Measuring the Topic-Level Impact of Experts in Scholarly Network

Lili Lin
College of Computer and Information, Hohai University
No.1 Xikang Road,
Nanjing, 210098, Jiangsu, China
linlili@hhu.edu.cn

ABSTRACT
Identifying the most influential scientific experts is of vital importance for exploring scientific collaborations to increase productivity by sharing and transferring knowledge within and across different research areas. However, most state-of-the-art expert finding approaches have usually studied candidates' personal information and network information separately. In this dissertation research, we propose a Topical and Weighted Factor Graph (TWFG) model that simultaneously combines all the possible information in a unified way. In addition, we also design the Loopy Max-Product algorithm and related message-passing schedules to perform approximate inference on our cycle-containing factor graph model. Information Retrieval will be chosen as the test field to identify representative scholars for different topics within this area. Finally, we will compare our approach with three baseline methods in terms of topic sensitivity, coverage rate of SIGIR PC members (e.g. Program Committees or Program Chairs) and NDCG (Normalized Discounted Cumulated Gain) scores for different rankings on each topic.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Digital Libraries

General Terms
Algorithms, Design, Experimentation

Keywords
Expert Finding; Factor Graph; Topic Relevance; Scholarly Network

1. INTRODUCTION
In order to make good use of expertise and knowledge, one important task in scientific research area named *expert finding* has received a significant amount of attention in recent years. The goal of expert finding is to return a ranked list of knowledgeable experts with relevant expertise on a specific topic or research area. The expert finding process can help solve many challenging but practical problems, such as assign the appropriate program committee members or reviewers for an international conference, search the potential collaborators for a critical project, recruit the talented employees for some jobs or roles, find important experts for consultation by researchers embarking on a new research field etc. However, manually identifying these experts in a large research area or organization is obviously labor intensive and time consuming.

Some researchers use content-based methods to detect persons who are experts on a specific research topic. However, these kinds of methods mostly concentrate on providing relevance scores between candidates and a user's query topic or an inferred topic, while neglecting the social relationships between candidates for more precise expert identification. Another option is to use link analysis algorithms such as PageRank [11] and HITS [8] to address expert-finding tasks. But PageRank and HITS have a common problem: topic drift, which tends to make most in-links in the network dominant [13]. Due to the limitations of content-based methods and traditional link structure-based methods, some previous works not only consider the relevance of a candidate on a specific topic, but also analyze networks between candidates in order to improve expert finding efficiency. To the best of our knowledge, however, most of these methods model possible information separately or combine them in a specific order, which may cause possible experts to be ignored.

Based on the above motivations and my previous literature reviews, this dissertation will have some major contributions which could be summarized as follows:

(1) The first one is to explore what are the fundamental features or factors that influence the measurement of topic-level experts.
(2) The second challenge is how to construct a model by simultaneously combine all relevant features in a unified way to detect persons who are experts on a specific research topic.

2. RELATED WORK
Several studies have investigated approaches for expert finding. The existing approaches can be divided into three main categories according to their focuses.

2.1 Content-based Methods
The first kind of content-based methodology is treated as an information retrieval (IR) task by Text REtrieval Conference (TREC). Such methods are basically variations of two kinds: profile-centric methods (also referred to as candidate-centric or query-independent approaches) and document-centric methods (also referred to as query-dependent approaches). In profile-centric methodologies [15, 16] all documents or texts related to a candidate are first merged into a single personal profile, where the ranking score for each candidate is then estimated according to the profile in response to a given query. Nevertheless, the document-centric methods [15, 17] analyze the content of each document separately instead of creating a single expertise profile. In order to make use of the advantages of both the profile-centric and document-centric methods, some existing approaches [18, 19] combine the two methods to improve expert-finding performance. However, these kinds of studies generally concentrate on aligning search results with user queries, which is different from our concentrations, which conduct topic-dependent expert finding based on automatically inferred latent topics.

The second kind of content-based methods is known as topic modeling. An early topic model, named Probabilistic Latent Semantic Indexing (PLSI), was proposed by Hofmann [6] to calculate the probability of generating a word from a document based on the latent topic layer. Blei, Ng, and Jordan [2] addressed some limitations of PLSI by proposing a three-level hierarchical Bayesian model called latent Dirichlet allocation (LDA). As the inability of LDA to model topic correlation, Blei and Lafferty [14] proposed a Correlated Topic Model (CTM) which explicitly models the correlation between the latent topics in the collection. Moreover, a novel Author-Persona-Topic (APT) model was proposed by Mimno and McCallum [21] for matching reviewers to submitted papers by modeling the expertise of a person based on documents. As another follow-up effort of the LDA model, Tang, Jin and Zhang [12] further extended the LDA and proposed the Author-Conference-Topic (ACT) model to organize different types of information concurrently in academic networks.

2.2 Link Structure-based Methods

As content related to candidates cannot serve as direct evidence of their expertise, a few studies have tried to employ link structure among candidates to address the expert-finding problem. Link structure-based algorithms, such as PageRank [11] and HITS [8], can be used to analyze relationships in a scholarly network in order to find authorized experts. Some other works aimed at applying variations of HITS or PageRank algorithms in order to alleviate the limitations of some classical indicators (e.g. citation counts) for ranking in bibliometrics. Liu, Bollen, Nelson and Sompel [20] developed AuthorRank for this purpose, a modification of PageRank that considers link weights among the coauthorship links. Jurczyk and Agichtein [22] explored the HITS algorithm to estimate the authority of users that can be potentially used for finding experts in Question Answer portals. A weighted PageRank algorithm that considers citation and coauthorship network topology was proposed by Yan and Ding [23] to measure author impact. However, they are not effective for finding the top "experts" without considering content features. Moreover, all of them are topic-independent, and include certain classical indicators such as impact factor, H-index, and citation counts.

2.3 Combination of Content-based and Link Structure-based Methods

Some researchers have used documents or snippet-level content to provide topic relevance for each candidate, and then applied link analysis to further refine the ranking results. Campbell, Maglio, Cozzi, and Dom [3] used text analysis and network analysis to sort individuals within an email network. Specifically, they collected all emails related to a topic and analyzed emails between every pair of people for whom there was relevant correspondence to build an "expertise graph." They finally applied a modified HITS algorithm to obtain ratings for all senders and recipients on that topic. Zhang, Tang, and Li [24] first used candidates' personal information (e.g. personal profile, contact information, and publications) to estimate an initial expert score for each candidate and selected the top ranked candidates to construct a subgraph. They then proposed a propagation-based approach to improve the accuracy of expert finding within the subgraph. Jiao, Yan, Zhao, and Fan [25] used expert relevance scores to generate a subset of experts, and also used a modified PageRank algorithm to calculate the authority scores of experts. They then combined expert relevance and expert authority with a linear formula to express the final expertise of a candidate. Ding [5] proposed topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm. Two ways for combining the ACT model with the PageRank algorithm are proposed in her work: simple combination or using a topic distribution as a weighted vector for PageRank. However, most of the above methods do not simultaneously model all the possible information in a unified way. Furthermore, most of them tend to use a subset of candidates for identifying representative authors, which may filter out some potential experts.

3. PROPOSED METHODS

3.1 Feature Selection

Motivated by observations on certain common characteristics of judgments people make to find experts, we define two important features, topic relevance and expert authority, as personal information. In addition, we also extract the citation relationships between authors to build the citation network in order to acquire the topic-level influences as the network information.

Topic relevance. This local feature can be used to model the relevance between an author and a specific topic. Given a topic, the amount of information that an author's publications contain contributes to the presentation of how much of the required knowledge that author has. We assume that if an author possesses a higher probability on a given topic, he/she is more likely to be an expert on this topic. The topic relevance of a candidate on a specific topic can be inferred by the ACT model [12].

Expert authority. This local feature, calculated by PageRank [11] within an author citation network, can be used to model the popularity of an author. We make an assumption, in that among those authors with the same relevance on a given topic, the author with higher PageRank value is more likely to be an expert since he/she tends to be a popular author in scientific research areas.

Topic-level influences. Even with the same citation network structure, mutual influences between authors will vary on different topics. More precisely, when calculating author expertise on different topics, dissimilarities or similarities between authors can result in different contributions. This mutual influence between authors could be calculated based on Kullback-Leibler divergence [10].

3.2 Factor Graph Model

In order to make the proposed approach easier to describe and understand, the notations are first given in Table 1. As factor graphs have the potential to unify modeling with great generality and flexibility [9], we propose a Topical and Weighted Factor Graph (TWFG) model to leverage topic relevance, expert authority, and topic-level influence. For simplicity, we make an assumption that topics are independent of each other. Hence we can decompose our factor graph model into a set of factor graphs with the same topological structure on different topics. Figure 1 shows a simple TWFG on a given topic z corresponding to the example we have been used.

As each observed variable $v_i \in V$ corresponds to a hidden vector $\{\mathbf{y}_i\} \in Y$, thus the factor graph can be regarded as the composition of a set of hidden variables $Y = \{\mathbf{y}_i\}_{i=1}^{N}$ and a set of functions. Concretely, the functions in our model fall into node function g and edge function f. The former is used to model the personal information (i.e., topic relevance and expert authority) and the latter is used to model the network information (i.e.,

topic-level influences). Here, we define the **node function** as equation (1) on the intuition that authors with higher topic relevance on a given topic z are more likely to be experts on that topic and authors with higher citation counts tend to be experts even they have the same topic relevance.

$$g_i(\mathbf{y}_i, z) = g_i^z(y_i^z) = \begin{cases} \exp(p_i \alpha_{iz} y_i^z), & \alpha_{iz} \geq \lambda \\ \exp(-p_i \alpha_{iz} y_i^z), & \alpha_{iz} \leq \lambda \end{cases} \quad (1)$$

where p_i represents the PageRank value of a given author v_i; α_{iz} denotes the probability of an author α_{iz} on a given topic z; $y_i^z \in \{0,1\}$ reflects the importance of an author for topic z, $y_i^z = 0$ indicates author v_i is not important for topic z and $y_i^z = 1$ indicates author v_i is important for topic z; and λ specifies the relevance threshold between an author and a topic, $\alpha_{iz} \geq \lambda$ indicates author v_i is more relevant with a given topic z and $\alpha_{iz} \leq \lambda$ indicates author v_i is less relevant with a given topic z.

Obviously, an important author v_i on a given topic z may have a high influence on one of his/her neighboring author node v_j if they have a high similarity on their research interests/topics. Then, author v_j may have high probability to become an important author on topic z too. In order to capture the topic-level influences between neighboring author nodes, we define **edge function** as equation (2).

Table 1. Notations

Symbol	Description
N	the number of authors in the citation network
V	the set of authors in the citation network
E	the set of edges in the citation network
Y	the set of hidden vectors for all author nodes
z	a research topic within a research area
t	the number of topics
v_i	an author node in the citation network
$\mathbf{y}_i$	the hidden vector for all topics on author v_i
y_i^z	author v_i 's importance weight on topic z
p_i	the PageRank value of a given author v_i
α_{iz}	the probability of author v_i on topic z
e_{ij}	an edge between author v_i and author v_j
θ_{ij}^z	the dissimilarity weight associated with edge e_{ij} on topic z

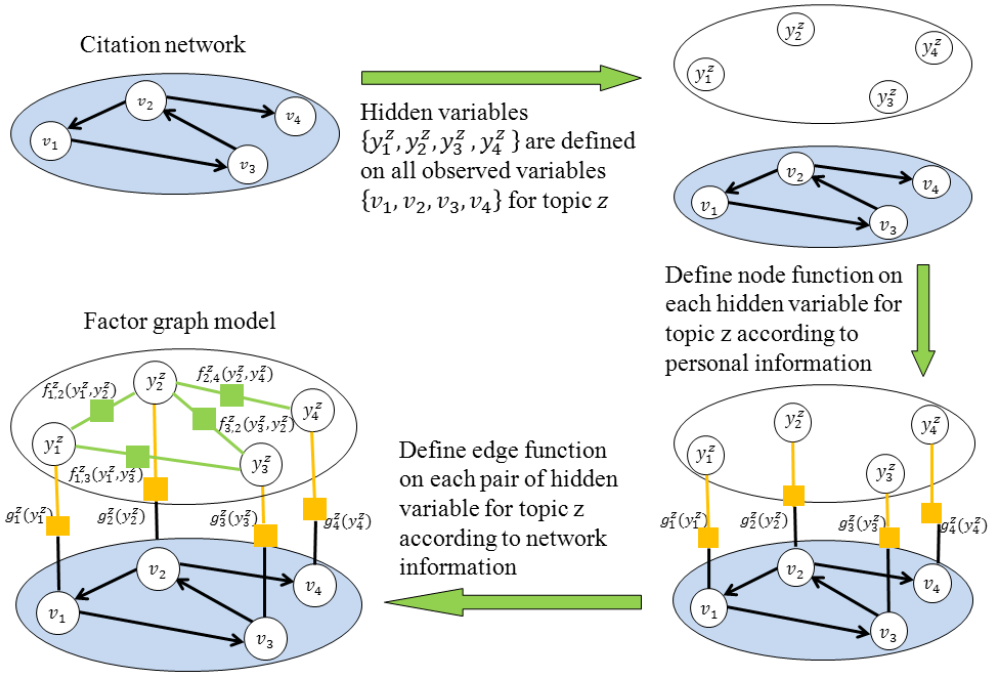

Figure 1 Graphical representation of a topical and weighted factor graph on a given topic z, where $\{y_1^z, y_2^z, y_3^z, y_4^z\}$ are hidden variables defined on all observed variables $\{v_1, v_2, v_3, v_4\}$ for topic z; $g_i^z(.)$ represents a node function and $f_{ij}^z(.)$ represents an edge function.

$$f_{ij}(\mathbf{y}_i, \mathbf{y}_j, z) = f_{ij}^z(y_i^z, y_j^z) = \begin{cases} \exp(\theta_{ij}^z), & \text{if } \theta_{ij}^z \leq \theta \text{ and } y_i^z = y_j^z \\ \exp(-\theta_{ij}^z), & \text{if } \theta_{ij}^z > \theta \text{ and } y_i^z = y_j^z \\ 1, & \text{if } y_i^z \neq y_j^z \end{cases} \quad (2)$$

where $y_i^z \in \{0,1\}$ and $y_j^z \in \{0,1\}$ represent the importance weight of author v_i and author v_j on a given topic z, respectively; θ_{ij}^z indicates the dissimilarity weight between author v_i and author v_j on topic z, which is calculated based on

Kullback-Leibler divergence [10] shown in equation (3); and θ specifies the dissimilarity threshold between author v_i and author v_j, $\theta_{ij}^z \le \theta$ indicates author v_i and author v_j have more similar research interests on topic z and $\theta_{ij}^z > \theta$ indicates less similar.

$$\theta_{ij}^z = \alpha_{iz} \ln \frac{\alpha_{iz}}{\alpha_{jz}} + \alpha_{jz} \ln \frac{\alpha_{jz}}{\alpha_{iz}} \qquad (3)$$

Based on above, we finally define the **objective function** by considering all the functions based on factor graph theory [1, 9] as equation (4).

$$\begin{aligned} p(Y) &= \frac{1}{S} \prod_{z=1}^{t} \prod_{i=1}^{N} g_i(\mathbf{y}_i, z) \prod_{z=1}^{t} \prod_{e_{ij} \in E} f_{ij}(\mathbf{y}_i, \mathbf{y}_j, z) \\ &= \frac{1}{S} \prod_{z=1}^{t} \prod_{i=1}^{N} g_i^z(y_i^z) \prod_{z=1}^{t} \prod_{e_{ij} \in E} f_{ij}^z(y_i^z, y_j^z) \\ &= \frac{1}{S} \prod_{z=1}^{t} (\prod_{i=1}^{N} g_i^z(y_i^z) \prod_{e_{ij} \in E} f_{ij}^z(y_i^z, y_j^z)) \qquad (4) \end{aligned}$$

where $Y = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ corresponds to all hidden variables; $g_i^z(y_i^z)$ is the node function and $f_{ij}^z(y_i^z, y_j^z)$ is the edge function; and S is a normalizing factor. As we have assumed that topics are independent, so that

$$p(Y) = \prod_{z=1}^{t} p(Y_z) \qquad (5)$$

Thus, once the topic is specified, the **objective function** for topic z can be defined as equation (6).

$$p(Y_z) = \frac{1}{S'} \prod_{i=1}^{N} g_i^z(y_i^z) \prod_{e_{ij} \in E} f_{ij}^z(y_i^z, y_j^z) \qquad (6)$$

where $Y_z = \{y_1^z, y_2^z, \cdots, y_N^z\}$ corresponds to the hidden variables for topic z; and S' is a normalizing factor.

3.3 Inference Algorithm

As a generic message-passing algorithm, the Sum-Product algorithm [9] has often been applied to compute the marginals of all variable nodes efficiently and exactly for the factor graph-based model. The algorithm involves passing messages between variable nodes (i.e., hidden variables) and function nodes on the built factor graph [9]. Message passing is initiated at the leaves. Each node v remains idle until messages have arrived on all but one of the edges incident on v. Once these messages have arrived, v is able to compute a message to be sent on the one remaining edge to its neighbor w (temporarily regarded as the parent). After sending a message to w, node v returns to the idle state, waiting for a "return message" to arrive from w. Once this message has arrived, the node v is able to compute and send message to each of its neighbors (other than w), each being regarded, in turn, as a parent. The algorithm terminates once two messages have been passed over every edge, one in each direction. However, the Sum-Product algorithm cannot address the problems to find the state configuration that has the largest probability and calculate the corresponding marginal probability under the most likely state configuration. Moreover, as Sum-Product algorithm cannot be directly applied for factor graph model with cycles, we finally use Loopy Max-Product algorithm to address the inference tasks. Hereby, we need to modify the Sum-Product algorithm into Max-Product algorithm [1] to find the state configuration $Y_z^{\max}$ that maximizes the objective function $p(Y_z)$ for a specified topic z, so that

$$Y_z^{\max} = \arg\max_{Y_z} p(Y_z) \qquad (7)$$

for which the corresponding value of the largest probability will be given by

$$p(Y_z^{\max}) = \max_{Y_z} p(Y_z) \qquad (8)$$

Due to the cycles in our factor graph model, the proposed Loopy Max-Product algorithm firstly initializes the message on every link between variable node and function node in each direction as 1 and then passes messages iteratively with serial schedule using random sequences until convergence. Here, update rules of the message passing for each topic z in our factor graph model can be defined as equation (9) - (12).

$$\mu_{f_{ij}^z \to y_i^z}(y_i^z) = \max_{y_j^z} [f_{ij}^z(y_i^z, y_j^z) \mu_{y_j^z \to f_{ij}^z}(y_j^z)] \qquad (9)$$

$$\mu_{f_{ij}^z \to y_j^z}(y_j^z) = \max_{y_i^z} [f_{ij}^z(y_i^z, y_j^z) \mu_{y_i^z \to f_{ij}^z}(y_i^z)] \qquad (10)$$

$$\mu_{y_i^z \to f_{ij}^z}(y_i^z) = \mu_{g_i^z \to y_i^z}(y_i^z) \prod_{f_h^z \in ne(y_i^z) \setminus f_{ij}^z, g_i^z} \mu_{f_h^z \to y_i^z}(y_i^z) \qquad (11)$$

$$\mu_{y_j^z \to f_{ij}^z}(y_j^z) = \mu_{g_j^z \to y_j^z}(y_j^z) \prod_{f_h^z \in ne(y_j^z) \setminus f_{ij}^z, g_j^z} \mu_{f_h^z \to y_j^z}(y_j^z) \qquad (12)$$

where $\mu_{f_{ij}^z \to y_i^z}(y_i^z)$ denotes the message sent from edge function node f_{ij}^z to variable node y_i^z and $\mu_{y_i^z \to f_{ij}^z}(y_i^z)$ denotes the message sent from variable node y_i^z to edge function node f_{ij}^z; $f_h^z \in ne(y_i^z) \setminus f_{ij}^z, g_i^z$ denotes the set of neighbor nodes of a given variable node y_i^z on the factor graph, excluding f_{ij}^z and g_i^z.

As every leaf node in the built factor graph is always a node function node g_i^z, its message to a variable node y_i^z is shown in equation (13). Thus, equation (11) and (12) can be further changed into equation (14) and (15).

$$\mu_{g_i^z \to y_i^z}(y_i^z) = g_i^z(y_i^z) \qquad (13)$$

$$\mu_{y_i^z \to f_{ij}^z}(y_i^z) = g_i^z(y_i^z) \prod_{f_h^z \in ne(y_i^z) \setminus f_{ij}^z, g_i^z} \mu_{f_h^z \to y_i^z}(y_i^z) \qquad (14)$$

$$\mu_{y_j^z \to f_{ij}^z}(y_j^z) = g_j^z(y_j^z) \prod_{f_h^z \in ne(y_j^z) \setminus f_{ij}^z, g_j^z} \mu_{f_h^z \to y_j^z}(y_j^z) \qquad (15)$$

So far, the maximal joint probability for the specified topic z can be obtained using equation (16) by propagating message from the leaves to an arbitrarily chosen root node y_i^z.

$$p(Y_z)^{\max} = \max_{y_i^z} (g_i^z(y_i^z) \prod_{f_h^z \in ne(y_i^z) \setminus g_i^z} \mu_{f_h^z \to y_i^z}(y_i^z)) \qquad (16)$$

Furthermore, we can compute the marginal probability for each author by multiplying all the incoming messages as equation (17).

$$p(y_i^z) = g_i^z(y_i^z) \prod_{f_h^z \in ne(y_i^z) \setminus g_i^z} \mu_{f_h^z \to y_i^z}(y_i^z) \qquad (17)$$

4. EXPERIMENTS

4.1 Data Collection

In this dissertation, we choose Information Retrieval (IR) as the test field. Papers and their citations were collected from the Web of Science (WOS) covering the period from 2001 to 2008, including 8,395 papers and 14,593 authors with 211,560 citations. Each paper contains related authors, title, source, published year, abstract, reference, citation counts, and so forth. The titles are preprocessed using a stemming algorithm and a stop word list. Citation records include the first author, published year, source, volume, and page number. Citations are used to generate a citation network. Details of our data collection are provided in [4].

4.2 Baseline Methods

As this dissertation focuses on finding topic-based experts, it is unreasonable to directly compare our results with other classical indicators or measures for author ranking, such as H-index, citation counts, and impact factor, which are all topic-dependent. Hence we choose three topic-level baseline methods to evaluate our approach (denoted as **TWFG**), including one method that combines topic model with citation counts (**TMCC)** and two topic-based PageRank algorithms [5].

4.3 Evaluation Methods

Empirically, fifty topics are extracted by the ACT model. A list of top 10 words is used to represent each extracted topic and to locate the emphasis of each topic during the period from 2001 to 2008. For evaluation, we will use the method of pooled relevance judgments together with human judgments to generate "ground truth" from different perspectives. Assessments will be firstly carried out firstly in terms of topic sensitivity and then the coverage rate of SIGIR PC members. Finally, we will use the Normalized Discounted Cumulated Gain (NDCG) [7] as a metric to compare different rankings of authors based on our approach and baseline methods.

4.3.1 Comparison of topic sensitivity

As the TMCC method is always topically sensitive, due to the fact that only authors whose topic relevance is above a defined topic relevance threshold will be chosen to rank on the top, we will evaluate topic sensitivity based on our approach and two topic-based PageRank algorithms [5]. Based on these three approaches, an illustration on the ratio of authors whose topic probabilities exceed the value of the given topic relevance threshold for each topic will be given to depict the topic sensitivity of our approach and the two other baseline methods.

4.3.2 Comparison of coverage rate of the SIGIR PC members

As the ACM's Special Interest Group on Information Retrieval (SIGIR) is one of the most important international conferences for the presentation of new research results and demonstration of new systems and techniques in the field of information retrieval (IR), it is reasonable to suppose that only persons who have made significant contributions to research in information retrieval are chosen as SIGIR PC members. As suggested, a second assessment will be carried out based on the professional achievement of authors selected as SIGIR PC members. We choose SIGIR PC members (i.e. Program Chairs, Program Committees, and Conference Committees) from 2001 to 2008 as the ground truth for evaluation. In order to conduct a topic-level comparison on SIGIR PC members, we tailor our evaluation data that corresponds with each topic. In other words, only SIGIR PC members whose topic probabilities are higher than a given topic relevance threshold will be picked out as the ground truth on that topic. Similar to the first assessment, we will conduct a comparison of the ranking results based on four topic relevance thresholds with a specific dissimilarity threshold. The result for the coverage rate of the SIGIR PC members among the top *k* (i.e. *k*= 5, 10, 20, 50, 100) authors with different topic relevance thresholds will be presented.

4.3.3 Comparison of NDCG scores

Finally, in order to compare ranking results through NDCG metric, a list of corresponding review/survey papers for each topic along with their citation records, will be recommended by ACT model by setting a topic relevance threshold. Here, we will use the number of citations (named as citation score) by topic-related review papers to depict the importance of an author on the target topic. We make an assumption that if an author writes *n* papers which are separately cited by the list of review papers under a given topic, then the citation score of the author for the target topic is *n*. By calculating citation scores for each author within each topic, we will acquire the "ground truth" for evaluation. The comparison of NDCG scores of author rankings for each topic will be demonstrated in order to compare the average performance of our method and three baseline methods.

5. CONCLUSION AND FUTURE WORK

By now, we have finished constructing the factor graph model and developing the Loopy Max-Product algorithm with corresponding message-passing schedules on the constructed cycle-containing factor graph model. The experiments for evaluating the proposed approach and the baseline methods are still in progress. Future work includes identifying how to incorporate temporal information into our model in order to conduct systematical analysis of author expertise on different topics over time.

6. REFERENCE

[1] Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York, NY: Springer Publications, 359- 419.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent Dirichle t allocation. *Journal of Machine Learning Research*, 3, 993-1033

[3] Campbell, C., Maglio, P., Cozzi, A., & Dom, B. 2003. Expertise identification using email communications. *In Proceedings of 12th International Conference on Information and Knowledge Management* (New Orleans, LA, USA, November 02-08, 2003). CIKM '03. ACM, New York, NY, USA, 528-531.

[4] Ding, Y., & Cronin, B. 2010. Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*(1), 80-96.

[5] Ding, Y. 2011. Topic-based PageRank on author co-citation networks. *Journal of the American Society for Information Science and Technology*, *62*(3), 449-466.

[6] Hofmann, T. 1999. Probabilistic Latent Semantic Indexing. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berkeley, CA, USA, August 15-19, 1999). SIGIR '99. ACM, New York, NY, USA, 50-57.

[7] Jarvelin, K., & Kekalainen, J. 2002. Cumulated gain-based evaluation of IR techniques. *Association for Computing Machinery Transactions on Information Systems*, *20*(4), 422-446.

[8] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, *46*(5), 604-632.

[9] Kschischang, F. R., Frey, B. J. & Loeliger, H. 2001. Factor graphs and the sum-product algorithm. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, *47*(2), 498-519.

[10] Kullback, S., Burnham, K. P., & Laubscher, N. F., Dallal, et al. 1987. Letter to the editor: The Kullback-Leibler distance. *The American Statistician*, *41*(4), 338-341.

[11] Page, L., Brin, S., Motwani, R., & Winograd, T. 1999. *The PageRank citation ranking: Bringing order to the Web.* Technical Report, Stanford InfoLab, 1999-0120.

[12] Tang, J., Jin, R., & Zhang, J. 2008. A topic modeling approach and its integration into the random walk framework for academic search. *In Proceedings of 2008 Institute of Electrical and Electronics Engineers International Conference on Data Mining* (Pisa, Italy, December, 2008). IEEE Computer Society Washington, DC, USA, 1055-1060.

[13] Zhang, J., Tang J., & Li J. 2007. Expert Finding in a Social network. *Advances in Databases: Concepts, Systems and Applications*, Lecture Notes in Computer Science 4443, 1066-1069.

[14] Blei, D. M., & Lafferty, J. D. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1), 17–35.

[15] Balog, K., Azzopardi, L., & Rijke, M. D. 2006. Formal models for expert finding in enterprise corpora. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, WA, USA, August 6-11, 2006). SIGIR '06. ACM, New York, NY, USA, 43-50.

[16] Karimzadehgan, M., Belford, G. G., & Oroumchian, F. 2008. Expert finding by means of plausible inferences. *Proceedings of 2008 International Conference on Information and Knowledge Engineering* (Las Vegas, Nevada, USA, July 14-17, 2008), IKE '08. CSREA Press, 23-29.

[17] Wu, H., Pei, Y., & Yu, J. 2009. Hidden topic analysis based formal framework for finding experts in metadata corpus. *Proceedings of the 8th IEEE/ACIS International Conference on Computer and Information Science* (Shanghai, China, June 1-3, 2009), ACIS-ICIS '09. IEEE Computer Society, Washington, DC, USA, 369-374.

[18] Petkova, D., & Croft, W. B. 2006. Hierarchical language models for expert finding in enterprise corpora. *Proceedings of the 18th Institute of Electrical and Electronics Engineers International Conference on Tools with Artificial Intelligence* (Arlington, Virginia, November 13-15, 2006). IEEE Computer Society, Washington, DC, USA, 599-608.

[19] Serdyukov, P., Henning, R., & Hiemstra, D. 2008. Modeling multi-step relevance propagation for expert finding. *Proceedings of the 17th Association for Computing Machinery Conference on Information and Knowledge Management* (Napa Valley, CA, USA, October 26-30, 2008), CIKM '08. ACM, New York, NY, USA, 1133-1142.

[20] Liu, X., Bollen, J., Nelson, M. L. & Sompel, H. V. 2005. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6), 1462–1480.

[21] Mimno, D. & McCallum, A. 2007. Expertise Modeling for Matching Papers with Reviewers. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Jose, CA, USA, August 12-15, 2007), KDD '07. ACM, New York, NY, USA, 500-509.

[22] Jurczyk, P., & Agichtein, E. 2007. Hits on question answer portals: Exploration of link analysis for Author Ranking. *Proceedings of the 30th Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, Holland, July 23-27, 2007), SIGIR'07. ACM, New York, NY, USA, 845-846.

[23] Yan, E., & Ding, Y. 2011. Discovering Author Impact: A PageRank Perspective. *Information Processing and Management*, 47(1), 125-134.

[24] Zhang, J., Tang J., & Li J. 2007. Expert Finding in a Scholarly network. *Advances in Databases: Concepts, Systems and Applications*, 4443, 1066-1069.

[25] Jiao, J., Yan, J., Zhao, H., & Fan, W. 2009. ExpertRank: An expert user ranking algorithm in online communities. *Proceedings of the 2009 International Conference on New Trends in Information and Service Science* (Beijing, China, June 30 - July 2, 2009), NISS '09. 674-679.