant as the subject's mark is high. Each subject was assigned 6 source queries so that each source query and its returned experts were evaluated at three times ($6 \times 15/30 = 3$). As a result, we could use the obtained feedback data to evaluate the effectiveness and accuracy of our proposed approach.

To verify the scalability of the proposed method, RAF, we employ two synthetic datasets (200M and 500M) to conduct a series of experiments. These datasets are built by extracting information from ScholarMate and they contain all the information related to experts (such as publications, projects, collaboration networks and so on) in the experiments. The experiment is conducted respectively in a cluster of nodes ranging from 1 to 8. Therefore, the execution time can be recorded to measure the scalability of the MapReduce algorithm.

5.2. Evaluation metrics

Similar to traditional recommendation and search system, we recommend a list of experts based on researchers' search queries and ask them to rate recommendations.

To verify the accuracy of proposed recommendation method, the Average Rating score (AR) and Normalized Discounted Cumulative Gain (NDCG) are selected as the performance metrics (Adomavicius and Tuzhilin, 2005). These metrics are computed over the top 1 and 5 recommended experts. AR is computed among the ratings from all the users and it indicates the average rating of all the recommendations. NDCG is a commonly adopted metric for evaluating a search engine's performance and it is for gradual judgments (i.e., documents are non-relevant or more or less relevant to the query). They are defined as follows:

$$\mathrm{AR} = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{1}{N}\sum_{j=1}^{N} r_{ij} \tag{19}$$

$$\mathrm{NDCG} = \frac{1}{|\mathrm{U}|}\sum_{i=1}^{|\mathrm{U}|} Z\sum_{j=1}^{N}\frac{2^{r_{ij}}-1}{\log(1+j)} \tag{20}$$

where $|\mathrm{U}|$ denotes the number of researchers in the survey; N is the number of recommended project opportunities and in this setting, $N=1$ or 5;r_{ij} represents the rating of researcher i on project opportunity j; Z is a normalization constant and is chosen so that a perfect ranking's NDCG value is 1.

To measure the performance of the scalability of the proposed method, we employ the Speedup metric, which is a widely used scalability metric to demonstrate how much a parallel algorithm is faster than a corresponding sequential algorithm (Amdahl, 1967; Meng et al. 2014). It is defined as follows:

$$S_p = \frac{T_s}{T_p} \tag{21}$$

where S_p the Speedup metric, T_s denotes the sequential execution time, T_p is the parallel execution time and p is the number of processors. With the fixed dataset, if the Speedup, S_p, has a linear relation with the number of nodes, we can consider that the algorithm will have good scalability.

5.3. Results and analysis

In this section, we present the accuracy and scalability results from experiments. In accuracy, the performance comparison of M2, SM2 and RAF methods is listed in Table 2. The robustness of this practice was shown by Buckley and Voorhees (2000) that evaluating the search engine should guarantee at least $n=25$ queries. We have 30 valid responses, which is more robust for statistical analysis.

Table 2
Evaluation and comparison of different recommendation methods.

		M2	SM2	RAF	Improvements over best baseline (%)
AR	Top 5	3.59	3.63	3.90	7.4*
	Top 10	3.31	3.49	3.74	7.2**
NDCG	Top 5	0.63	0.60	0.70	11.1**
	Top 10	0.58	0.64	0.74	15.6**

* P value significant at alpha=0.05.
** P value significant at alpha=0.01.

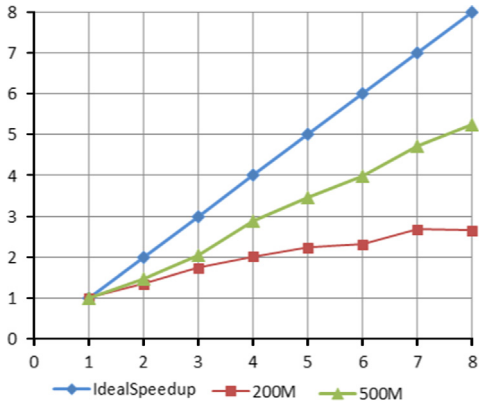

Fig. 4. Speedup performance of RAF.

We can see that our proposed method achieves the best performance in terms of both the AR metric and the NDCG metric. The AR scores obtained by using the M2 method and the SM2 method are 3.59 and 3.63in terms of returning top five experts and these results are acceptable. However, our proposed method, RAF, achieves more than 7.4% improvements over baselines. The improvements of AR at top 10 returned experts also indicate that our method can find more relevant experts than the other two methods. We further evaluate the rank performance of the three methods. NDCG scores reflect the browsing efforts of the researchers before locating the relevant experts. In terms of NDCG values, the proposed method achieves over 11.1% improvements when returning top 5 experts and over 15.6% improvements when returning top 10 experts. The improvements on the NDCG value clearly show that our method is more effective than the M2 and SM2 methods, which gives a higher ranking for finding relevant experts.

Moreover, we test the improvement significance of the results of the proposed method over baseline methods by means of the paired t-tests. The statistical significance of an improvement is represented as a p-Value with Student's paired bilateral t-test (Student, 1908). Although requiring a normal data distribution in theory, Hull (1993) points that it is robust to violations of this requirement in practice. Moreover, previous work shows that this test to be more accurate than other ones, such as Wilcoxon's signed rank test (Sanderson and Zobel, 2005). When $p < a$, with $a=0.05$ difference between the two systems is deemed to be statistically significant. The smaller the p-Value, the more significant the difference. From Table 2, it is indicated that improvements of our approach in AR and NDCG are all statistically significant.

In scalability, the Speedup performance is demonstrated in Fig. 4. We can observe that the speedup of RAF increases relatively linearly when the number of processor nodes increases. Furthermore, when the parallel algorithm is applied in larger dataset (500M), it can obtain better speedup. With 8 nodes in 500M dataset, the speedup value reaches 5.23, which is 65.4% (5.23/8=65.4%) of the Ideal Speedup (Let S_p be the speedup for p processors. Ideal Speedup or linear speedup is obtained when $S_p = p$. When running an algorithm with linear speedup, doubling the number of processors doubles the speed. As this is ideal, it is considered very good scalability.) The experimental result indicates that our proposed method RAF on Map-Reduce performs better with larger dataset and has good scalability over big data context.

In general, these experimental results show that our proposed RAF method outperforms baseline methods and achieve higher accuracy, and RAF on MapReduce framework has good scalability in big data environment. Our proposed RAF framework can be leveraged with big data analytics tools to find domain expert in research social network services effectively and efficiently.

6. Conclusions and future work

In this paper, a novel research analytics framework approach combining relevance, quality and connectivity is proposed for expert finding in research communities. Expertise profile of researchers is built from three aspects: topic relevance, expert quality and researcher connectivity. The effectiveness and efficiency of proposed approach is verified in the designed experiment. The proposed algorithm and designed recommender system have been incorporated into the existing research social network websites to facilitate expert search and potential collaboration. Also, the algorithm and system can be generalized to other personalization applications in digital libraries and other scientific websites.

Several research questions can be further investigated in future. Firstly, we compute keyword similarity to expand query profile. We are aware that the use of domain ontology will greatly help to resolve semantic ambiguity in keyword matching. Thus in future, research domain ontology can be constructed to support extended profile matching. Secondly, this paper adopts AHP technique as the rank aggregation method. Possibly, other data fusion techniques can be considered such as Condorcet fusion (Montague and Aslam, 2002) and other techniques that model score distribution (Nandakumar et al., 2008). Thirdly, it is better to consider other human factors when recommending experts for given query topics. Therefore, we will extend this work by considering these factors in the future.

Acknowledgments

This work was supported in part by 973 Project (Grant no. 2012CB316205), National Natural Science Foundation of China (Nos. 71001103, 71171172, 91224008, 71371062, 91324015, and 71490725), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No. Z131101000413058), Program for Excellent Talents in Beijing, General Research Fund of Hong Kong (No. CityU 148012, CityU 119611), and CityU Teaching Development Grant (no. 6000201).

References

Adomavicius, G., Tuzhilin, A., 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. 17, 734–749.
Afzal, M.T., Maurer, H., 2011. Expertise recommender system for scientific community. J. Univers. Comput. Sci. 17, 1529–1549.
Amdahl, G.M., 1967. Validity of the single processor approach to achieving large scale computing capabilities. In: Proceedings of the Spring Joint Computer Conference. ACM, April 18–20, pp. 483–485.
Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., 2010. A view of cloud computing. Commun. ACM 53, 50–58.
Balog, K., Azzopardi, L., de Rijke, M., 2009. A language modeling framework for expert finding. Inf. Process. Manag. 45, 1–19.
Balog, K., Bogers, T., Azzopardi, L., De Rijke, M., Van Den Bosch, A., 2007. Broad expertise retrieval in sparse data environments. In: Proceedings of the 30th

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 551–558.
Brian, H., Brunschwiler, T., Dill, H., Christ, H., Falsafi, B., Fischer, M., Grivas, S.G., Giovanoli, C., Gisi, R.E., Gutmann, R., 2008. Cloud computing. Commun. ACM 51, 9–11.
Buckley, C., Voorhees, E.M., 2000. Evaluating evaluation measure stability. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 33–40.
Cao, Y., Liu, J., Bao, S., Li, H., 2005. Research on expert search at enterprise track of TREC 2005, Proc. of TREC.
Chen, H., Chiang, R.H., Storey, V.C., 2012. Business intelligence and analytics: from big data to big impact. MIS Q. 36, 1165–1188.
Cheng, Y., Qin, C., Rusu, F., 2012. GLADE: big data analytics made easy. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, pp. 697–700.
Craswell, N., de Vries, A.P., Soboroff, I., 2005. Overview of the trec-2005 enterprise track, TREC 2005 Conference Notebook, pp. 199–205.
Deng, H., Han, J., Lyu, M.R., King, I., 2012. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, pp. 71–80.
Deng, H., King, I., Lyu, M.R., 2008. Formal models for expert finding on DBLP bibliography data. In: Proceedings of the Eighth IEEE International Conference on Data Mining. IEEE, pp. 163–172.
Dom, B., Eiron, I., Cozzi, A., Zhang, Y., 2003. Graph-based ranking algorithms for e-mail expertise analysis. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. ACM, pp. 42–48.
Dyer, R.F., Forman, E.H., 1992. Group decision support with the analytic hierarchy process. Decis. Support Syst. 8, 99–124.
Elsayed, T., Lin, J., Oard, D.W., 2008. Pairwise document similarity in large collections with MapReduce. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pp. 265–268.
Fang, H., Zhai, C., 2007. Probabilistic models for expert finding, Advances in Information Retrieval. Springer, Berlin, Heidelberg, pp. 418–430.
Gollapalli, S.D., Mitra, P., Giles, C.L., 2011. Ranking authors in digital libraries. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. ACM, pp. 251–254.
Hofmann, K., Balog, K., Bogers, T., De Rijke, M., 2010. Contextual factors for finding similar experts. J. Am. Soc. Inf. Sci. Technol. 61, 994–1014.
Hull, D., 1993. Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 329–338.
Karimzadehgan, M., Zhai, C., Belford, G., 2008. Multi-aspect expertise matching for review assignment. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM, pp. 1113–1122.
Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 604–632.
Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 140, 1–55.
Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H., 2005. Co-authorship networks in the digital library research community. Inf. Process. Manag. 41, 1462–1480.
Macdonald, C., Ounis, I., 2008. Voting techniques for expert search. Knowl. Inf. Syst. 16, 259–280.
Meng, S., Dou, W., Zhang, X., Chen, J., 2014. KASR: A keyword-aware service recommendation method on MapReduce for big data application. IEEE Transactions on Parallel and Distributed Systems 25 (12), 3221–3231.
Montague, M., Aslam, J.A., 2002. Condorcet fusion for improved retrieval. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management. ACM, pp. 538-548.
Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.K., 2008. Likelihood ratio-based biometric score fusion. IEEE Trans. Pattern Anal. Mach. Intell. 30, 342–347.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank citation ranking: bringing order to the web.
Park, Y.-J., Chang, K.-N., 2009. Individual and group behavior-based customer profile model for personalized product recommendation. Expert Syst. Appl. 36, 1932–1939.
Piwowar, H., 2013. Value all research products. Nature 493, 159.
Quattrone, G., Capra, L., De Meo, P., Ferrara, E., Ursino, D., 2011. Effective retrieval of resources in folksonomies using a new tag similarity measure. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, pp. 545–550.
Sanderson, M., Zobel, J., 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 162–169.
Serdyukov, P., Hiemstra, D., Fokkinga, M., Apers, P.M., 2007. Generative modeling of persons and documents for expert search. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 827–828.
Serdyukov, P., Rode, H., Hiemstra, D., 2008. Modeling multi-step relevance propagation for expert finding. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM, pp. 1133–1142.
Shami, N.S., Ehrlich, K., Millen, D.R., 2008. Pick me!: link selection in expertise search results. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 1089–1092.
Shi, L., Que, J., Zhong, Z., Meyer, B., Crenshaw, P., He, Y., 2013. A scalable implementation of malware detection based on network connection behaviors. In: Proceedings of the 2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). IEEE, pp. 59–66.
Silva, T., Guo, Z., Ma, J., Jiang, H., Chen, H., 2013. A social network-empowered research analytics framework for project selection. Decis. Support Syst. 55, 957–968.
Student, A., 1908. The probable error of a mean. Biometrika 6, 1–25.
Sun, J., Ma, J., Liu, Z., Miao, Y., 2014. Leveraging content and connections for scientific article recommendation in social computing contexts. Comput. J. 57, 1331–1342.
Sun, Y.-H., Ma, J., Fan, Z.-P., Wang, J., 2008. A group decision support approach to evaluate experts for R&D project selection. IEEE Trans. Eng. Manag. 55, 158–170.
Wang, G.A., Jiao, J., Abrahams, A.S., Fan, W., Zhang, Z., 2013. ExpertRank: a topic-aware expert finding algorithm for online knowledge communities. Decis. Support Syst. 54, 1442–1451.
Wu, H., Hua, Y., Li, B., Pei, Y., 2013. Towards recommendation to trust-based user groups in social tagging systems. In: Proceedings of the 10th International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, pp. 893–897.
Zhang, J., Tang, J., Li, J., 2007. Expert finding in a social network, Advances in Databases: Concepts Systems and Applications. Springer, Berlin, Heidelberg, pp. 1066–1069.
Zhou, D., Orshanskiy, S.A., Zha, H., Giles, C.L., 2007. Co-ranking authors and documents in a heterogeneous network. In: Proceedings of the Seventh IEEE International Conference on Data Mining. IEEE, pp. 739–744.