Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

A TOPIC SENSITIVE SIMRANK (TSSR) MODEL FOR EXPERTS FINDING ON ONLINE RESEARCH SOCIAL PLATFORMS

Wenping Zhang
City University of Hong Kong, wzhang23-c@my.cityu.edu.hk

Liying Ye
City University of Hong Kong, liyingye2-c@my.cityu.edu.hk

Wei Du
City University of Hong Kong, weidu7-c@my.cityu.edu.hk

Jian Ma
City University of Hong Kong, isjian@cityu.edu.hk

Wei Xu
Renmin University of China, weixu@ruc.edu.cn

See next page for additional authors

Recommended Citation

Zhang, Wenping; Ye, Liying; Du, Wei; Ma, Jian; Xu, Wei; and Tang, Shengtao, "A TOPIC SENSITIVE SIMRANK (TSSR) MODEL FOR EXPERTS FINDING ON ONLINE RESEARCH SOCIAL PLATFORMS" (2016). *PACIS 2016 Proceedings*. 306.
http://aisel.aisnet.org/pacis2016/306

Authors
Wenping Zhang, Liying Ye, Wei Du, Jian Ma, Wei Xu, and Shengtao Tang

A TOPIC SENSITIVE SIMRANK (TSSR) MODEL FOR EXPERTS FINDING ON ONLINE RESEARCH SOCIAL PLATFORMS

Wenping Zhang, City University of Hong Kong, wzhang23-c@my.cityu.edu.hk

Liying Ye, City University of Hong Kong, liyingye2-c@my.cityu.edu.hk

Wei Du, City University of Hong Kong, weidu7-c@my.cityu.edu.hk

Jian Ma, City University of Hong Kong, isjian@cityu.edu.hk

Wei Xu, Renmin University of China, weixu@ruc.edu.cn

Shengtao Tang, China Academy of Information and Communications Technology, tangshengtao@caict.ac.cn

Abstract

As an efficient online academic information repository and information channel with crowds' contribution, online research social platforms have become an efficient tool for various kinds of research & management applications. Social network platforms have also become a major source to seek for field experts. They have advantages of crowd contributions, easy to access without geographic restrictions and avoiding conflict of interests over traditional database and search engine based approaches. However, current research attempts to find experts based on features such as published research work, social relationships, and online behaviours (e.g. reads and downloads of publications) on social platforms, they ignore to verify the reliability of identified experts. To bridge this gap, this research proposes an innovative Topic Sensitive SimRank (TSSR) model to identify "real" experts on social network platforms. TSSR model includes three components: LDA for Expertise Extension, Topic Sensitive Network for Reputation Measurement, and Topic Sensitive SimRank for unsuitable experts detection. We also design a parallel computing strategy to improve the efficiency of the proposed methods. Last, to verify the effectiveness of the proposed model, we design an experiment on one of the research social platforms-ScholarMate to seek for experts for companies that need academic-industry collaboration.

Keywords: Experts finding, Social platform, Topic sensitive SimRank, Social network analysis.

1 INTRODUCTION

With the development of Web 2.0 and Science 2.0, online research social network platform plays a significant role in expert finding applications. On social network platforms such as Facebook[1], LinkedIn[2] and ScholarMate[3], enormous amount of user-generated contents (UGC) provide rich information about people's research work, social activities and collaboration relationships. Besides the role of information repository, online research social network platforms also act as efficient information channels that spread R&D news quickly. With the advantage of online social platforms in collecting and disseminating research information with crowds' contribution, researchers have begun to leverage online social platforms for experts finding (Bozzon, Brambilla, Ceri, Silvestri, & Vesci, 2013; Davoodi, Kianmehr, & Afsharchi, 2013; X. Liu, Wang, Johri, Zhou, & Fan, 2014; Sun, Xu, Ma, & Sun, 2015).

Compared with the traditional way, using online social platforms to find experts is more efficient and effective. First, online social platforms provide a channel where you can find experts all over the world without geographical barrier. Second, online social platforms facilitate the communication. People can easily get access to experts and analyse different responses in order to find the most suitable ones. Last, online social platforms can reduce the cost of maintaining local databases of experts for certain emergencies such as natural disasters (e.g. earthquake, hurricane, etc.) and terrorist attacks, or at least, supplement the limited local databases when necessary. As we know, local expertise databases are only utilized when there is a need, and the local information may not be updated in time. Based on these advantages, finding experts through online social platforms is generally accepted by the public now. For example, LinkedIn, as a professional networking site, is a good place for employees and headhunting firms to find field experts, and ScholarMate also provides a function specifically for finding experts from various domains.

However, tons of information provided by millions of users on an online research social platform confuses people when seeking aids and solutions from experts. How to effectively and efficiently identify a real expert regarding a specific domain/topic has become a challenging task. Especially due to the low information quality on social platforms, the verification of the experts is critically important.

Previous research focuses on finding experts based on features such as published research work, social relationships, and online behaviours (e.g. reads and downloads of publications) on social platforms for specific purposes, while no further measures are designed to verify the reliability of identified experts. The unsuitable experts can be experts who are loosely connected with other experts and thereby influence the future teamwork, and can even be some fraud information maker due to the lack of review on low online information quality. Identifying them is critically important to improve the effectiveness of experts finding. Current research aims to find the experts with relevant domain knowledge, while ignore the implicit relationship among researchers in the same domain. According to Role Equivalence theory (Guler, Guillén, & Macpherson, 2002), actors who play similar roles tend to have similar kinds of relationships with the same set of third parties. Even if they are not cohesive, they tend to behave alike and substitutes for each other in the social structure. These actors are said to be role equivalent. Role Equivalence theory lays foundation for us to identify unsuitable experts who are assumed to be loosely connected with "real" experts. In this paper, we are trying to distinguish such unsuitable experts from the real ones by integrating topic modelling and social network analysis to further improve the effectiveness of experts finding. Topic Sensitive SimRank (TSSR) model is

[1] https://www.facebook.com/

[2] https://www.linkedin.com/

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

proposed to identify reliable experts on social platforms. The model aims to solve two critical problems in expertise seeking on social platforms: 1) who has relevant expertise? and 2) who are probably unsuitable experts loosely connected with the domain community?. TSSR model includes three components: *LDA for Expertise Extension, Topic Sensitive Network for Reputation Measurement, and Topic Sensitive SimRank for unsuitable experts detection*. We will test the proposed model on one of the biggest research social platforms in China: ScholarMate. The dataset is collected and described briefly in this paper. The results will be listed in our future research.

The rest of paper is organized as follows. Section 2 provides the related work including topic models and social network analysis techniques. Section 3 gives the details of the proposed model. We introduce the experiment design and dataset in Section 4. This research concludes with a discussion of the research and future improvements.

2 RELATED WORK

2.1 Topic models

Topic models are probabilistic generative models to represent and analyse semantic structure of textual corpora based on hierarchical Bayesian analysis. One of the most famous and widely used topic models is Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003) based on latent semantic indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and probabilistic LSI (Hofmann, 1999). The basic idea of LDA is to model documents as an infinite mixture of topics (i.e. multinomial distribution), where topics refer to a distribution of a fixed vocabulary of terms from these documents. To generate a document, LDA firstly draws a topic from the topics and then draws a term from the collection of terms in the selected topic. Compared with the classical LSI and probabilistic LSI, LDA supports semantically richer representations of high-level concepts. As a result, LDA has been widely used in various tasks. For instance, Wei et al. applied LDA-based document models to enhance ad-hoc information retrieval (Wei & Croft, 2006). Krestel used LDA to recommend tags of resources to improve the web search (Krestel, Fankhauser, & Nejdl, 2009). Rasiwasia adopted LDA for image classification (Rasiwasia & Vasconcelos, 2013). Due to its high scalability and characteristics of unsupervised learning, LDA has also become one of the hot topics in big data analysis. For instance, Tirunllai and Tellis applied LDA to extract latent dimensions of consumer satisfaction from large scale user-generated contents (e.g. product reviews) (Tirunillai & Tellis, 2014).

2.2 Social network analysis

With the popularity of social media, social network analysis has been widely used in many applications. For example, Sun et al. (2015) leveraged research social network to find domain experts with RAF (Research Analytics Framework) (Sun et al. 2015), Xu et al. (2012) combined social network analysis and semantic concept analysis to recommend academic researchers (Xu et al., 2012), and Davoodi et al. (2013) proposed an expert recommender system by using semantic social network analysis (Davoodi et al. 2013). Social network analysis aims to mine useful information from constructed graph where nodes represent objects and links represent their relationships. We divide the existing measurement for social network analysis into three streams: SNA-network, SNA-node, and SNA-relation. SNA-network measurement such as *cohesion*, *density*, and *centralization* describes the attributes of the whole network (Haythornthwaite, 1996). SNA-node measures the attributes of network on the node level. Node centrality is a typical SNA-node measure that identifies the most important nodes in a network (Brandes, 2001; Freeman, 1977; Opsahl, Agneessens, & Skvoretz, 2010). According to the different meaning of "importance", node centrality can be further defined as *betweenness centrality*, *closeness centrality*, *degree centrality*, *Eigenvector centrality*, and *Katz centrality*. SNA-relation aims to quantify the relation (e.g. closeness or similarity) between two nodes in a connected network. Typical measures such as Jaccard's coefficient (Leicht, Holme, & Newman, 2006), SimRank and its variants (Jeh & Widom, 2002; H. Liu, He, Zhu, Ling, & Du, 2013), Katz similarity (Katz, 1953), and algebraic distance (Chen & Safro, 2011) belong to the third stream.

Among the social network analysis measures, SimRank is more and more popular because it can be applied in any field to measure the closeness/similarity of two nodes in a network. SimRank measure is based on the simple intuition that "two nodes are similar if they are linked to similar nodes" (Jeh & Widom, 2002). However, it has several drawbacks such as high computation costs, failure in similarity calculation of directed neighbors, not applicable in heterogeneous context and uniform link weights. Variants of SimRank are proposed to improve SimRank in different aspects. SimRank redefines similarity in weighted network (Antonellis, Molina, & Chang, 2008). SimFusion (Xi et al., 2005) and LinkClus (Yin, Han, & Yu, 2006) are applicable in heterogeneous network that include different types of objects and relationships. BlockSimRank (Li, Cai, Liu, He, & Du, 2009) leverages the block structure of network, and improves the computational efficiency by computing the similarity between blocks and objects within blocks. In our research, we introduce a parallel computing strategy and design topic sensitive SimRank method due to the low efficiency of SimRank on a large network.

3 PROPOSED MODEL

We propose a Topic Sensitive SimRank (TSSR) model to identify reliable on social platforms based on topic modelling and social network analysis. Intuitions behind the model design are: 1) experts' research work (i.e. research projects, research publications, patents, etc.) will harvest a crowd of online attention such as like, comment and forward via research social platforms, and 2) a potential expert has high probability being a reliable one if one's characteristics are similar or close to other identified experts. Based on this, the proposed model mainly includes three components as shown in Figure 1, namely *expertise extension*, *reputation measurement*, and *unsuitable experts detection*.

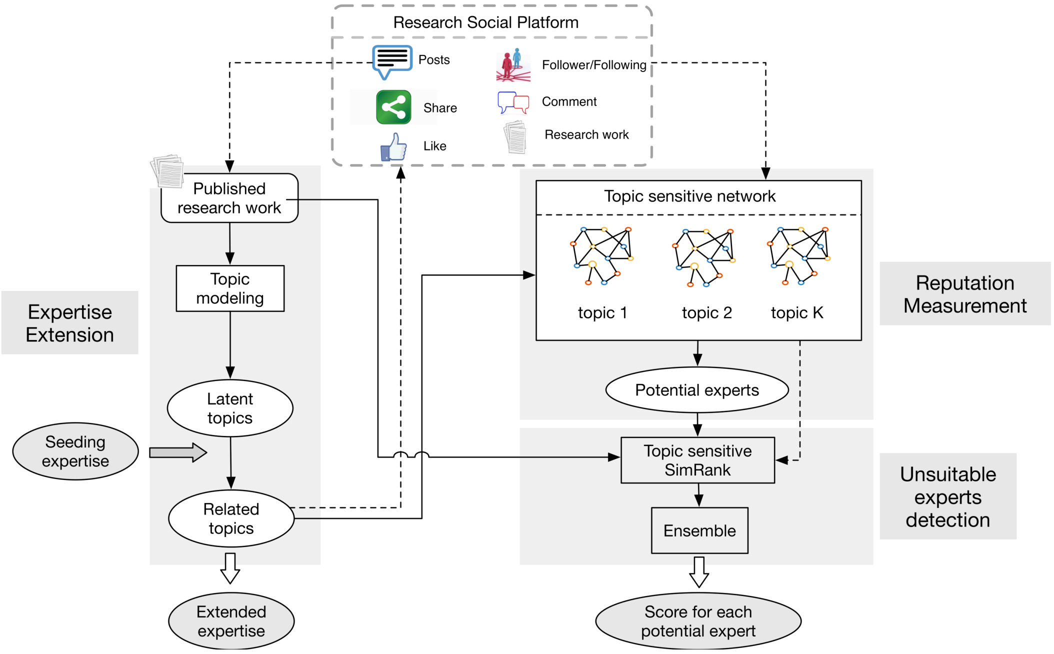

Figure 1. Framework of Topic Sensitive SimRank Model

To perform experts finding tasks, a set of expertise/keywords is provided. Even not, the target project/research that need experts can also be transformed into a set of expertise.

The provided expertise is often limited. First, to extend manually provided seeding expertise, we introduce a topic modelling method-LDA (Latent Dirichlet Allocation) to detect latent topics from published research work as a reliable information source. Latent topics that include keyword(s) from seeding expertise are named related topics. Related topics reveal different aspects of expertise under supervision of seeding expertise. Keywords in related topics are used to extent the seeding expertise, and are then used to identify professional content and target experts on research social platform.

Second, to measure one's reputation regarding a specific expertise, we construct a social network based on target experts' social relationship (e.g. like, comment or share on corresponding research work) and collaboration network (e.g. collaboration relationship in corresponding research work). According to the topics included in the research work, the social network can be transformed into topic sensitive sub-networks to measure one's reputation on each topic as well as one's integrated reputation. We measure one's reputation based on the centrality the target user get from other users.

Third, users with high reputation in terms of expertise may be unsuitable experts that are loosely connected with other experts. Unsuitable experts detection aims to identify the unsuitable experts by using topic sensitive SimRank based on the intuition that real experts tend to have similar profiles in both expertise and social relations. Users who have small SimRank values with other experts have high probability being unsuitable experts. Unsuitable experts detection enhances the reliability of experts finding.

Last, a parallel computing strategy is designed to reduce the computation cost when implementing the model on research social platforms. In the second component, topic sensitive sub-networks constructed on each topic are independent with each other. Given the complexity of network construction and analysis, we introduce big data analysis to improve the efficiency.

3.1 LDA for Expertise extension

To identify the potential experts, we first use a set of seeding keywords (i.e. expertise) as queries to retrieve relevant research work and corresponding users. The set of expertise is important to differentiate research work from other general documents or reports. Therefore, the first step for experts seeking is to determine the queries, namely a set of expertise that includes the potential knowledge or skills. In reality, the set of seeding expertise is provided manually by invited experts based on their experience, which is usually limited in revealing various aspects of solutions. Due to this, LDA is introduced to detect latent topics and enrich the seeding expertise by mining the corresponding research work.

Published research works are viewed as credible. To extend the seeding expertise, we implement LDA on content posted by selected N_0 published research work to detect K latent topics. Suppose the set of seeding expertise is $S = \{key_1, ..., key_{|S|}\}$. Each research work is viewed as a document.

In LDA model (Blei, Ng, & Jordan, 2003), each document (piece of content posted by verified experts) $d \in D$ is characterized by a multinomial distribution θ over a set of K topics, while each topic z can be represented by a multinomial distribution ϕ over a set of keywords W . The document-topic parameter and topic-keyword parameter are controlled by the hyper-parameter, the Dirichlet prior, α and β . The plate illustration of the LDA is shown in figure 2.

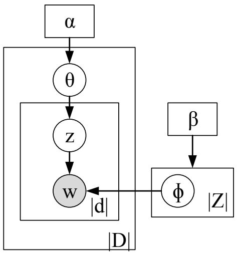

Figure 2. Illustration of LDA model

The generative process of LDA for each document d in a corpus D is summarized as firstly drawing a topic z from the document-topic distribution, then drawing a keyword w from the topic-keyword distribution. More precisely, the generative process can be formally described as:

$$p(w_i) = \sum_{k=1}^{K} p(w_i|z_k)p(z_k) \tag{1}$$

where $p(w_i)$ is the probability of keyword w_i in a document, $p(w_i|z_k)$ is the probability of sampling keyword w_i from topic z_k, and $p(z_k)$ is the probability of topic z_k chosen for keyword w_i.

According to the illustration in Figure 2, $\phi = P(W|Z) = \{p(w_i|z_k)\}$ is the topic-keyword distribution, and $\theta = P(Z|D) = \{p(z_k|d_j)\}$ is the document-topic distribution. Note that ϕ and θ are controlled by hyper-parameters α and β. In LDA we estimate the hyper-parameters α and β by using Gibbs sampling instead of estimating ϕ and θ directly. After the estimation by Gibbs sampling, we can obtain the topic-keyword distribution and the document-topic distribution. After the sampling, the keywords describing the similar concept are contextually clustered into the same topics.

The way of expanding manually selected seeding expertise is listed as follows. 1) Detect K latent topics from verified experts' posts by using LDA. 2) Extract K^* topics that are highly related with the keywords in seeding expertise $S = \{key_1,...,key_{|S|}\}$. In LDA, each topic is a distribution over words. Each word plays different roles (owns different probability) in the topic. Only topics where seeding expertise plays the dominant roles are selected as potentials. To further distinguish the related topics and unrelated topics, we treat the seeding expertise and each topic as weighted vectors. Then the highly related topics are selected from the potential ones according to their relationship (similarity) with the seeding expertise. In this way, the topics with multiple seeding expertise playing important roles will have higher probability to be selected. 3) Expand the seeding expertise by adding the new keywords from related topics. To rule out the noise words in the topics, only words with high probability in the selected topics are added as new keywords. Through these three steps, the limited seeding keywords are contextually expanded, and the result is represented as $S' = \{t_1,...,t_T\}$. These automatically contextually expanded field expertise indicators (i.e. keywords) lay the foundation for the next step.

3.2 Topic sensitive network for reputation measurement

In this section, we construct a directed network to measure users' reputation in terms of the set of generated expertise in Section 3.1. To scale down the network size, the keywords in generated expertise are first used to retrieve relevant research work as well as its authors. We name the retrieved research work as expertise related content. Besides, users' *like*, *comment* or *share* on such expertise related content are also crawled as relationships between online users. By viewing the users as nodes and users' frequency of *like*, *comment* or *share* on expertise related content as links, the directed weighted network can be constructed.

As shown in Figure 2, the directed network is transformed into topic sensitive networks by classifying the crawled content into each topic. We use a simple classification method: if a piece of content includes at least one keyword in a generated topic, then this piece of content is classified into this topic. A piece of content can be classified into several topics, which is consistent with the fact that the content may reveal different aspects of the solutions.

Let $G = (V, E)$ represent the constructed network where V represents the set of nodes and $E = \{e(v_k, v_l)\}$ denotes the set of directed relations between the nodes. We divide the constructed network into sub-networks by extracting topic related content and corresponding relations. Let $G_i = (V_i, E_i)$ $(i = 1,...,K)$ denote the sub-network on ith generated topic. We compute the reputation based on weighted degree centrality measure (Opsahl et al., 2010; Tang, Wang, Zhong, & Pan, 2014). For each user $v_m^i \in V_i$, the reputation on i th generated topic can be computed as

$$rep(v_m^i) = \sum_{v_n^i \in I(v_m^i)} \sum_{all\ j} w_j e_j(v_n^i, v_m^i) \tag{2}$$

where $I(v_m^i)$ denotes the in-neighbours of user v_m^i and $e_j(v_n^i, v_m^i) = 1$. w_j denotes the given weight for defined relations between nodes such as *like*, *comment* or *share*. The weight can be set as uniform or different according the importance of different relations.

Obviously, the integrated reputation of one user can be computed as the summation of reputation value on each topic: $total\_rep(v) = \sum_{i=1}^{K} rep(v^i)$.

3.3 Topic Sensitive SimRank for unsuitable experts detection

By ranking users' reputation in descending order, we can get a set of experts and corresponding reputation values. However, some of them are possibly unsuitable experts. Identifying them is critically important. Real experts tend to have similar profiles in both expertise and social relations. In real world, according to role equivalence theory, real experts who play the same role tend to connect to the same set of third parties with similar kinds of relationships. Even if they are not cohesive, they tend to behave alike and substitutes for each other in the social structure. Therefore, we leverage SimRank to measure the structural similarity/closeness among the identified experts. Users who have small SimRank values with others have high probability as unsuitable experts.

The SimRank is computed based on the topic sensitive network constructed in Section 3.2. Note that we only output the SimRank values between the set of experts obtained in Section 3.2. For each topic sensitive network, SimRank value between two identified experts reveals the similarity between their profiles as well as social structure regarding a specific topic. Note that SimRank is only applicable in connected network. If two identified experts are not directly/indirectly connected in a network, we view their SimRank value as 0. Links in the topic sensitive network $G_i = (V_i, E_i)$ $(i = 1, ..., K)$ here are viewed as uniform. Suppose G_i has identified experts represented as V_0. The rationale of SimRank is that two nodes are similar if they are linked to similar neighbours (Jeh & Widom, 2002). For a normal expert and each verified expert, their SimRank value can be computed recursively based on following formula:

$$sim_t(u,v) = \frac{C}{|I(u)||I(v)|} \sum_{i=1}^{|I(u)|} \sum_{j=1}^{|I(v)|} sim_{t-1}(I_i(u), I_j(v)) \tag{3}$$

where C is a parameter between 0 and 1, and $I(u)$ denote the set of in-neighbours of u. If $u = v$, $sim_t(u,v) = 1$. The similarity between u and v at t th iteration is the average similarity of their neighbours at $t-1$ th iteration multiplying parameter C.

For an identified expert, we define the topic sensitive SimRank on each topic as the average SimRank value with other identified experts. Therefore, an expert's SimRank value on i th generated topic can be represented as: $SimRank(v^i) = \frac{1}{|V_0|} \sum_{v^j \in V_0} sim(v^i, v^j)$. Topic sensitive SimRank measures one's credibility in terms of this specific topic.

The integrated SimRank of one user can be computed as the maximum SimRank values on K topics: $SimRank(v) = \max_{1 \le i \le K}(SimRank(v^i))$. Experts who have high SimRank values are viewed as firmly connected with other identified experts and therefore more reliable.

4 EXPERIMENT DESIGN

To evaluate the effectiveness of our model, we collected data from ScholarMate, one of the most popular and influential online research social platforms in China. The experiment aims to seek for experts for companies that need academia-industry collaboration. High-tech companies, especially the start-ups, need to seek help from experts who possess domain knowledge. Through our method, the companies are able to find a set of suitable experts who not only possess relevant expertise but also are expected to be cohesive in potential teamwork.

To verify the effectiveness of our model, baseline methods: content-based method and social network based method are selected. The selected methods are compared in terms of accuracy and error rate. Note that in proposed model, potential experts are selected by giving an appropriate threshold for reputation. Unsuitable experts are identified by setting an appropriate threshold for topic sensitive SimRank values. We will use the content-based method to test the accuracy and error rate of our identified experts in the second step. The formulas of accuracy and error rate are listed as follows. In addition, to test the effectiveness of unsuitable experts detection, we will ask domain experts to manually judge the detection result.

$$accuracy = \frac{|\text{verified experts}|}{|\text{identified real experts}|} \tag{4}$$

$$error\ rate = \frac{|\text{verified experts}|}{|\text{identified unsuitable experts}|} \tag{5}$$

5 DISCUSSION AND FUTURE WORK

Online research social platform has become an efficient tool for experts finding in different contexts. Current research focuses on finding experts based on various features including published research work, social relationships, and online behaviours on social platforms. However, no further measures are designed to verify the reliability of identified experts. Detecting unsuitable experts during experts finding process is critically important. To solve this problem, this research proposes an innovative TSSR model that can identify potential experts on online social platforms as well as differentiate “real” experts and unsuitable experts. The major contribution of this research is designing a Topic Sensitive SimRank method in identifying real experts as well as unsuitable experts online. Besides, we also design a topic modelling based method to enrich seeding expertise, which may also apply in other applications. Third, the future implementation of the proposed method will help government and institutions find experts in many situations such as emergencies and disasters.

We will continue improving our research in terms of several aspects. First, more sophisticated method such as advanced variant of SimRank will be adopted to improve the performance of our novel proposed model. Second, larger dataset will be collected to further verify the efficiency and effectiveness of the proposed model. Third, more evaluation metrics such as invited experts’ judgement will be introduced to measure the performance of the proposed model. Last, we still need to investigate existing research and select more general baseline methods to compare the performance of the proposed model with current research.

Acknowledgement

The authors gratefully thank the Editor and all reviewers. The authors also acknowledge with gratitude the generous support of the NSFC (National Natural Science Foundation of China) M1552003 and 71371164, and Beijing Municipal Science and Technology Project (Conflicts of Interests Identification and Reviewer Assignment for Municipal Science and Technology projects).

References

Antonellis, I., Molina, H. G., & Chang, C. C. (2008). Simrank++: query rewriting through link analysis of the click graph. *Proceedings of the VLDB Endowment, 1*(1), 408-421.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., & Vesci, G. (2013). *Choosing the right crowd: expert finding in social networks.* Paper presented at the Proceedings of the 16th International Conference on Extending Database Technology.

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology, 25*(2), 163-177.

Chen, J., & Safro, I. (2011). Algebraic distance on graphs. *SIAM Journal on Scientific Computing, 33*(6), 3468-3490.

Davoodi, E., Kianmehr, K., & Afsharchi, M. (2013). A semantic social network-based expert recommender system. *Applied Intelligence, 39*(1), 1-13.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Guler, I., Guillén, M. F., & Macpherson, J. M. (2002). Global competition, institutions, and the diffusion of organizational practices: The international spread of ISO 9000 quality certificates. *Administrative science quarterly, 47*(2), 207-232.

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research, 18*(4), 323-342.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Jeh, G., & Widom, J. (2002). *SimRank: a measure of structural-context similarity.* Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39-43.

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). *Latent dirichlet allocation for tag recommendation.* Paper presented at the Proceedings of the third ACM conference on Recommender systems.

Leicht, E. A., Holme, P., & Newman, M. E. (2006). Vertex similarity in networks. *Physical Review E, 73*(2), 026120.

Li, P., Cai, Y., Liu, H., He, J., & Du, X. (2009). Exploiting the block structure of link graph for efficient similarity computation *Advances in Knowledge Discovery and Data Mining* (pp. 389-400): Springer.

Liu, H., He, J., Zhu, D., Ling, C. X., & Du, X. (2013). Measuring similarity based on link information: A comparative study. *Knowledge and Data Engineering, IEEE Transactions on, 25*(12), 2823-2840.

Liu, X., Wang, G. A., Johri, A., Zhou, M., & Fan, W. (2014). Harnessing global expertise: A comparative study of expertise profiling methods for online communities. *Information Systems Frontiers, 16*(4), 715-727.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*(3), 245-251.

Rasiwasia, N., & Vasconcelos, N. (2013). Latent dirichlet allocation models for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35*(11), 2665-2679.

Sun, J., Xu, W., Ma, J., & Sun, J. (2015). Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework. *International Journal of Production Economics, 165*, 185-193.

Tang, X., Wang, J., Zhong, J., & Pan, Y. (2014). Predicting essential proteins based on weighted degree centrality. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 11*(2), 407-418.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of marketing research, 51*(4), 463-479.

Wei, X., & Croft, W. B. (2006). *LDA-based document models for ad-hoc retrieval.* Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.

Xi, W., Fox, E. A., Fan, W., Zhang, B., Chen, Z., Yan, J., & Zhuang, D. (2005). *Simfusion: measuring similarity using unified relationship matrix.* Paper presented at the Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.

Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y. K., & Xu, W. (2012). Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decision Support Systems, 54*(1), 564-573. doi: 10.1016/j.dss.2012.08.003

Yin, X., Han, J., & Yu, P. S. (2006). *LinkClus: efficient clustering via heterogeneous semantic links.* Paper presented at the Proceedings of the 32nd international conference on Very large data bases.