ISSN 2321 3361 © 2018 IJESC

Research Article Volume 8 Issue No.5

Finding the Topic-Specific Twitter Expert using JGibbLDA and TF-IDF

Ashily .M Baby[1], Sarju .S[2]
M.Tech Scholar[1], Assistant Professor[2]
Department of Computer Science and Technology
St. Joseph College of Engineering and Technology, Palai, India

Abstract:
Twitter, a famous long range interpersonal communication stage, gives a medium to individuals to impart data and insights with their supporters. Discovering experts on Twitter is an essential issue since tweets from experts are significant sources that convey rich data in different spaces. In any case, past strategies can't be specifically connected to the Twitter expert discovering issue. As of late, a few endeavours utilize the relations among users and Twitter Lists for expert finding. Nevertheless, these approaches only partially utilize such relations. In this paper, propose a new strategy for expert finding which is based on JGibbLDA and TF-IDF. Using JGibbLDA done topic modelling, topics and related words are extracted from the documents. For finding topics name, the related words analyse or search in the Wikipedia contents. This enables to identify words which are semantically related to the topic query along with the other words which are directly related to the topic query. This strategy double folds the chances of selecting words which are related to the topic query. Then using term frequency-inverse document frequency (TF-IDF), the weight of words in the documents are calculated and then find the K no. of experts in a particular domain. The static dataset collected via Twitter API. Analyses on true information will exhibit the adequacy of proposed approach for the subject particular expert finding on Twitter.

Keywords: Expert search, twitter, Wikipedia, JGibbLDA, TF-IDF

I. INTRODUCTION

Expert finding addresses the task of identifying the right person with the appropriate skills and knowledge. Experts can be required for a variety of purposes: problem solving, question answering, providing more detailed information on a topic, to name a few. The expert finding task has attracted a great deal of interest within the Information Retrieval (IR) community over the past few years. Recent years, expert finding has attracted much attention due to the rapid flourish of the Web 2.0 applications and the advancement of information retrieval technologies from the traditional document-level to the object-level. The need in finding a well informed person may be critical for any kind of project. Much research works have been done to solve the expert finding problems. Ideally, given a topic of interest, one would hope to find users who provide credible information about the topic. Credibility is often conceived as a function of expertise and trust, and expertise is commonly defined by the support and nomination of other professionals in the domain. Therefore, to find credible sources of information, one should seek out people who not only frequently publish topically relevant tweets but also are trusted by their peers. Unfortunately, there is no simple way for a normal user to observe how trusted someone is in a particular field. Micro-blogging sites, out of which Twitter is the most popular, have emerged as an important platform for exchanging real-time information on the Web. Recent estimates suggest that 200 million active Twitter users post 150 million tweets (messages) daily. Expert finding has become a hot topic with the growth of social network. The expert finding task can be described as follow: given a set of documents or tweets, a list of users names, and a set of topics, the goal is to find experts from the list of users names for each of these topics. In this paper, done the topic modelling using JGibbLDA. For finding topics name, the related words analyse or search in the Wikipedia contents. And using TF-IDF ranking the persons who have expertise in particular topics, then which expert may be rank higher, which is closer to the truth.

II. RELATED WORKS

In the paper [1], twitter rank: Finding point delicate persuasive Twitterers, it center around recognizing powerful clients of small scale blogging administrations. The Twitter which is a small scale blogging administration utilizes a long range interpersonal communication demonstrate called following, influencing the clients to pick whom they to need to take after to get tweets. It was discovered that 72.4% of the clients of Twitter take after over 80% of their devotees and 80.5% of the clients have 80% of clients they are following tail them back. The Twitter rank calculation is utilized to quantify the impact of clients in twitter .This is an expansion of Page rank calculation which just measures the impact in view of the connection structure of system. The Twitter rank calculation measures the impact in view of connection structure and topical comparability between clients. It is superior to anything the Page rank calculation. The strategy comprises of three procedures: Topic refining, Relationship development and Ranking. The subject refining naturally distinguish points that twitterers are intrigued in view of the tweets distributed by them. At that point a system is built in view of the theme particular connection amongst clients and their supporters. At long last a point delicate client impact positioning procedure happens which gives us important rundown of clients who is affected on a specific subject in twitter. Twitter Rank works in two stages; initial one is utilizes Latent Dirichlet Allocation (LDA) show. It sees the themes of free in light of their tweets. Second one for every theme it assembles a chart weighted by taking both the topical comparability in the middle of two clients and supporter diagram, at that point additionally enlist page rank calculation for discover subject particular powerful clients. A.Pal and S.Counts [2] proposed Identifying topical authorities in micro-blogs which proposed a set of features for authors that includes nodal and topical metrics. A probabilistic clustering and within clustering procedures gives a final list of top authors for a particular topic. The algorithm used was found more flexible in the real world scenarios. The tweets are classified into three: Original Tweet (OT), Conversational Tweet (CT) Repeated Tweet (RT). The original tweets are those tweets produced by authors. The CT is directed at another user. The RT is produced by someone else but user copies or forwards it. Around a user some metrics like number of original tweets; number of links shared etc are computed. A self similarity score is determined which gives the measure of how many words a user borrows from the previous tweets. Some of the textual features extracted for a user include topical signal (TS) and signal strength (SS).The TS gives a measure of the involvement of user in a topic. The SS gives the strength of TS i.e. true authority of a user on a topic. A Gaussian mixture model is used to cluster users. This clustering aims at reducing the target cluster which consists of authoritative users. They also showed that probabilistic clustering it is a way to filter a large chunk of outliers in the feature space. At last they allow that Gaussian based ranking it is helpful to rank users and more effective way for finding top1 ranked authors. In this paper proposed features and methods that could be used to produce a ranked list of top authors for a given topic for identifying topical authorities in micro blogging environments. Here proposed a number of features of authors and observe that topical signal and mention impact are slightly more important than other features. Here also showed that probabilistic clustering is an effective way to filter a large chunk of outliers in the feature space (either long tail or celebrities) and select high authority users on which ranking can be applied more robustly. Finding topic experts on micro blogging sites with millions of users, such as Twitter, is a hard and challenging problem. In this paper[3], propose and investigate a new methodology for discovering topic experts in the popular Twitter social network. The paper, Cognos: Crowd sourcing search for topic experts in micro-blogs make use of twitter lists which are created by individual users that includes experts and their topics interested by them. These metadata provides information regarding experts and their domain of expertise. The list information is mined to build a system called cognos to find topic experts in twitter. The twitter list can be seen in the form of a list graph and it can be connected to a follower graph via member of relation and subject to relation. The twitter list can be observed in the form of a table which gives information like list name, description and members. The list name gives the relevant topic, description gives details of topics and members gives the name of experts in the relevant topic. Since cognos act as a list feature it is indeed a who-to-follow system in twitter .Here ranking procedure is based on list feature. It was found that the performance of cognos was better compared to the conventional methods. The crowd sourced search helps to build future content search. This have mined List information to build Cognos, a system for finding topic experts in Twitter. Cognos infers a users expertise more accurately and comprehensively than state-of-the-art systems that rely on the users bio or tweet content. N. k.sharma has concentrated [5] on gathering who will be who in the twitter informal organization twitter list: they propose to utilize twitter rundown to distinguish the nature of twitter clients by twitter crowed. Rundown contain the clients and to register comparability between every client and given subject question. This is utilized to pursuit and rank every one of the clients. Utilizing congnos move to pick the clients that clients contained in more number of records those Meta information contain the inquiry. It utilize twitter rundown to recognize the nature of twitter clients. In this paper, they outline and assess a novel who-will be who benefit for surmising traits that portray singular Twitter clients. This procedure abuses the Lists include, which enables a client to assemble different clients who tend to tweet on a point that is important to her, and take after their aggregate tweets. Our key understanding is that the List meta-information (names and depictions) gives profitable semantic signals about who the clients incorporated into the Lists are, including their subjects of mastery and how they are seen by the general population. Therefore, we can construe a clients aptitude by dissecting the meta-information of crowd sourced Lists that contain the client. In [6], Users interface with online news from multiple points of view, one of them being sharing substance through online person to person communication destinations, for example, Twitter. There is a little however imperative gathering of clients that dedicate a significant measure of exertion and care to this action. These clients screen a substantial assortment of sources on a theme or around a story, precisely select fascinating material on this subject, and scatter it to an intrigued gathering of people running from thousands to millions. These clients are news guardians, and are the primary subject of investigation of this paper. In this paper, receive the point of view of a writer or news supervisor who needs to find news keepers among the gathering of people drew in with a news site. Here, take a gander at the clients who shared a news story on Twitter and endeavor to distinguish news custodians who may give more data identified with that story. In this paper, depict how to locate this particular class of guardians, which they allude to as news story keepers. Henceforth, they continue to process an arrangement of highlights for every client, and exhibit that they can be utilized to naturally discover important custodians among the group of onlookers of two expansive news associations. There are distinctive sorts of clients that total substance. They vary in the quantity of points they cover (centered/unfocused), in how much analysis they incorporate (just URLs or URLs and remarks) and in the way they post data (human/programmed). These experiences enable us to make a first portrayal of news story caretakers. This paper [7] centers around discovering specialists utilizing Email correspondence. Email records are seen most suited to this undertaking of skill area as individuals routinely impart what they know. Email give a simple to mine vault of correspondence between individuals in the interpersonal organization and it contains real exhibits of skill and in addition information of mastery. In this, points are created through unsupervised bunching of message substance and watchword seeking is empowered through standard data recovery procedures keep running on client provided catchphrases and message content. A typical strategy for discovering data in an association is to utilize social net works. Email records appear to be especially appropriate to this assignment of aptitude area , as individuals routinely convey what they know. In addition, since individuals expressly guide email to each other, informal organizations are probably going to be contained in the examples of correspondence. Two calculation for deciding ability from email were thought about: a substance based approach that considers just of email content and a chart based positioning calculation (HITS) that considers both of content and correspondence designs. Email demonstrates who speaks with whom and what those interchanges are about adequately giving a window onto casual interpersonal organizations. Email shows exhibitions of skill as well as information of who comprehends what. The decision of who to send an inquiry to depends on the senders learning of mastery, and in addition information of topic.

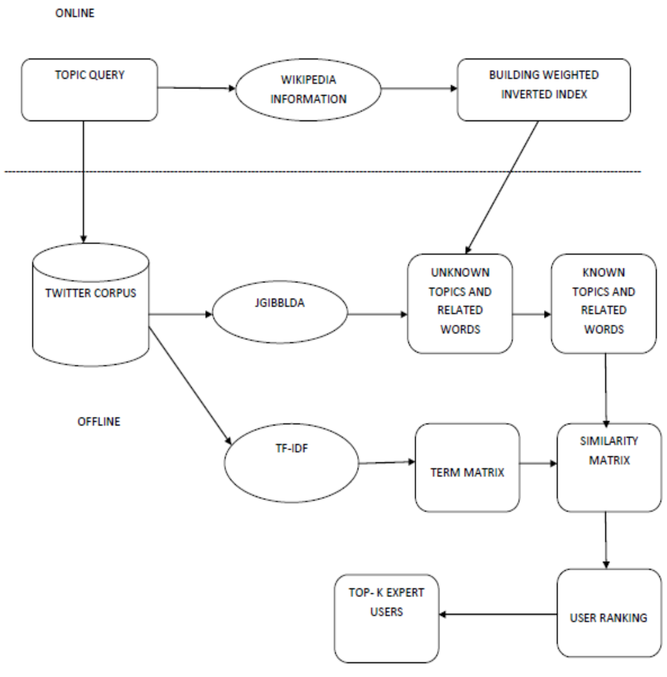
III. PROPOSED METHODOLOGY

In this project, propose a new strategy for expert finding which is based on JGibbLDA and TF-IDF. Using JGibbLDA done topic modelling, topics and related words are extracted from the documents. For finding topics name, the related words analyse or search in the Wikipedia contents. This enables to identify words which are semantically related to the topic query along with the other words which are directly related to the topic query. This strategy double folds the chances of selecting words which are related to the topic query. Then using term frequency-inverse document frequency (TF-IDF), the weight of words in the documents are calculated and then find the K no.of experts in a particular domain. Details are given in figure.

A. System Architecture

The overall architecture of the system is given in the below. This project uses the JGibbLDA for topic modelling and uses TF-IDF for ranking the persons who have expertise in particular topics. From the input dataset, different topics are determined using JGibbLDA. Under each topics various related words are contained. For finding topics name, the related words analyse or search in the Wikipedia contents. This enables to identify words which are semantically related to the topic query along with the other words which are directly related to the topic query. Then create term matrix and find the top K-users in a particular domain. The data corpus that has been pre-processed by cleaning, tokenization and stop word removal. The proposed architecture consists of mainly three steps:

- Extracts various topics and related words using JGibbLDA from dataset corresponding to the topic query.
- Extracts contents online from Wikipedia corresponding to the topic query.
- Ranking the persons who have relevant expertise in particular topics using TF-IDF.



B. JGibbLDA for topic modelling

In this venture the GibbsLDA separate subjects and related words from the pre-handled information dataset, which acquired from the twitter. Gibbs inspecting begins with allotting esteems for all factors included. At that point one of these factors is selected and its esteem is recalculated expecting the various esteems are right. A next factor is picked, until the point when the whole arrangement of factors meets to specific esteems. Also, even that isn't vital. After a certain"burn-in period" it does the trick to run the Gibbs sampler for some time and normal the estimations of the factors after that period. Since the updates of the Gibbs sampler just modify the arrangement each progression for only a smidgen it is prudent to just consider tests every S steps. Example based portrayals are viewed as more important and more exact to speak to subjects than word based portrayals.

C. Gibbs sampling algorithm for LDA

```
zero all count variables
for all documents m
  for all words n in document m
    sample topic index z_{mn} = k \sim Mult(1/K)
    increment document-topic count: n_m^k + 1
    increment document-topic sum: n_m + 1
    increment term-topic count: n_k^t + 1
    increment term-topic sum: n_k + 1
while not converged
  for all documents m
    for all words n in document m
      get current topic-term assignment: (k, t)
      decrement counts and sums: n_m^k - 1, n_m - 1, n_k^t - 1, n_k - 1
      sample topic: k \sim p(z_i|z_-i, w)
      set new topic-term assignment (k, t)
      increment counts and sums
  set converged to true if converged
set parameters \phi and \theta
```

D. Term Frequency-Inverse Document Frequency

Users in Twitter have rich aptitude on different themes and finding these subject particular specialists clears an approach to empower others to recover or take after the pertinent and reliable data on a particular point in smaller scale blogging administrations. For instance, on the off chance that some individual is tweeting about the motion picture survey, there will be heaps of tweets and we need to discover who is the master in giving audits. Additionally there are numerous calculations for finding the master audit. A calculation utilized for this is tf-idf (term recurrence opposite report recurrence). In data recovery, tf-idf, short for term recurrence converse archive recurrence, is a numerical measurement that is planned to reflect how vital a word is to a record in an accumulation or corpus. It is frequently utilized as a weighting factor in data recovery, content mining, and client demonstrating. After subject demonstrating utilizing TF-IDF positioning the specialists, for that make a term grid.

- TF(t)= (no. of times term t appears in a document)/(total no. of terms in the document)
- IDF(t)=log(total no. of documents/no. of documents with term in it)

The tf-idf value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Now a days, tf-idf is one of the most popular term-weighting schemes. Calculating the weight of the terms using

NLP based TF-IDF method. Using tf-idf, weight will be calculated using the equation

W i, j = TF i, j *log(N/DFi)

Where TFi,j is the no .of occurance of i in j, DFi is the no. of documents containing in i and N is the total no.of documents. Then by using TF-IDF method ranking the N-top experts in a particular domain.

IV. RESULT AND DISCUSSION

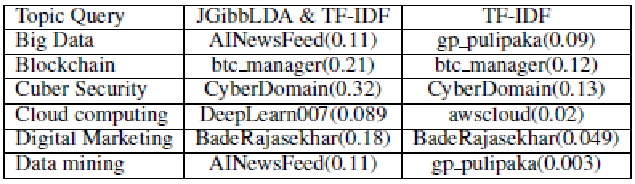
A. Theoretical Evaluation

This section contains a number of theoretical evaluations carried out in the various steps of the system. Note that theoretical analysis means evaluating the system theoretically. It is not always true that practical evaluation follows theoretical evaluation. Here evaluation is done manually by evaluating the results of expert search based on information in Wikipedia, The data set used in this evaluation was obtained by crawling using Twitter API. The information extracted for each user are user profiles, tweets, follower list and user-list subscribe information. In total 20,000 sample documents were extracted for this project. documents also consisted of special characters along with English words. Queries The queries used for experiment include various domain topics such as Big data, Cloud Computing, Blockchain and so on. There are around 10 topics used for the experiment. These queries are then used for expert finding and is used to evaluate the effectiveness of proposed method. To evaluate the quality of expert search result of proposed method, aggregate the top 10 users returned by each evaluated method. By manual evaluation, proposed method gave right expert finding for each topic query documents also consisted of special characters along with English words This proves the effectiveness of our proposed method.

B. Experimental Evaluation

Compare the result that obtained from the TF-IDF with JGibbLDA method and result that obtained from the TF-IDF without JGibbLDA method. Here six topics are taken for the experimental evaluation. The below table shown the comparison between JGibbLDA, TF-IDF method and TF-IDF method. The table 1 consist of Topic Query, experts in different methods. From the score of experts, we clearly get that which method is better one. Here the TF-IDF with JGibbLDA method is better than the other method. The figure 1 is a Topic-Accuracy graph of TF-IDF with JGibbLDA method and figure 2 is a Topic-Accuracy graph of TF-IDF without JGibbLDA method. From the accuracy graphs clearly understood that which one performance is better. The figure 3 the mean accuracy graph. The TF-IDF with JGibbLDA method have 70% accuracy and The TF-IDF without JGibbLDA method have 23% accuracy. TF-IDF with JGibbLDA method performance is better.

Table.1. Comparison between jgibblda, tf-idf method and tf-idf method



Topic Query	JGibbLDA & TF-IDF	TF-IDF
Big Data	AINewsFeed(0.11)	gp_pulipaka(0.09)
Blockchain	btc_manager(0.21)	btc_manager(0.12)
Cuber Security	CyberDomain(0.32)	CyberDomain(0.13)
Cloud computing	DeepLearn007(0.089	awscloud(0.02)
Digital Marketing	BadeRajasekhar(0.18)	BadeRajasekhar(0.049)
Data mining	AINewsFeed(0.11)	gp_pulipaka(0.003)

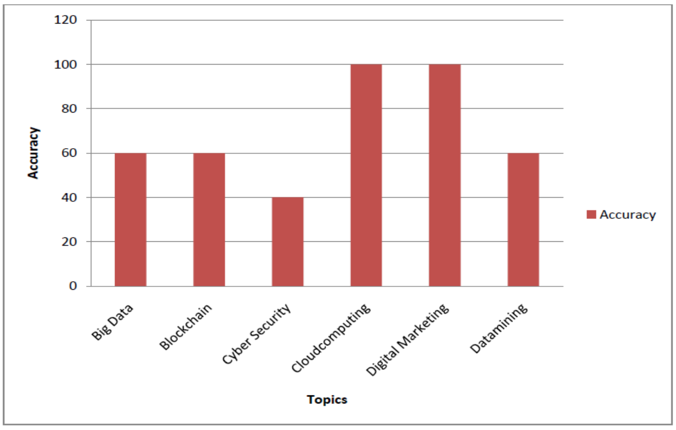


Figure.1. Topic-Accuracy graph of TF-IDF with JGibbLDA method

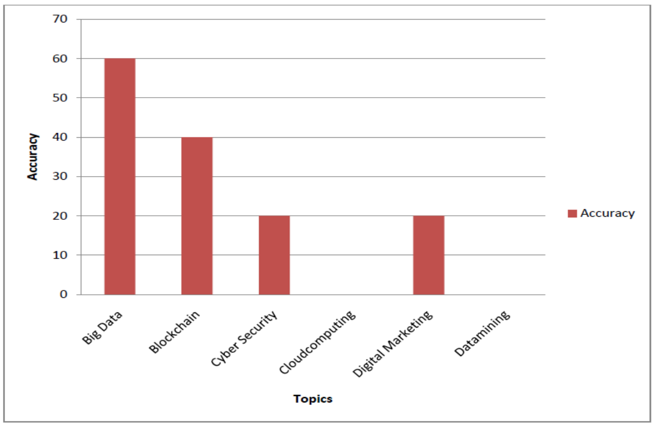


Figure.2. Topic-Accuracy graph of TF-IDF withot JGibbLDA method

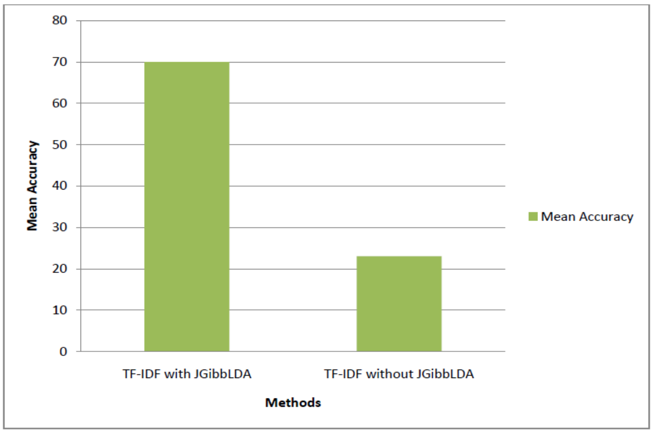


Figure 3: Mean Accuracy Graph

V. CONCLUSION

Expert finding addresses the task of identifying the right person with the appropriate skills and knowledge. Earlier expert finding method uses local relevance between the topic query and documents which reduces the chances of finding expert. For finding expert, it uses the occurrence of the topic query. In this paper, we propose a new strategy for expert finding which is based on JGibbLDA and TF-IDF. Using JGibbLDA done topic modelling, topics and related words are extracted from the documents. For finding topics name, the related words analyse or search in the Wikipedia contents. This enables to identify words which are semantically related to the topic query along with the other words which are directly related to the topic query. This strategy double folds the chances of selecting words which are related to the topic query.Then using term frequency-inverse document frequency (TFIDF), the weight of words in the documents are calculated and then find the K no.of experts in a particular domain. To the best of knowledge, this is the first attempt that targets expert finding problem in Twitter by utilizing all of such information. Then by using TF-IDF method determine the experts in particular domain. This project proposes a new strategy for expert finding using JGibbLDA and TF-IDF. Using JGibbLDA find the various topics from the data that collected from the Twitter API, then by using TF-IDF, ranking the persons who have experts in particular topics.

ACKNOWLEDGEMENT

I express sincere gratitude to my project guide, Mr. Sarju S for his guidance and support.

VI. REFERENCES

[1]. J. Weng, E.-P. Lim, J. Jiang, and Q. He, Twitterrank: Finding topic-sensitive influential Twitterers, in Proc. ACM Int. Conf. Web Search Data Mining, 2010, pp. 261270.

[2]. A. Pal and S. Counts, Identifying topical authorities in microblogs, in Proc. ACM Int. Conf. Web Search Data Mining, 2011, pp. 4554.

[3]. S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, Cognos: Crowdsourcing search for topic experts in microblogs, in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2012, pp. 575590.

[4].] X. Liu, S. Zhang, F. Wei, and M. Zhou, Recognizing named entities in tweets, in Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 359-367.

[5]. N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi, Inferring who-is-who in the Twitter social network, ACM SIGCOMM Comput. Commun. Rev., vol. 42, no. 4, pp. 533538, 2012.

[6]. J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman, Finding news curators in Twitter, in Proc. Int. Conf. World Wide Web, 2013, pp. 469478.

[7]. C. S. Campbell, P.-P. Maglio, A. Cozzi, and B. Dom, Expertise identification using email communications, in Proc. ACM Conf. Inf. Knowl. Manag., 2003, pp. 528531.

[8]. B. Gao, T.-Y. Liu, W. Wei, T.-F. Wang, and H. Li, Semi-supervised ranking on very large graphs with rich metadata, in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 96104.

[9]. D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learning Res., vol. 3, pp. 9931022, 2003.

[10]. G. Salton and C. Buckley, Term weighting approaches in automatic text retrieval, Inform. Process. Manage. vol. 24, no. 5, pp. 513523, 1987.

[11]. V. Qazvinian, E. Rosengren, D.-R. Radev, and Q.-Z. Mei, Rumor has it: Identifying misinformation in microblogs, in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 15891599.

[12]. Chen, Z.-Y. Liu, and M.-S. Sun, Expert finding for microblog misinformation identification, in Proc. Int. Conf. Comput. Linguistics, 2012, pp. 703-712.