Constructing Effective and Efficient Topic-Specific Authority Networks for Expert Finding in Social Media

Reyyan Yeniterzi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
reyyan@cs.cmu.edu

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
callan@cs.cmu.edu

ABSTRACT

Authority-based approaches are widely used in expert retrieval from social media. However, most of these approaches are applied to either topic-independent networks, or more topic-dependent networks which still contain topic-irrelevant users as nodes and interactions as edges. Therefore, authority estimation over these graphs is still not topic-specific enough. This paper proposes a more topic-focused authority network construction approach which provides more effective topic-specific authority modeling of users. Focusing the computational effort to more topic-specific authority networks also leads to significant gains in running time for authority estimation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

authority networks; social media; expert finding

1. INTRODUCTION

In social media environments anybody can create content on any topic, and so in order to be considered an expert on a specific topic, it is not just enough to author content on that particular topic. One also needs to have a topic-specific influence over other users. Therefore, in addition to writing, getting attention from other users by either being read or commented is also important. This influence can be estimated by applying authority-based approaches to user authority networks which are constructed by using the authoritative interactions between users.

Popular authority estimation approaches, like *PageRank* and *HITS* based approaches, are commonly applied to social networks in order to estimate the authority of the users. However, depending on the graph and the approach being used, not all the estimated authority scores are topic-specific that can be used to improve the ranking of expert candidates on the particular topic. For instance, *PageRank* [3] is a topic-independent algorithm which is applied to the whole user network. *Topic-Sensitive PageRank* [6] is a more topic-specific approach which allows teleportations only to topic-relevant users but still iterates over a topic-independent authority network. *HITS* [9] approach focuses more on topic-dependent sub-graph of web pages, but its application to user nodes does not provide topic-specific subgraphs, but instead returns authority networks which contain topic-irrelevant user nodes and user interaction edges.

This paper proposes a more topic-specific authority graph construction approach, which only uses topic-relevant users, and interactions among them, which are originated from topic-relevant content. This proposed graph, called *Topic-Candidate* (TC) graph, is used to estimate more topic-specific authority scores, which provide statistically significant improvements over the authority scores estimated from PR and $HITS$ graphs. The proposed graph also drastically decreases the computational effort that is required to estimate the authority scores.

SoMeRA'14, July 6–11, 2014, Gold Coast, QLD, Australia.

http://dx.doi.org/10.1145/2632188.2632208 .

2. RELATED WORK

Topic-specific expert retrieval is a widely studied topic. A recent literature review by Balog et al. [2], provided a detailed overview of the prior research on expert finding. Among these systems, the most effective expert retrieval systems are characterized by their use of document, profile, or graph-based techniques. *Document-based* models [1, 10] initially retrieve topic-relevant documents, associate them with their authors, and finally rank the candidates based on the aggregated expertise scores. *Profile-based* models [1] develop a model (profile) of each user using the text associated with that user, and given a topic, standard text retrieval is used to rank the profiles. *Graph-based* models [13] go beyond using only the text content by exploiting the link structure between documents and entities.

In addition to the user created content, the underlying social network structure of the social media is a valuable source for graph-based methods in order to identify more authoritative topic-specific experts. For instance, the following network of microblogs was used to identify topical authorities in microblogs [14] and similarly the feed subscription lists of blogs were used to identify influential blogs given a topic [7]. User interactions between email senders and receivers [4] or askers and responders in question answer communities [8, 15] were also explored for authority estimation.

Prior work used these existing social networks by con-

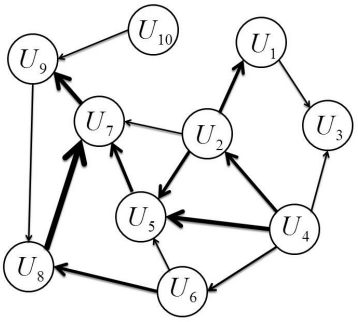

Figure 1: An example user authority network in which the direction of the link is an indication of the more authoritative user and the thickness of link represents the degree of authority.

structing authority graphs and applying link-analysis approaches such as *PageRank* or *HITS* [5, 4, 8]. Furthermore Zhang et al. [15] proposed a variation of PageRank, called *ExpertiseRank*, to identify experts in online communities. Weng et al. [14] also proposed *TwitterRank*, which is also an extension of PageRank, to identify topic-sensitive influential twitterers.

Among these authority estimation algorithms, there is not a clear winner. Their effects seem to depend on the environment and network structure. However, in general, approaches that are more topic-dependent work better than ones that are topic-independent.

3. APPROACH

This work expands the related work by exploring the effects of topic-dependency of the authority graphs on authority estimation, and proposes a more topic-specific authority network construction approach. Initially, background information on standard authority-based approaches, *PageRank* and *HITS*, and their authority graphs are described and then the proposed authority graph is discussed.

Authority-based approaches use the relationship and activities among entities to measure the influence and importance of each entity. Authority is measured over a graph in which nodes are the users, and directed edges indicate a relation among users where the direction of the edge is an indication of the more authoritative user. An example user authority network is presented in Figure 1. As can be seen in the figure, the edges can also be weighted by the frequency of the activity between the users.

3.1 PageRank and Topic-Sensitive PageRank

For user-authority estimation task, *PageRank (PR)* [3] is the probability distribution representing the likelihood that a user will find an authority by randomly following authoritative links among users. *PageRank* is a topic-independent algorithm that considers all users and their activities over all the documents, therefore it is applied to the whole user authority network as shown in Figure 1. It is normally applied to unweighted web graphs for estimating authority of web pages. Its customized version for estimating authority among users is as shown:

$$PR(u) = \frac{1-d}{|U|} + d \sum_{i \in IL_u} \frac{PR(i)}{OL(i)} \quad (1)$$

where $PR(u)$ is the *PageRank* score of user u, IL_u is the set of nodes that are linked to u (incoming links), $PR(i)$ is the *PageRank* score of node i, and $OL(i)$ is the number of outgoing links from node i. The d in the Equation 1 refers to damping factor. The teleportation probability is uniformly distributed between all users, $1/|U|$ where $|U|$ is the number of users in the graph.

Same activity can occur between same users multiple times over different posts in different times. Therefore, same type of activities among users can be aggregated to determine the weight of the edge which can be used to calculate weighted *PageRank* scores. In weighted *PageRank*, compared to unweighted *PageRank*, instead of just using the number of outgoing links from a node, the probability of following a link depends on the proportion of the weight of the edge to sum of weights of all the outgoing edges.

Topic-Sensitive PageRank (TSPR) [6] assumes that teleportation is possible only to users that are associated with topic-relevant content. Therefore, in *TSPR*, unlike the *PR*, the teleportation probabilities are distributed uniformly among users who have created topic-relevant content which has been retrieved as a result of searching the topic over the document collection. Instead of using a teleportation probability of $1/|U|$ for every user, the probability $1/|U_t|$ where U_t is the set of users associated with topic t, is used for users whose content have been retrieved for the particular topic. For the rest of the users, 0 is used as the teleportation probability.

Both *PR* and *TSPR* algorithms are applied to the whole network which consists of all users and all interactions among them. Such a network is useful for identifying authorities in general, however it may not be very effective in identifying more topic-specific authorities. *TSPR* favors users that are associated with topic-relevant content, but it still does not differentiate whether the edges are topic-relevant or not.

3.2 Hyperlink-Induced Topic Search (HITS)

Unlike *PageRank*, *Hyperlink-Induced Topic Search (HITS)* [9] algorithm is using a topic-specific subgraph instead of the whole graph and for each node it calculates two types of scores, *authority* and *hub*. The algorithm consists of several iterations and at each step first the *authority* and then the *hub* scores are updated. *Authority* score of a node is equal to the sum of the *hub* scores of the nodes of incoming edges. Similarly *hub* score is equal to the sum of the *authority* scores of the nodes of outgoing edges. The default *HITS* algorithm is applied to unweighted graphs, but a customized version of *HITS* can also be applied to graphs with weighted edges as in user graphs. In such a graph, the *auth* and *hub* scores are calculated by using the weights of edges by multiplying them with *hub* scores of the incoming edges or *auth* scores of the outgoing edges.

With respect to applying *HITS* to social networks, one can think of the nodes with high *authority* scores as the authoritative users whose content attracts attention of many other users who interact a lot with other users and their content. Similarly, the nodes with high *hub* scores are the active users who interact a lot with other users and their content. For instance in a blogosphere, a good *hub* is a user who reads or comments to many blog posts that also receives attention from other users, and a good *authority* is a user whose posts have been read or commented by other users who also interact with many other users. In such a scenario,

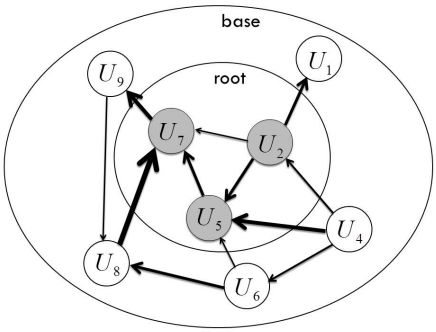

Figure 2: Expanding the root set (in grey) into a base set and constructing a HITS graph.

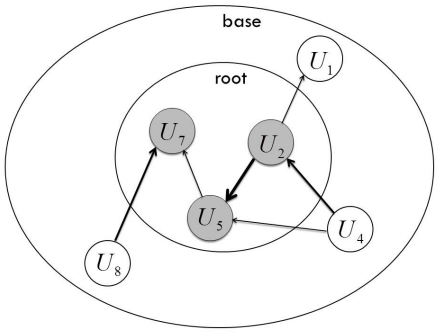

Figure 3: Removing topic irrelevant edges from HITS graph and constructing a Topic Candidate graph.

using *authority* score directly for estimating authority score of the user is a perfect fit.

Kleinberg applied *HITS* approach to more topic-specific authority sub-graphs with the aim to focus the computational effort on highly relevant documents [9], instead of using all web pages as in compared to *PageRank*. Kleinberg's approach for constructing the *HITS* authority network is also used in this paper in order to construct more topic-specific user networks. Such a sub-graph is constructed by initially retrieving top n topic-specific expert candidates, which is called the *root set*. Later on this *root set* is expanded into a *base set* which consists of users who have interacted with these candidates in the *root set*, either by being connected to or connected from. Such a *base set* contains all the users within the *root set*. After creating this *base set*, a graph is constructed by using all the candidates within this set as nodes and existing interactions among them as edges. An example *root set*, *base set* and the constructed graph is given in Figure 2. Compared to the *PageRank* graph in Figure 1, this is a more topic-dependent authority network.

3.3 The Proposed Authority Network

The *HITS* network is well suited for web graphs in which each node (web page) is mainly about one topic, and so using *base set* nodes and edges among them creates a topic specific sub-graph. However, there is an issue that needs to be considered in using *HITS* graphs for users and interactions among them. Unlike web pages, users are not interested in or knowledgeable on only one topic, instead they can be experts on or interacting with several topics, which are either related or not related to the particular topic. Because of these different types of interactions, during *base set* construction not all the inserted users and interactions between these users and the *root set* users will be topic-relevant. Since all the interactions of users are used during this expansion, the final constructed *HITS* graph will still contain many topic-irrelevant user nodes and interactions. Existence of such nodes and interactions may cause iterating the authority to topic-irrelevant users and favoring users who may be authorities on some topics but not on the particular given topic.

In order to prevent this, we propose constructing more topic focused authority sub-graphs. Graphs, called *Topic-Candidate (TC)* graphs, are proposed, which are constructed by using user interactions from only topic-relevant posts, rather than using all user interactions from all posts. Using the same notation used as in *HITS*, candidates retrieved with the content-based approach can be referred as the *root set*, while the set of nodes in the *root set* together with the nodes directly connected with them construct the *base set*. However, unlike *HITS* graphs, only the nodes in base set that are connected to/from nodes in root set due to topic-relevant activities are used. All the other topic-irrelevant users or interactions are ignored. An example *Topic-Candidate* graph is given in Figure 3. Compared to *HITS* graph in Figure 2, the number of nodes and edges are less, and in some cases the weight of some edges are also lower.

If we take a closer look, U_6 from Figure 2 does not exist in Figure 3, because its only connection to root set (through node U_5) is not originated from topic-relevant activities. Therefore, U_6 does not exist in *TC* graph. Similarly U_9 from Figure 2 does not exist in Figure 3, due to not using the topic-irrelevant edge $U_7 \rightarrow U_9$. The weight of the other edges, that are connected to or from any *root set* node, are also decreased due to removing topic-irrelevant interaction counts, such edges include $U_8 \rightarrow U_7$, $U_2 \rightarrow U_1$, $U_5 \rightarrow U_7$ and $U_4 \rightarrow U_5$. Edges, in which all the edge weight is coming from topic-irrelevant posts, are also removed from the graph, such as $U_2 \rightarrow U_7$. In *PageRank* and *HITS* graphs (Figure 1 and 2), due to using all user interactions, U_7 favors U_9 with its authority. But when only the topic-relevant activities are used, U_7 no more influences its authority to node U_9 which prevents U_9 to be estimated as an authority on the particular topic. To sum up, as can be observed in Figure 3, in *TC* graphs all the edges are originated from user actions performed on topic-relevant content and they are either directed to or from one of the topic-relevant content authors (the nodes in *root set*).

4. DATASET

Research on how an organization can use its internal social media for locating experts necessarily involves data that is difficult to share. Our research used blog and related data provided by a large multinational IT firm. This blog collection has been previously used [11, 12]. Although the dataset is not public due to the personal and company-internal information it contains, we believe that it is typical of such datasets. The dataset characteristics are summarized below so that the dataset can be compared to other blog datasets.

The collection consists of blog data (posts and comments) and employee metadata covering a 56-month timespan. This dataset also includes access logs - which employees read

# Posts	165,414
# Comments	783,356
# Employees	>100,000
# Posters	20,354
# Commenters	42,169
# Readers	92,360

Table 1: Statistics of the corporate blog collection.

Approach	NDCG@1	NDCG@3	NDCG@10
Profile	0.7000	0.6689	0.6494
Votes	0.3667	0.4090	0.4140
ReciprocalRank	0.7083	0.7003	0.7281
CombSUM	0.6417	0.6334	0.6168
CombMNZ	0.5333	0.5295	0.5124
IRW	0.5167	0.5189	0.5159

Table 2: Content-based baseline results.

which blog entries for 44 of the 56 months. Statistics related to this dataset are summarized in Table 1.

Employees must login to corporate information systems, therefore users are not anonymous in this environment. All posts and comments created have the authorship information available. Only this information is used to associate posts and comments with the corresponding candidates. The access logs contain the employee ID of the visitor, the timestamp of the visit, the URL of the blog post visited, and the employee ID of the author of the blog post. Employees also have access to a corporate blog search engine. We were also provided with this search engine's access logs.

4.1 Evaluation Data

40 work-related topics were created for testing. Some of these were selected from search queries in the access logs of the corporate blog search engine and the rest of them were created by the company employees. The topics from the access logs were selected to mirror task-specific expert-seeking behavior such as 'oracle performance tuning' and 'websphere process server'. On the other hand, topics created by the employees were considerably more general like 'mainframe' and 'cloud computing'.

A sample-based approach was used to create the pool of candidate experts to be assessed. The top 10 candidates returned by several content-based expert-finding algorithms were combined to create a candidate pool. Deeper pools are desirable, of course, but an explicit goal was to produce pools small enough for an assessor to assess in less than an hour. Due to data confidentiality agreement, the manual assessments were performed by the author of this paper. Candidate experts were displayed in a random order and for each candidate top 3 most topic-relevant posts or comments were displayed during assessment. Expertise was measured on a 4-point scale (not expert, some expertise, an expert, very expert) depending on candidate's documents.

4.2 Authority Networks

This blog collection contains two types of user interactions: reading and commenting. These interactions are compared in terms of their effectiveness on estimating topic specific authority. Commenting may be viewed as a stronger form of evidence because it requires individual to take an action. Commenting information is also more generally available because most blog applications display user ids next to comments. Reading may be viewed as a weaker form of evidence because it requires less effort and may even be accidental. Typically the user ids of readers are not displayed, which makes this form of evidence somewhat unique to organizational blogs. However, reading is much more common than commenting, which might compensate for its weaknesses or make it more useful for low-traffic situations.

Separate graphs are created for reading and commenting interactions. In these graphs, the edges are from readers or commenters to authors; if a post attracts many comments, the author benefits, not the users who participate in the discussion. The edges are weighted by the number of blog posts written by $user_i$ and read (or commented) by $user_j$. This model does not consider whether $user_j$ read a specific post once or several times; only the total number of posts that were read is important.

5. EXPERIMENTS AND RESULTS

The *HITS* graph and the proposed *TC* network require a list of topic-relevant candidates, referred as *root set*, in order to construct the authority graphs. Therefore, content-based approaches are used initially to retrieve an initial good ranking of experts. The estimated authority scores are later interpolated with these content-based scores in order to improve the performance of expertise ranking.

During data processing, all html tags are removed and krovetz stemmer is applied to these html tag free documents. Indri [1] search engine is used for indexing and retrieval.

5.1 Content-based Experiments

Applied content-based expert finding approaches include:

- Profile-based: A single profile (big document) is built for each user using all blog posts written by the particular user. Given a query, the relevancy ranking of the profiles are used as the users' expertise ranking. This approach is very similar to Balog's Model 1 [1].
- Document-based: Voting Models [10] provide several options for aggregating the documents, therefore they have been chosen as the document-based approaches. *Votes*, *ReciprocalRank*, *CombSUM* and *CombMNZ* approaches are applied to the bloggers of the retrieved blog posts. During retrieval only the top n documents are retrieved to identify the expert candidates for a given topic. Initial experiments performed on the data revealed that retrieving the top 1000 blog posts provides high baseline scores.
- Graph-based: The *Infinite Random Walk* (IRW) model from multi-step relevance propagation algorithms [13] is applied. Different λ values are tested, $\lambda = 0.01$ resulted in high scores, therefore used in this paper.

The results of these experiments are summarized in Table 2.

According to Table 2, the *Reciprocal Rank* approach outperforms all the other models which suggests that highly ranked documents contribute more to the expertise of a candidate. Due to its effective performance, the *Reciprocal Rank* approach was used in the rest of the experiments as the content-based baseline approach.

[1]http://www.lemurproject.org/indri/

Approach	Graph	R	C	NDCG@1	NDCG@3	NDCG@10	MAP_{VE}	MRR_{VE}
RR	-	-	-	0.7083	0.7003	0.7281	0.3621	0.5156
PR	PR	0.10	0.10	0.7500	0.7085	0.7176	0.4653	0.5622
	HITS	0.10	0.10	0.7500	0.7150	0.7164	0.4662	0.5655
	TC	0.20	0.30	0.7833	0.7029	0.6643	$0.4701_{s'}$	0.6612^{r}
TSPR	PR	0.80	0.10	0.7583	0.6944	0.7093	0.4375	0.5429
	HITS	0.30	0.60	0.7583	0.7023	0.7155	0.4420	0.5535
	TC	0.40	0.30	0.7917	0.7139	0.7005	0.4792	0.6299^{r}
HITS	PR	0.00	0.10	0.7417	0.6903	0.7163	0.4533	0.5536
	HITS	0.00	0.10	0.7417	0.6943	0.7194	0.4591	0.5535
	TC	0.10	0.00	0.7333	0.6891	0.7006	0.4392	0.5379

Table 3: Results after re-ranking experts with authority scores calculated over unweighted networks (R: Reading; C: Commenting). $1-(R+C)$ is the weight of content-based approach. MAD of the weights are 0 for all.

5.2 Authority-based Experiments

After retrieving an initial list of expert candidates with content-based methods, authority measures were used to favor authoritative experts. The initially retrieved candidates were re-ranked with the estimated authority-based scores. A weighted combination of normalized content-based expertise, reading and commenting authority scores are used. The following equation is used to calculate the final scores:

$$finalScore = content^{\lambda} * reading^{\beta} * commenting^{\theta} \quad (2)$$

where $\lambda + \beta + \theta = 1$. 5-fold cross-validation is used to find the optimum parameter setting for the interpolation. The optimum parameter setting is identified by using the median value, and reported together with the experimental results. *Median Absolute Deviation (MAD)* value, which is a summary statistic of the statistical dispersion of the data, is used to see the variability of the estimated optimum weights. MAD is the median of the absolute deviations from the data's median and calculated as shown:

$$MAD = median(|x_i - median(X)|) \quad (3)$$

5.2.1 Experiments with Topic-Candidate Graph

Standard authority estimation algorithms *PageRank*, *Topic Sensitive PageRank* and *HITS* were applied to the proposed *Topic Candidate (TC)* graph, and other commonly used authority graphs, *PageRank* and *HITS*, in order to see the effects of the proposed topic-specific authority network. For each *PR*, *HITS* and *TC* graphs, the experiments were performed with weighted (by the frequency of the activity between two users) and unweighted graphs. The experimental results of weighted and unweighted graphs were very similar, therefore due to space restrictions only the results of unweighted graphs are presented in Table 3.

In Table 3 the first column presents the authority approach used and the second column shows the authority graph that was used in iterations. The next two columns reports the weights of the different authority signals, where R stands for *reading* and C stands for *commenting*. $1-(weight_R+weight_C)$ is the weight of the content-based baseline approach, which has not been shown explicitly. The first row of Table 3 reports the scores of content-based baseline *Reciprocal Rank* (RR) expert retrieval approach without any authority re-ranking ($weight_R = 0,\ weight_C = 0$) and the rest of the rows summarize the results after re-ranking. The middle three columns summarize the *NDCG* scores after re-ranking. *NDCG* metric is a graded relevance metric which takes into account all the four relevance degrees, *very expert*, *an expert*, *some expertise* and *not an expert*, within the assessments. Among these, only the *not an expert* assessed ones are assumed as irrelevant but all the others are considered as relevant. Due to high number of relevant categories, the number of relevant expert candidates within top 1, 3 and 10 is high, and therefore the effects of authority-based re-ranking may not be obviously observed with *NDCG* metrics. Since our aim is to rank the *very expert* candidates in higher ranks, a detailed analysis on the effects of re-ranking was performed on only the *very expert* (VE) assessed candidates, while considering all other relevance categories as irrelevant. With such an experimental evaluation, the assessment values are not graded anymore but instead they are binary; therefore *Mean Average Precision (MAP)* and *Mean Reciprocal Rank (MRR)* metrics are used to present the results, which are summarized in the last two columns in Table 3.

Two statistical significance tests; (1) randomization test and (2) sign test, are applied in order to see the effects of the proposed approaches. Results that are significant with $p < 0.05$ are presented with r (randomization test) and s (sign test) symbols and results which are significant with $0.05 < p < 0.1$ are presented with r' and s' symbols.

According to Table 3, *PR* approach performs similarly when applied to *PR* and *HITS* graphs. The estimated authorities resulted in improvements over the best content-based baseline in terms of *NDCG@1*, MAP_{VE} and MRR_{VE}. When *PR* is applied to *TC* graphs, the improvements are much higher. Compared to *PR* and *HITS* graphs, statistically significant improvements were observed with *TC* graph with respect to MAP_{VE} and MRR_{VE}. Similar results are also observed with the *TSPR* approach. The trend of results with *PR* and *HITS* approaches are also similar in weighted graphs which are not explicitly shown in this paper.

However, the *NDCG@10* scores are actually getting worse after authority-based re-ranking. A detailed analysis revealed that the decrease in *NDCG@10* metric is caused by introducing new candidates which have not been assessed in manual assessments due not getting into top 10 expert candidates with any of the content-based approaches. In *PR* and *HITS* graphs, the average number of newly introduced candidates are 0.125, while this number is 0.85 for *TC* graphs. This is the reason why *NDCG@10* scores are lower with *TC* graphs compared to *PR* and *HITS* graphs.

In Table 3 with unweighted graph results, the *TC* graphs did not outperform the *PR* and *HITS* graphs with the *HITS* approach. However, with the weighted graphs, the perfor-

Graph Type	# Nodes		# Edges		Running Time	
	R	C	R	C	R	C
PR	92K	42K	1,631K	214K	1,202	85
HITS	57K	14K	1,480K	138K	1,116	49
TC	7K	1K	9K	2K	4	1

Table 4: Approximate average number of nodes, edges and running times (in seconds) in different authority graphs (R: Reading; C: Commenting).

mances of *TC*, *PR* and *HITS* graphs are very compatible which shows that there is not a clear and consistent winner between these tested graphs when *HITS* is applied.

Analyzing the ranking of *very expert* users after authority-based re-ranking suggests that the estimated authorities are overall useful for them, since both the MAP_{VE} and MRR_{VE} scores have improved statistically significantly compared to best content-based baseline for all graphs (even not shown explicitly in the table with r and s symbols). The increase in MAP_{VE} score suggests that authority scores are generally useful for favoring *very expert* candidates.

With respect to comparing the reading and commenting activities, there is not a clear winner. One may expect commenting to be more powerful since it is a more explicit form of authority signal compared to the more implicit reading signal. However, the results suggest that the higher frequency of reading signal compensates its weakness.

Other than the improvements in accuracy, using these sub-graphs can also improve the running time performance of the applied authority estimation approaches. Table 4 presents the change in the average number of nodes and edges within the used graphs, and the approximate running time (in seconds) of applying authority-based approaches to these graphs. According to Table 4, the number of nodes and edges decrease drastically in *TC* graphs compared to *PR* and *HITS* graphs. However, the graphs have still hundreds of nodes and edges which makes them hardly sparse. Due to this decrease in size, the iterations take less time and so the overall running time of the approaches also drop significantly as shown. These numbers indicate that using *TC* graphs does not only provide more effective results but also improves the efficiency. Given any query, an expert blogger search engine should be also efficient in returning a ranked list of expert candidates. Since both *TSPR* and *HITS* algorithms are topic-dependent, they need to be run for each given query in real time. Compared to *PR* and *HITS* graphs, our proposed *TC* graph is also a better fit for these real time applications.

6. CONCLUSION

This paper analyzed very commonly used authority estimation approaches, like *PageRank*, *Topic-Sensitive PageRank* and *HITS*, with respect to the authority networks they are applied. Using these approaches, which are developed for web pages, directly on user authority networks, may not always return the expected outcomes. It has been shown that even the more topic-dependent *HITS* graphs may still consist of many topic-irrelevant users and interactions.

Topic-Candidate graph, which is a more topic-specific authority network, is proposed. For *PR* and *TSPR* approaches, this proposed *TC* graph provided statistically significant improvements over the more topic-independent graphs, *PR* and *HITS*. The experimental results showed that using topic-specific authority graphs provides more accurate estimates of topic-specific authority scores which can be used to improve the performance of expert finding approaches. The proposed graph also provided significant improvements in running time of the *PR*, *TSPR* and *HITS* approaches, which is also very important in terms of practical purposes.

7. ACKNOWLEDGMENTS

This research was in part supported by National Science Foundation (NSF) grant IIS-1302206. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR*, 2006.

[2] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3):127–256, 2012.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, 1998.

[4] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM*, 2003.

[5] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *DMKD*, 2003.

[6] T. H. Haveliwala. Topic-sensitive PageRank. In *WWW*, 2002.

[7] A. Java, P. Kolari, T. Finin, A. Joshi, and T. Oates. Feeds That Matter: A Study of Bloglines Subscriptions. In *ICWSM*, 2007.

[8] P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM*, 2007.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM*, 2006.

[11] N. Sahoo and R. Krishnan. Socio-temporal analysis of conversation themes in blogs by tensor factorization. In *WITS*, 2008.

[12] N. Sahoo, R. Krishnan, and J. Callan. Sampling online social networks guided by node classification. In *SCECR*, 2008.

[13] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM*, 2008.

[14] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[15] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, 2007.