Locating Query-oriented Experts in Microblog Search

Qin Chen, Yan Yang, Qinmin Hu, Liang He
Department of Computer Science and Technology
East China Normal University, China
qinchen199000@gmail.com, {yanyang,qmhu,lhe}@cs.ecnu.edu.cn

ABSTRACT

In this paper, we propose an approach to locating query-oriented experts in Microblog. We first define the experts by social influence and content relevance. Then, we adopt the BM25 model to calculate the content relevance of each account. For the social influence, we present a global-ranking algorithm as GUserRank and a topic-ranking algorithm as TUserRank after applying the LDA topic model. After that, we output the ranking expertise degree of each candidate for evaluation. Our experimental results show that the proposed approach is effective and promising. Especially, the topic-ranking algorithm achieves an improvement with 40.11% over the baseline. Furthermore, our approach does not rely on the data sets such that it can be duplicated in many fields.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval model*

General Terms

Algorithms, Experimentation

Keywords

Expert Search, Query-oriented, Microblog

1. INTRODUCTION

Nowadays, the microblog has become globally hot, which makes people conveniently and freely publish statuses, upload pictures and attach video links. More importantly, people, especially the professionals/experts, spread their personal influence by the "social-networking" service. However, it is still a big challenge to follow the experts if the users do not subscribe or "follow" them, since there are millions of active users and tons of statuses published every day.

Previous work has been done on expert search in the microblog field. For example, Ghosh et al. [6] observed that users often utilized groups or lists to manage experts in various topics. Based on the meta-data (list names and descriptions) with valuable cues for expertise, they built the Cognos system for expert search in Twitter. However, it was very vulnerable to list spamming, since their system highly depended on the accuracy of the lists created by the users, and only considered the user's personal profiles.[11] presented the expertise oriented search (EOS) system for experts ranking and associations mining on a given topic in the researcher social network. Since this system was based on co-authorship and only incorporated the user's local information, it is not a good fit for the microblogging expert search. Also, tools have been released by the service providers, such as the Who-To-Follow (WTF) by Twitter[1] and Person-Finder by Sina Weibo[2], a popular Chinese microblog platform. However, the concept of expert has not been explicitly defined.

Here we propose a novel approach to locating the microblogging experts on a given query. First, the definition of an expert is given on its social influence and content relevance, based on the observation of the experts in the social networks. Then, we adopt the classic BM25 function [13] to compute the weights as the content relevance of each account to the given query. The accounts are treated as the expert candidates sorted by the weights. After that, in order to measure the social influence globally and topically, we propose a global-ranking algorithm as GUserRank and a topic-ranking algorithm as TUserRank, based on the "forward" and "mention" interaction topologies. Finally, we score the expertise degree of each candidate by combining the social influence with the content relevance.

The main contributions of this paper are drawn as follows: (1)we propose a novel approach by combining the social influence with content relevance for expert location; (2)we integrate the LDA model into our topic-ranking algorithm, since it is necessary to promote the topic diversity on an expert's social influence; (3) the objective information, i.e., the link structure (follow topology) and the interaction structure (forward and mention topology), is fully taken into account, if we treat the account profile as subjective; (4) our experimental results confirm that it is successful to evaluate an expert on its social influence and content relevance.
The rest of this paper is organized as follows: The related work is discussed in Section 2. Our proposed approach is introduced in Section 3. The experimental details are de-

[1]http://twitter.com/#!/who_to_follow
[2]http://t.sina.com.cn/

scribed in Section 4. In Section 5, we show our experimental results, followed by the corresponding analysis and discussions in Section 6. Finally, we present some conclusions and plans for future work in Section 7.

2. RELATED WORK

A lot of related work has been done in recent years. Here we present our related work on expert search in Section 2.1, social influence in Section 2.2 and topic model in Section 2.3.

2.1 Expert Search

Expert search has been well studied in the email-based social network. Campbell et. al. [3] compared two algorithms for expert location: a content based approach that only incorporates the email text, and a graph-based ranking algorithm (HITS) that takes both of text and communication patterns into account. The better performance of the graph-based algorithm inspired us to leverage both topological relations and status content for microbloging expert search. Zhang et. al. [20] also proposed three families of expert searching strategies based on graph characteristics (e.g., degree distribution) and social characteristics (e.g., user interactions) of the email simulated social network. In [5], an expertise propagation algorithm for email and web based social networks was proposed. First, some candidates were ranked according to their probability of being experts for a certain topic. Then, a small set of the top ones were selected as seed to discover other potential experts. Though the algorithm performed well with the actual seed experts, it was not robust to the noise.

However, expert search in microblogging sites has rarely been studied. Weng et. al. [16] proposed a TwitterRank algorithm that used both link structure and tweet contents to identify influential users under particular topics. In addition, Ghosh et. al. [6] noticed that many twitter users often carefully created lists or groups to manage other users whom they considered as experts on a given topic. Thus, the information of the generated list, such as names and descriptions, often provided valuable semantic cues to help model the grouped(listed) users' topic expertise. Based on the list meta-data in Twitter, they built the Cognos systems to help find experts on a specific topic. However, this system relied much on the accuracy of the lists created by users and only considered the basic information such as users' profiles.

In addition, some application systems were developed. In a Twitter's released project report of implicit semantic indexing, the "Who Knows" system was introduced to help find appropriate experts [15]. [10] described the "Referral Web" system, which could recommend personalized experts to users by combining social networks and collaborative filtering. Instead of studying the searching strategy, [8] focused on the interface design of expert search systems. It also investigated what information should be displayed and how to display it efficiently.

2.2 Social Influence

Microblogging research has become a hot spot these years. Many researchers are engaged in how to measure user's social influence in the network effectively. In general, social influence is used to indicate the importance of users in microblogging platforms. In another word, the higher a user's social influence, the greater impact she/he will have on others. Various approaches have been proposed for social influence measurement. Most of them fall into two categories. One is to calculate social influence based on statistics, such as the number of followers and total amount of tweets [1, 4, 18]. It is ease of implementation, but only fits some simple scenarios. The alternative approach is to take user's entire topology graph and diffusion behaviors into account. Specially, it expands the graph algorithms, such as PageRank or HITS, to calculate users' social influence [14, 17]. It is usually more accurate, but harder to be implemented.

Intuitively, users' social influence may vary in different topics. For example, Bill Gates' social influence on IT is very high, but with respect to food or fashion, it may be lower than the gourmet or fashion designer. Thus, it is necessary to consider different levels of social influence. Weng et. al. [16] presented their research on topic social influence measurement. Their proposed Twitterrank algorithm was mainly based on link structure (i.e., follow) in twitter. In this paper, we also considered the topic-level social influence for microblogging expert search . What's more, we utilized the more comprehensive topology. In addition to the link structure, the interaction structure, i.e., forward or mention, has been incorporated. Also, Weng et. al. [16] computed the topic similarity between users for social influence transfer, whereas we concern the amount of topic related information a user responses to by forwarding or mentioning.

2.3 Topic Model

Topic model is often employed to mine "latent topics" from high dimensionality of terms in text. Since the probabilistic latent semantic indexing (PLSI) [8] was presented in 1999, many extensions have been proposed. Among these, LDA[2] is a well-known topic model and has been widely used in text processing for its outstanding performance. However, it is not good at processing short text [12], such as the microblogging status with 140 characters limited.

Many researchers have tried to improve the LDA model to fit the microblogging settings. There are mainly two approaches. The first is by aggregations. [9] analyzed all aggregation strategies and discovered that merging all the published statuses of an account as an input document could yield better results. The other one is by extension over the traditional LDA model. [19] expanded it by utilizing both the structured and unstructured content, which was more effective for analysis. In [21], a background model was presented to incorporate the "forward" topology. When generating status content, there was a probability to choose words contained in the background model or in the particular subject.

In summary, the first approach is easy to be implemented, since it does not involve additional model derivations. Though it neglects some information such as emotions and tags, it proves to be sufficient for most scenarios[9]. The second one incorporates the scenario specific elements for modeling, which is more difficult for derivations. In this paper, we apply the aggregation approach for topic extractions in our experiment.

3. THE PROPOSED APPROACH

In this section, we present our proposed approach to locate query-oriented experts on microblog search. Section 3.1 gives an overview of the expert search procedure. Section 3.2 demonstrates how to calculate account's content relevance for candidate selection. For each candidate, either his/her global or topic influence can be measured by the algorithms proposed in Section 3.3. By combining the content relevance and social influence, the expertise degree of each candidate is calculated in Section 3.4.

3.1 Overview

Let $U = \{u_1, u_2, ..., u_{|U|})\}$ be the account set, where $|U|$ denotes the total account number. $S = \{S_1, S_2, ..., S_{|U|}\}$ denotes the set of all accounts'published statuses and each element S_i represents u_i 's status collection. $R = \{(u_i, u_j)|$ u_i $follows$ $u_j\}$ denotes the set of follow relationships among these accounts. In particular, if u_i follows u_j, u_i is called the "follower"and u_j is called the "friend". Q is a given query and E_Q is the set of experts specific to query Q. Thus the problem can be defined as this: for the input U, S, R and Q, the goal is to output the corresponding E_Q.

Figure 1 shows an overview of our expert search approach. All the data has been stored in the database and can be accessed efficiently. Our goal is to locate the related microblogging experts on a given query.

As is shown in the figure, our proposed approach mainly includes three steps. First, most obviously irrelevant accounts are filtered out due to their low status content relevance to the query. The remaining accounts are selected as the candidate experts. Then the social influence (either global or topical) of these candidates is measured. Finally, each candidate's expertise degree is determined by both content relevance and social influence. Based on the ranking, the corresponding experts can be located and displayed.

3.2 Content Relevance

To measure the content relevance of each account to the query, all published statuses of an account are aggregated into a document. Then the BM25 model [13] is adopted to calculate the weights for the content relevance . For a given query Q, the content relevance of the document d is defined as follows:

$$BM(d,Q) = \sum_{i=1}^{w} \frac{TF(q_i)(1+k)}{TF(q_i) + k(1-b+b\frac{dl}{avgdl})} IDF(q_i) \quad (1)$$

where q_i is the i-th query keyword in Q and w is total keywords number. $TF(q_i)$ denotes the term frequency of q_i. dl represents the length of document d and $avgdl$ is the average length. k and b are two regulatory factors and usually set as $k \in [1.2, 2.0]$ and $b = 0.75$ by experience. $IDF(q_i)$ denotes the inversed document frequency of q_i which can be calculated by Eq.(2):

$$IDF(q_i) = \frac{\log(\frac{N-n(q_i)+1}{n(q_i)})}{\log(N)} \quad (2)$$

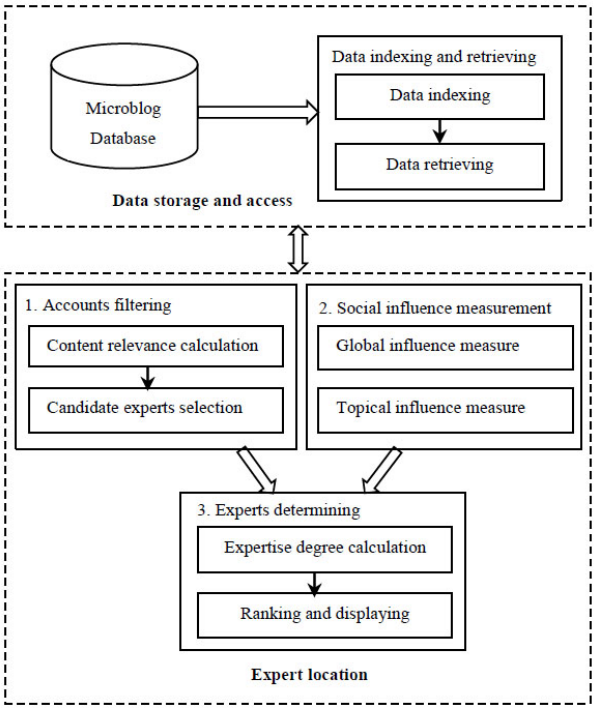

Figure 1: Architectural overview of expert search

where N is the total number of documents and $n(q_i)$ is the number of documents which contain the query keyword q_i.

With the calculated content relevance, all the accounts are ranked accordingly. The top-ranked accouts are selected as the candidates on the query.

3.3 Social Influence

For each candidate, his/her social influence is further measured. Intuitively, a microblogging account is similar to a web page in some extent. For example, a web page transfers its importance via link relationship while a microblogging account can transfer his influence by topological relations such as "follow", "forward" or "mention". What's more, the social influence may vary with topics due to different backgrounds. Thus, we are motivated to propose a global-ranking algorithm and a topic-ranking algorithm to measure the global and topical social influence of a candidate respectively.

3.3.1 Global Influence

We expand the well-known PageRank algorithm to measure the global influence. Similar to the prior TwitterRank algorithm[16], the influence is also calculated by a random surfer model according to "follow" relations. In other words, the random surfer visits each account with certain probabilities based on the "follow" relationship in R, which means the influence is transited from the candidates to their friends. Since not all published statuses of a candidate will have an impact on his followers, [16] combined the similarities with published status number for influence calculation. Different from [16], we consider the more explicit influential actions, i.e., forward and mention. Specifically, we assume that the more the statuses are forwarded, the larger influence the can-

didate will obtain. Similarly, if the candidate is mentioned frequently, he/she will also get a high influence.

Based on the above assumptions, each element in the influence transition matrix P can be defined as follows:

$$P_{ij} = \frac{ActionNum(i,j)}{\sum_{u \in F(i)} ActionNum(i,u)} \quad (3)$$

where $F(i)$is the candidate i's friends set, P_{ij} denotes the influence transition probability from i to her/his friend j, and $ActionNum(i,j)$ indicates the frequency of i's response action to j. Specifically, it can be either $ForwardNum(i,j)$ or $MentionNum(i,j)$.The former refers to the number of candidate i's forwarding statuses from j and the latter represents the number of times candidate i has mentioned candidate j.

As mentioned in [16], some candidates may follow each other in a loop with no friends outside, which will lead their influence sink without distribution. To solve this, a random jump vector E is introduced and each entry E_i indicates the probability of other candidates randomly jump to candidate i without follow relationship. In this paper, each E_i is equally set to $1/(|C|)$, where $|C|$ denotes the total candidate number.

With the transition probability matrix and random jump vector defined, the global influence of a candidate can be calculated iteratively by our global-ranking algorithm named GUserRank as follows:

$$R = dP^T R + (1-d)E \quad (4)$$

where P^T is a transpose of transition matrix P defined in Eq.(3), E is the random jump vector as set above. d is a damping factor which is between 0 and 1 to control the probability of random jump.

3.3.2 Topical Influence

In general, accounts with different backgrounds tend to publish statuses in different domains or topics, which results in a nonuniform distribution of social influence over topics for each account. For example, a researcher in computer science may have a large influence in IT domain. He may probably have thousands of followers who are interested in computer science and often forward his technical statuses or mention him. However, his influence in other topics such as food or sports may be weak since few of his statuses are related to these domains. Therefore, it is necessary to consider topical influence.

To measure the topical influence of a candidate, we obtain the topics using the LDA topic model. Section *a* describes how to extract topics from the published statuses. Then the topic-ranking algorithm as TUserRank is introduced in Section *b*.

a. Topic Extraction

The Latent Dirichlet Allocation (LDA) model [2] is applied for topic extraction due to its outstanding performance in many studies. LDA [2] is an unsupervised machine learning algorithm and is often used to extract potential semantic topics from a document collection.

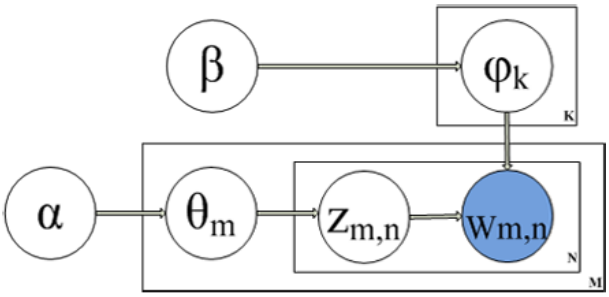

Figure 2: Graph model of LDA using plate notion

In LDA, each document is denoted as a probability distribution over topics while each topic is denoted as a probability distribution over words. This probability model can be expressed by the plate notion in Figure 2. Variables are represented by circles, shaded for observed and unshaded for latent. Arrows indicate the conditional probability dependency between variables. Boxes indicate repeated sampling and variables in the lower right corner of each box refer to the number of repetitions. α and β are the hyper parameters for symmetric dirichlet distribution.

LDA is a generation model. Assuming the number of topics, words and documents are K ,V and M respectively. For each document m, the multinomial probability distribution over topics is denoted as θ_m. And for each topic k, the multinomial probability distribution over words is denoted as φ_k. To generate a document with N words, its topic vector θ_m should firstly be chosen according to the symmetric dirichlet distribution $dir(\alpha)$. Then a topic $Z_{m,n}$ is selected based on θ_m to generate a word $W_{m,n}$ with the topic-word distribution φ_k.

As is mentioned in Section 3.2, each account corresponds to a merged document containing all his/her published statuses, which is treated as an input document in LDA[9]. Then gibbs sampling [7] is used to infer the document-topic distribution θ and topic-word distribution φ. For convenience, θ and φ can also be expressed as matrix Dt and Tw as follows:
1. Dt: Each row in matrix Dt represents one document (i.e., account) and each column represents one topic. $Dt[i][j]$ represents topic j's weight in document i.
2. Tw: Each row in matrix Tw represents one topic and each column represents one word. $Tw[i][j]$ represents word j's weight in topic i.

With matrix Dt, we get accounts' topic distribution which can further be utilized to measure the topic-specific influence. The other matrix Tw will be used to calculate expertise degree by mapping query keywords to related topics.

b. Topic-sensitive Rank

With the topics extracted by the LDA model, the topic-ranking algorithm named TUserRank, is proposed to focus on the topical influence of a candidate. It is based on the

global-ranking algorithm. In order to measure the topic-level social influence, we consider not only the number of the "forwarded" or "mentioned" statuses but also the topic relevance of these statuses. In particular, each component in the topic-sensitive transition matrix P^t for topic t is defined as:

$$P_{ij}^t = \frac{ActionNum(i,j) \times Dt[j][t]}{\sum_{u \in F(i)} ActionNum(i,u) \times Dt[u][t]} \quad (5)$$

where P_{ij}^t denotes the probability of the social influence of candidate i on topic t transferred to his friend j, Dt is the document-topic(i.e., account-topic) distribution matrix as described previously, $Dt[j][t]$ represents friend j's weight on topic t and other variables are the same as Formula(3). In addition, the random jump vector on topic t is defined as:

$$E^t = Dt'_{.t} \quad (6)$$

where Dt' is the column-normalized form of matrix Dt and $Dt'_{.t}$ represents the t-th column of Dt'. The intuition behind Formula (6) is that the bigger weight on topic t a candidate has, the higher probability other candidates will jump to him/her without follow relationship, and thus the higher influence under this topic the candidate will get.

With the newly defined transition matrix P^t and random jump vector E^t, formula (4) can be refined as follows:

$$R^t = d(P^t)^T R^t + (1-d)E^t \quad (7)$$

where R^t is the topic-specific social influence vector for all candidates on topic t.

3.4 Expertise Degree

We make a linear combination to integrate the content relevance and the social influence. The expertise degree stands for the ranking weights of a candidate on a given query. Since we observe that the social influence has a power law distribution, we adopt it as follows:

$$ED(u,q_i) = CR(u,q_i) + \log(SI(u,q_i)) \quad (8)$$

where $ED(u,q_i)$ denotes the expertise degree of a candidate u on a query keyword q_i; $CR(u,q_i)$ represents u's status content relevance with the keyword, which is calculated by the BM25 function defined in Eq.(1); $SI(u,q_i)$ indicates u's social influence, which can either be global social influence (GSI) or topical social influence(TSI).

Thus we further define the social influence in expertise degree as:

$$SI(u,q_i) = \begin{cases} R_u, & \text{(w.r.t.GSI);} \\ \frac{\sum_{t \in MT(q_i)} Tw[t][q_i] \times R_u^t}{\sum_{t \in MT(q_i)} Tw[t][q_i]}, & \text{(w.r.t.TSI).} \end{cases} \quad (9)$$

In fact, the global social influence of candidate u is independent on the query and can be calculated by Formula (4), where R_u denotes the u-th row in influence vector R. As for the topical social influence (TSI), we firstly map the query keyword $q_i$to its related topics set $MT(q_i)$ according to the topic-word distribution matrix Tw. By adding the influence of candidate u in these related topics, we can get his/her topical influence on the given query.

4. EXPERIMENTAL SETUP

We introduce our experimental setup in this section, including the description of the data, the preprocessing and the evaluation methods.

4.1 Datasets

We collected the data set via the API[3] provided by Sina Weibo, the Chinese microblog platform. Firstly, a seed account was randomly picked out. Then, the data of the seed's friends and his/her friends' friends was crawled. There are 20,445 accounts in total. However, we randomly selected 1000 accounts in the experiment for convenience. For each account, the data consists of the basic information, the published statuses and the "follow" relationships. Each account's interaction information such as "forward"(labeled with "//@") and "mention"(labeled with "@") can be extracted from status content. The statistics of our dataset are shown in Table 1.

Table 1: Statistics of Microblogging Dataset

Statistics	Value
Num. of Accounts	1,000
Num. of Relationship Pairs	5,355
Num. of Published Statuses	1,776,012

Figure 3 shows the distribution of each account's follower number. We observe that only 57% of the accounts have at least a follower in the dataset, while the remaining 43% have no followers. For the accounts who have followers, about half of them have more than 4 followers. It is not hard to understand because the amount of experimental accounts is quite small (1000) and they are randomly picked from the crawled data set. Figure 4 shows the distribution of microblog(status) number. 98% of the accounts have published more than 100 microblogs and about 20% accounts have published more than 2000.

4.2 Preprocessing

Before starting our experiment, we did some preprocessing for the raw data which mainly includes status mergence, data cleaning and word segmentation. In addition, the query set related to various domains is also prepared as follows: Q={"Internet ", "Food ", "Business ", "Fashion", "Market ", "Music ", "Phone"}, where Q denotes the query keyword set.

Then we performed our experiments with three approaches: BM25, BM25+GUserRank and BM25+TUserRank. The BM25 model is the baseline and the other two are proposed with either global or topical influence incorporated. The results and performance analysis are demonstrated in Section 5 and 6 respectively.

[3]http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9A API

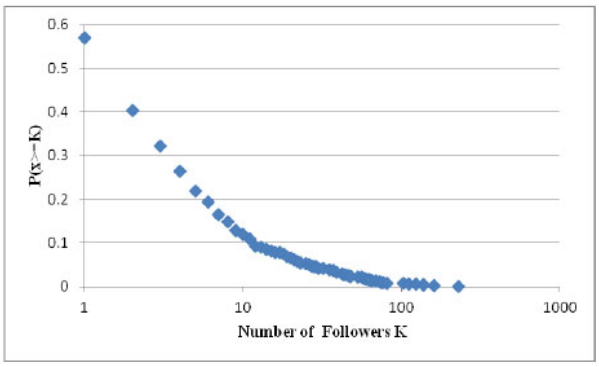

Figure 3: Distribution of follower number

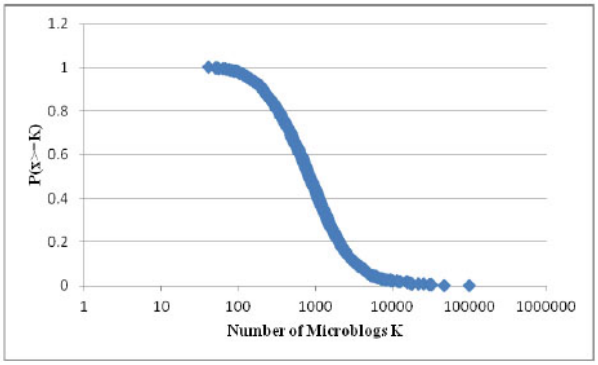

Figure 4: Distribution of microblog number

4.3 Evaluation

To evaluate the performance of our proposed approaches, an "Expert Voting" system was developed to gain the standard expertise degree from the volunteers in Sina Weibo. Each voter chose a keyword she/he was interested in. Then the corresponding candidate experts were displayed to the voter. The voter could click on each candidate's avatar to view related information (e.g., account profiles, published statuses) and then rated for the expertise degree specific to the keyword. The rating ranges from 1 to 10 and a larger value indicates a higher expertise degree of the candidate. Finally, the average rate for each candidate was normalized to [0,1] as the standard expertise degree on the specific keyword.

With the obtained artificial expertise rating, mean average precision(i.e.,MAP), a general rank-sensitive evaluation metric was applied. Mean average precision for the retrieval list length of N is defined as:

$$MAP@N = \frac{\sum_{i=1}^{N} score(i)P@i}{\sum_{i=1}^{N} score(i)} \quad (10)$$

where $score(i)$ represents the i-th standard expertise degree in the list and $P@i$ represents the top i average precision which can be defined as:

$$P@i = \sum_{j=1}^{i} \frac{score(j)}{i} \quad (11)$$

5. EXPERIMENTAL RESULTS

Table 2 shows the MAP performance of the BM25 model and the proposed approaches combined with either global-ranking algorithm or topic-ranking algorithm, for different top "N" retrieved experts. The former proposed denotes as BM25+GUserRank and the latter as BM25+TUserRank. The "forward" and "mention" topologies are applied in the global-ranking and the topic-ranking algorithms. Here we adopt the BM25 model as the baseline. The top "N" experts are tested as {10, 20, 30, 40, 50}.

Corresponding to Table 2, Table 3 presents the improvements of the proposed approaches over the BM25 model. The best relative rates are marked as bold.

Table 2: MAP@N values of different approaches

N	BM25	BM25+GUserRank		BM25+TUserRank	
		forward	mention	forward	mention
10	**0.5495**	**0.5487**	**0.5073**	**0.7061**	**0.7061**
20	0.4671	0.4828	0.4603	0.6335	0.6337
30	0.4227	0.4518	0.4333	0.5896	0.5898
40	0.4062	0.4404	0.4215	0.5690	0.5692
50	0.4025	0.4364	0.4175	0.5611	0.5612

Table 3: Improvements of the proposed approaches over the baseline

N	BM25+GUserRank		BM25+TUserRank	
	forward	mention	forward	mention
10	-0.14%	-7.67%	28.50%	28.50%
20	3.36%	-1.45%	35.62%	35.67%
30	6.87%	2.49%	39.48%	39.52%
40	**8.42%**	**3.77%**	**40.07%**	**40.11%**
50	8.42%	3.72%	39.38%	39.42%

6. ANALYSIS AND DISCUSSIONS

Here we analyze our experimental results in the following five aspects: (1) the performance comparisons among the BM25 model, the global-ranking combined approach and the topic-ranking combined one; (2) the effectiveness of social influence; (3) the effectiveness of content relevance ;(4) the impact of the topologies and (5)the investigation on the retrieval list length "N".

6.1 Performance on Different Approaches

In order to investigate the influence of our proposed global-ranking and the topic-ranking combined approaches, we analyze the experimental results in Section 5. To illustrate the results in Table 2 graphically, we re-plot these data in Figure 5. The x-axis represents the top N values and the y-axis shows the MAP performance. The "_ f " or "_ m" suffixation in the legend label indicates the topology applied, i.e., "forward" or "mention".

We observe that both of the proposed approaches outperform the baseline BM25 model. Specifically, when the retrieval list length equals to 10, the proposed BM25+GUserRank approach does not perform as well as expected. As shown in Table 3, its performance (MAP@10) with mention topology incorporated has reduced by 7.67% when compared with

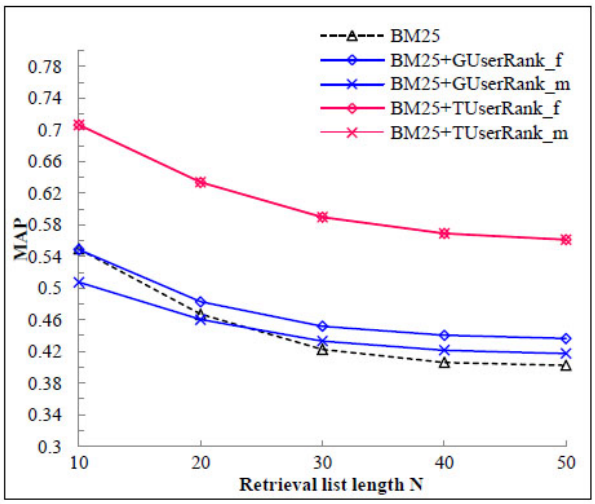

Figure 5: Performance of different approaches

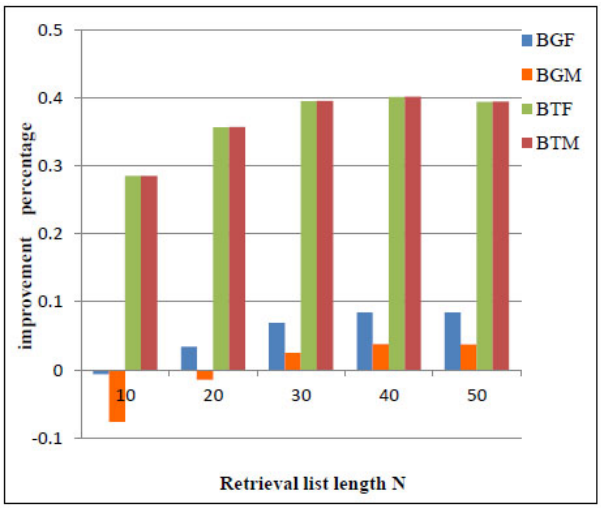

Figure 6: Investigation of Diffferent "N"

the baseline. However, as the list grows, it gradually outperforms the baseline. The best approach is the proposed BM25+TUserRank, which significantly outperforms the baseline with all top N values.

6.2 Effectiveness of Social Influence

Taking the social influence for expert search is the major contribution in our approaches. As is mentioned above, with either global or topical influence incorporated, our approaches can yield better performance. Hence, we can conclude that taking user's social influence into account can help improve the precision of expert search, which has also verified the appropriateness of our expert definition in Section 1.

In addition, it is worth noting that the proposed global-ranking combined approach(i.e.,BM25+GUserRank) gets a slight improvement over the baseline.Whereas, the other approach, BM25+TUserRank, has gained a significant promotion by over 25%.This can be explained by the different levels of influence they considered. Though the influence calculation algorithms, GUserRank and TUserRank, are both based on the well-known PageRank, their transition matrices are different. The former mainly focuses on account's topological relationship and the amount of interacted information by the topology. It does not care about the semantics of the current query. The latter has taken the topic relevance of the interacted information into account. In fact, for a given query, it first maps it into several related topics and then measures the topics related influence instead of the global one, which seems more in line with the intention of inquirers. The different performance between the two approaches also coincides with our intuition that users may have different influence on different topics. Therefore, topic-sensitive influence is more efficient than the global for expert search.

6.3 Effectiveness of Content Relevance

Both Table 2 and Figure 5 show that the mean average precision of all the approaches drops consistently when the retrieval list length grows. Noting that the baseline BM25 model only utilizes content relevance for expert search, thus its declined performance is completely caused by the reduced content relevance. It can be further concluded that in addition to social influence, the other component of our expert definition, content relevance, also plays an important role in expert search.

6.4 Impact of Topologies

In this paper, we mainly consider the interaction topology, i.e., "forward" or "mention", for social influence calculation. Then the calculated influence is utilized to locate experts. Due to the better performance of the proposed approaches, we can infer that either topology can help propagate social influence and further promote the expert search precision.

Whether the two topologies differ in the influence propagation ability is also analyzed. From Table 3, we observe that the performance of the BM25+GUserRank approach is affected by the topology used. Concretely, the best improvement based on "forward" is 8.42%. While based on "mention", it is only 3.77%. In other words, the forward topology seems to be more efficient than the mention topology for global influence propagation. This can be explained by the characteristics of accounts' activities in microblogging sites. Generally speaking, microblogging accounts tend to forward the statuses which are actually interesting. Once these statuses are forwarded, they will be probably read by more accounts and forwarded again since they are really valuable in some extent. Thus the influence of the accounts who originally create the statuses will propagate to the whole social network. When it comes to another topology, accounts usually mention those who are their close friends in real life with no regard to the content quality. Thus, it brings about the illusion of influence propagation. However, this phenomenon does not exist in the BM25+TUserRank approach since it introduces the topic relevance for content filtering, which helps reduce the side effect to some extent.

6.5 Investigation on "N"

In order to investigate the influence of the retrieval list length "N", we re-plot Table 3 in Figure 6. The x-axis represents the top N values and the y-axis shows the relative rates of improvements. The legend label "BGF" indicates the "B-

M25+GUserRank" approach with "forward" topology incorporated. Other labels are also denoted in this way.

We can observe that we achieve the best rate of 40.11% when "N" is around 40. It is very interesting that when "N" is as small as 10, we also get the best performance with an improvement of 28.50% by BM25+TUserRank approach. What's more, the larger of "N", the stable the improvements are.

7. CONCLUSIONS AND FUTURE WORK

The conclusions of our work are fourth-fold. First, we propose novel approaches to locating experts in microblog search, where the BM25 model is adopted as a baseline, the global-ranking and topic-ranking combined approaches are presented as the main contributions. Second, we take social influence into account to evaluate an expert, not only at the global level but also at the topic level. Third, the topic-sensitive social influence is well modeled in our approach as another unique contribution, through integrating a LDA model. Fourth, we investigate the social influence, the content relevance and the impact of topologies in the experiments and conclude that they are all positive on the results.

In the future, we will continue on studying the expert search comprehensively, especially on the topic-sensitive aspect. This is also our ongoing work.

ACKNOWLEDGMENTS

We would like to appreciate the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. This work is supported by the Shanghai Science and Technology Commission Foundation (No. 12dz1500205 and No. 13430710100).

8. REFERENCES

[1] I. Anger and C. Kittl. Measuring influence on twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, page 31. ACM, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 528–531, New York, NY, USA, 2003. ACM.

[4] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.

[5] Y. Fu, R. Xiang, Y. Liu, M. Zhang, and S. Ma. Finding experts using social network analysis. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 77–80. IEEE Computer Society, 2007.

[6] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.

[7] G. Heinrich. Parameter estimation for text analysis. Technical report, Technical report, 2005.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[9] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[10] H. Kautz, B. Selman, and M. Shah. Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[11] J. Li, J. Tang, J. Zhang, Q. Luo, Y. Liu, and M. Hong. Eos: expertise oriented search using social networks. In *Proceedings of the 16th international conference on World Wide Web*, pages 1271–1272. ACM, 2007.

[12] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web*, pages 121–130. ACM, 2008.

[13] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[14] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*, pages 18–33. Springer, 2011.

[15] L. A. Streeter and K. E. Lochbaum. Who knows: A system based on automatic representation of semantic structure. In *RIAO'88*, pages 379–389, 1988.

[16] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[17] Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa. Turank: Twitter user ranking based on user-tweet graph analysis. In *Web Information Systems Engineering–WISE 2010*, pages 240–253. Springer, 2010.

[18] S. Ye and S. F. Wu. Measuring message propagation and social influence on twitter. com. In *Social informatics*, pages 216–231. Springer, 2010.

[19] C. Zhang and J. Sun. Large scale microblog mining using distributed mb-lda. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1035–1042. ACM, 2012.

[20] J. Zhang and M. S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 71–80. ACM, 2005.

[21] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.