ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Finding the Topic-Specific Expert in Twitter using Pattern Mining and TF-IDF

Ashily M Baby
Computer Science and Engineering
St.Joseph College of Engineering and Technology
Kottayam, India
ashily77@gmail.com

Abstract—Twitter, a popular social networking platform, pro-vides a medium for people to share information and opinions with their followers. Finding experts in Twitter is an important problem because tweets from experts are valuable sources that carry rich information in various domains. However, previous methods cannot be directly applied to Twitter expert finding problem. Recently, several attempts use the relations among users and Twitter Lists for expert finding. Nevertheless, these approaches only partially utilize such relations. This project focuses on developing a probabilistic method to jointly exploit three types of relations such as follower relation, user-list relation, and list-list relation for finding experts. It consists of two components, namely, an offline graph-based ranking algorithm to learn the global authority of each candidate and an online ranking model to select top-N relevant experts on the given query. In this project, Latent Dirichlet Allocation (LDA) using JGibbLDA is used for extracting the topical features and Term FrequencyInverse Document Frequency (TF-IDF) based search analysis on word/s weightage will be conducted. The data set will be crawled via Twitter API. Experiments on real-world data will demonstrate the effectiveness of proposed approach for topic-specific expert finding in Twitter.

Index Terms—Expert search, twitter, list, graph based ranking, JGibbLDA, tf-idf.

I. INTRODUCTION

Recently, the Twitter microblogging site has emerged as an important source of real-time information on the Web. Millions of users with varying backgrounds and levels of expertise post about topics that interest them. As a result, the quality of information posted in Twitter is highly variable and finding the users that are recognized sources of relevant and trustworthy information on specific topics (i.e. topical experts) is a key challenge. Identifying topic experts is also the first step towards finding authoritative information on the topic. Twitter is an online social networking service and microblogging service that enables its users to send and read text-based posts of up to 140 characters, known as tweets. It was created in March 2006 by Jack Dorsey and launched in July the same year. Twitter is primarily an interest based social network. Users either join Twitter to speak about things that they are interested in, or to listen to others who are talking on topics which they are interested in. Tweets are publicly visible to everyone on the web by default; however, senders can restrict message delivery to just their followers by making their accounts private. Users can tweet via the Twitter website, compatible external applications (such as Tweetdeck or Echofon), or by Short Message Service (SMS). Users may subscribe to other users tweets, which is known as following and the subscribers are known as followers. Following someone implies that all of his/her tweets will be visible on your personal Twitter homepage also known as home time-line.

Experts finding has become a hot topic along with flourishing of social networks such as micro-blogging services like twitter. Find experts in twitter is an important problem because tweets from experts are valuable sources that carry rich information (eg: trends) in various domains. Expert finding which aims at identifying people with the relevant expertise or experiences on a given topic query. The tweets in twitter cover extremely wide and diverse topics, such as routine activity or experiences, top news, technology etc. And users in twitter have rich expertise on various topics and finding these topic specific experts paves a way to enable others to follow the relevant and trust worthy information on a specific topic in micro-blogging services. Recently several attempts use the relations

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

among users and twitter lists for expert finding. These approaches only partially utilize such relations. Probabilistic method to jointly exploit three types of relations for finding experts, follower relation, user-list relation, list-list relation. Followers are people who receive the tweets and if someone follows you they will show up in your follower list.They will see your tweets in their home timeline whenever they log into Twitter.

A list is a curated group of twitter accounts. We can create our own lists or subscribe to lists created by others. Viewing a list timeline will show you a stream of tweets from only the accounts on that list. It will improving the accuracy of finding a experts on a given topic in twitter.

In proposed system, using an expert finding algorithm we will get the topic specific experts by finding the similarity of different relations. The dataset is crawled via Twitter API,for each user in twitter, crawled five types of data, user name, followers, tweets, user-list membership information, user- list subscribe information. This method aims to assign similar ranking scores to the similar users and lists, and meanwhile the ranking scores are subjected to the supervised information from the wisdom of twitter crowds. Based on the computed ranking scores, select the top relevant users for any given topic. Inaddition by using JGibbLDA topic modelling were done and from that find the experts and also using the TF-IDF text mining method to find the topic specific experts and compared both one and determine which one has better accuracy for finding the topic specific twitter experts. Now a days we get information about various domains from different social networks. One of the notable social network is twitter. Expert finding has become a hot topic along with the flourishing of social network such as twitter. The need of this study is to find topic specific expert in twitter. Finding experts in twitter is an important problem because tweets from experts are valuable sources that carry rich information in various domains. Expert finding which aims at identifying people with the relevant expertise and experience on a given topic. Finding these topic specific experts paves a way to enable others to follow the relevant and trust worthy information on a specific topic in micro-blogging services.

II. LITERATURE REVIEW

Expert finding methods have an assumption that individuals published documents are relevant with respect to their expertise on the knowledge on that particular topic. So they focus on modeling the associations between documents and candidate experts. This task has a great influence on the information retrieval community.

In [1], V. Qazvinian, E. Rosengren, D.-R. Radev, and Q.-Z. Mei detailed about Rumor has it: Identifying misinformation in microblogs, Rumor is a statement used to spread false information. It is a statement whose true value is unverifiable. The rumor detection in micro-blogs is based on features like: content based, network based and micro-blog specific features. The content based feature focus on the lexical patterns and parts of speech patterns of tweets. In lexical patterns all symbols and segments are represented as they appear and they are broken down into tokens with the help of space character .In the later patterns, all words are replaced with their part of speech tags. The network based feature focus on the user behaviour on Twitter. This helps to identify re-tweeted messages which may contain more content than the original tweet and may sound rumorous. The micro-blog specific features focus on hash tags and URLs. The hash tags use a hash () symbol prefixed over words or phrases. The hash tags used by rumours are different from that used by the tweets. Hence they can be identified easily. The users of Twitter use URLs to refer to external sources .If the URL is related to a rumour then it will provide a negative instance.

L.Chen et.al [2] proposed Expert finding for micro-blog misinformation identification which involves integration of collective and machine intelligence [2]. According to the micro-blog contents, the users are indexed automatically. Then a matching process occurs among users and suspected misinformation. To know the credibility of misinformation, it is sent to respective experts to judge their assessment. A tag based method is used to index experts of micro-blog users with social tags. There are two classes of misinformation: Domain Knowledge Constrained (DKC) and Time Space Constrained (TSC).The

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

former talks about domain specific topics while the later is related to some events that occur in some places and time. Let E denote all micro-blog candidate experts. Then priori probability of expert, e given a misinformation, m [Pr(e—m)]is directly proportional to priori probability of m given e [Pr(m—e)] and priori probability of e[Pr(e)].Here Pr(m) indicating misinformation remains same for all users and hence removed from ranking. Therefore only Pr (e) is taken into account. The process of increasing in size of social media it provides a convenient communication scheme for people, at the same time cradle of misinformation. Spreading the misinformation over social media is harmful to public interest. So they design a framework, which intelligence and machine intelligence, it helpful for identify misinformation. The basic point is (1) list the expert users according to their microblog satisfied. (2) It matches the experts with specified presence misinformation by sending the truth misinformation to suitable experts. They collect the analysis of expert and to decide the quality of information, and it helpful for prove the misinformation have propose a tag based method to list the experts of microblog users with the social tags. And match guess misinformation it is based on a real world dataset indicate.

In the paper [3], J. Weng, E.-P. Lim, J. Jiang, and Q. He detailed about twitter rank: Finding topic-sensitive influential Twitterers, it focus on identifying influential users of micro-blogging services. The Twitter which is a micro-blogging service uses a social networking model called following, making the users to choose whom they want to follow to receive tweets. It was found that 72.4% of the users of Twitter follow more than 80% of their followers and 80.5% of the users have 80% of users they are following follow them back. The Twitter rank algorithm is used to measure the influence of users in twitter .This is an extension of Page rank algorithm which only measures the influence based on the link structure of network. The Twitter rank algorithm measures the influence based on link structure and topical similarity between users. It is better than the Page rank algorithm. The method consists of three processes: Topic distillation, Relationship construction and Ranking. The topic distillation automatically identify topics that twitterers are interested based on the tweets published by them. Then a network is constructed based on the topic specific relation between users and their followers. Finally a topic sensitive user influence ranking process occurs which gives us relevant list of users who is influenced on a particular topic in twitter. Twitter Rank works in two steps; first one is employs Latent Dirichlet Allocation (LDA) model. It notices the topics of independent based on their tweets. Second one for each topic it builds a graph weighted by taking both the topical similarity in between two users and follower graph, then also enrol page rank algorithm for find topic specific influential users.

In the paper [5],S. Ghosh detailed about, 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. One disadvantage with list based methodology is list spamming where malicious users create fake lists.

N. k.sharma has focused on the paper [7], inferring who is who in the twitter social network twitter list: they propose to use twitter list to identify the quality of twitter users by twitter crowed. List contain the users and to compute similarity between each user and given topic query. This is used to search and rank all the users. Using congnos move to choose the users that users contained in more number of lists those Meta data contain the query. It use twitter list to identify the quality of twitter users. In this paper,they design and evaluate a novel who-is-who

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

service for inferring attributes that characterize individual Twitter users. This methodology exploits the Lists feature, which allows a user to group other users who tend to tweet on a topic that is of interest to her, and follow their collective tweets. Our key insight is that the List meta-data (names and descriptions) provides valuable semantic cues about who the users included in the Lists are, including their topics of expertise and how they are perceived by the public. Thus, we can infer a users expertise by analyzing the meta-data of crowdsourced Lists that contain the user. The methodology can accurately and comprehensively infer attributes of millions of Twitter users, including a vast ma- jority of Twitters influential users (based on ranking metrics like number of followers). This work provides a foundation for building better search and recommendation services on Twitter.

In research paper [8], propose Ahmad Kardan a novel method to find experts who are members of social network by means of business intelligence approach. This model is first verified by real data from Friend Feed social network. First data is extracted, transformed and loaded into data warehouse with ETL processes. A new ranking algorithm has been proposed for ranking experts. This algorithm has been proposed for finding importance of people in a social network. Some changes were made in PageRank algorithm to make it possible to use in social network for expert finding. In PageRank algorithm, different pages were navigated and importance of each page is calculated using Marcov chain. In the proposed algorithm ,instead of webpages ,there are people in social network and connection between them are used are hyperlinks.

In paper [9], Alessandro Bozzon, focus on finding expert within the population of social networks, according to the information about the social activities of their users. This paper considers social network both as a source of expertise information and as a source of expertise information and as a route to reach expert users and define models and methods for evaluating people expertise by considering their profiles and by tracing their activities in social networks. For matching queries to social resources, it uses both text analysis and semantic annotation. In this paper, considers the problem of ranking the members of a social network according to the level of knowledge that they have about a given topic after such ranking, top-k experts are chosen. First step for this is analysis of resources is performed. This is done by extracting social data from different platforms through their APIs. For each considered resource, Resource Extraction module perform the analysis flow. Another main factor to consider is Language Identification that allows classification of resources according to their main language. Next step is to ensure matching expertise need to candidate need. For this, it uses a vector space model where resources, related entities and expertise needs are represented in same space. The score of each resource is calculated. Ranking is performed according to these score.

III. PROPOSED METHODOLOGY

Expert finding has become a hot topic along with the flourishing of social networks, such as micro-blogging services like Twitter. Finding experts in Twitter is an important problem because tweets from experts are valuable sources that carry rich information (e.g., trends) in various domains. Recently, several attempts use the relations among users and Twitter Lists for expert finding. Nevertheless, these approaches only partially utilize such relations. In this paper, develop a probabilistic method to jointly exploit three types of relations (i.e., follower relation, user-list relation, and list-list relation) for finding experts. Specifically, propose a Semi-Supervised Graph-based Ranking approach (SSGR) to offline calculate the global authority of users. In SSGR, employ a normalized Laplacian regularization term to jointly explore the three relations, which is subject to the supervised information derived from Twitter crowds. We then online compute the local relevance between users and the given query. By leveraging the global authority and local relevance of users, rank all of users and find top-N users with highest ranking scores.

$$W\ i;j = T\ F\ i;j \quad \log(N=DF\ i)$$

IV. METHODOLOGY

A. LDA and TF-IDF

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Topic modelling algorithms are used to discover a set of hidden topics from collections of documents, where a topic is represented as a distribution over words. Topic models provide an interpretable low-dimensional representation of documents (i.e. with a limited and manageable number of topics). LDA is a typical statistical topic modelling technique and the most common topic modelling tool currently in use. It can discover the hidden topics in collections of documents using the words that appear in the documents. Pattern-based representations are considered more meaningful and more accurate to represent topics than wordbased representations. Moreover, pattern-based representations contain structural information which can reveal the association between words. In order to discover semantically meaningful patterns to represent topics and documents, two steps are proposed: firstly, construct a new transactional dataset from the LDA model results of the document collection D; secondly, generate pattern-based representations from the transactional dataset to represent user needs of the collection D. Using JGibbLDA topic modelling were done and then find the topic experts.

Users in Twitter have rich expertise on various topics and finding these topic-specific experts paves a way to enable others to retrieve or follow the relevant and trustworthy information on a specific topic in micro-blogging services. For example, if somebody is tweeting about the movie review, there will be lots of tweets and we have to find out who is the expert in giving reviews. Also there are many algorithms for finding the expert review. An algorithm used for this is tf-idf (term frequencyinverse document frequency). In information retrieval, tfidf, short for term frequencyinverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval, text mining, and user modeling.

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. Nowadays, 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 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 determine the experts in particular domain. Then compared both the JGibbLDA and TF-IDF methods and evaluate which one have better accuracy to find the topic specific experts. This project aims to find the top-N relevant users for any given topic, then compare and determine the accuracy of different methods that are used for finding topic specific experts.

B. Methods of Data Collection and analysis:

Users: The data set used in this paper was crawled via Twitter API. Real time data is taken as dataset. For each user in Twitter, we crawled five types of data, i.e., user profiles, followers, tweets, user-list membership information, and user-list subscribe information. In particular, we used a user-centric strategy to collect data as a brute-force crawling of all users for all lists would be prohibitively expensive and would not scale6 . More specifically, to be unbiased to the users, we randomly crawled the information of users by utilizing a publicly available user collection as the seed set. From dataset, filtered non-English characters, stopwords, punctuation as well as the high-frequency words in Twitter (e.g., RT), and employed Porters stemmer [36] for remaining words. After the processing, the users without any context information are removed.

Queries: In this project sample queries are used for evaluation, whose topics are from general to specific, e.g., a general per-sonal hobby like traveling or a specific TopNews like Boston Marathon bombings, which can be used to comprehensively evaluate the effectiveness of our proposed approach. A semi-supervised graph-based ranking method used for com-puting the global authority of a user on the given topic. The regularization term in the semi-supervised graph-based ranking method used to smooth the ranking scores on the graph. By using the approach of the

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

semi-supervised graph-based ranking method improves the effectiveness of finding the topic-specific experts. Ranking method can effectively exploit the three different types of relations among users and lists (i.e. follower relation, user-list relation, list-list relation).

The input data or real-time dataset were get from the twitter API. By creating Consumer Key (API Key), Consumer Secret (API Secret), Access Token,Access Token Secret, get the real time data from the twitter accounts. After getting the data, the data stored in the database called SQLYog. Then using expert find algorithm, find the similarity and weight of different relations and find the expert.Using JGibbLDA topic modelling were done using Maximum matched Pattern-based Topic Model (MPBTM)and then find the topic experts. Then by using TF-IDF method determine the experts in particular

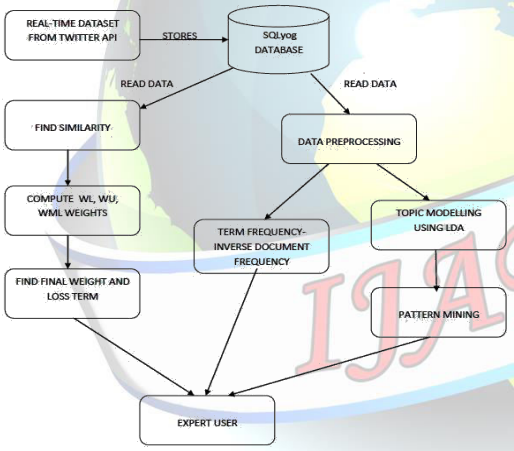

Fig. 1. Architecture of Proposed System

domain. Then compared both the JGibbLDA and TF-IDF methods and evaluate which one have better accuracy to find the topic specific experts. This project aims to find the top-N relevant users for any given Query or Topic.

V. CONCLUSION

In this paper, address the problem of topic-specific expert finding in Twitter and successfully integrate different types of user-related information (i.e., the crowdsourced Lists informa-tion, follower graph and users profiles) into a unified ranking framework for accurately inferring the topical expertise of users. To the best of knowledge, this is the first attempt that targets expert finding problem in Twitter by utilizing all of such information. Specifically, within the framework and develop a semi-supervised graph-based ranking method, comprising a loss term. Using JGibbLDA topic modelling were done and then find the topic experts. Then by using TF-IDF method determine the experts in particular domain. Then compared both the JGibbLDA and TF-IDF methods and evaluate which one have better accuracy to find the topic spe-cific experts. This project aims to find the top-N relevant users for any given topic, then compare and determine the accuracy of different methods that are used for finding topic specific experts. The experiments conducted on real-world Twitter data set demonstrate that our method significantly outperforms the state-of-the art methods. The following potential directions: First, we would like to improve the efficiency of the learning of the global authority of users for expert search; Second, we also interest in studying the diversity issue in the expert finding problem.

ACKNOWLEDGMENT

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

REFERENCES

[1] V. Qazvinian, E. Rosengren, D.-R. Radev, and Q.-Z. Mei, Rumor has it: Identifying misinformation in microblogs, 2011.
[2] L. Chen, Z.-Y. Liu, and M.-S. Sun, Expert finding for microblog misin-formation identification, 2012.
[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, Twitterrank: Finding topic-sensitive influential Twitterers, 2010.
[4] A. Pal and S. Counts, Identifying topical authorities in microblogs, 2011.
[5] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, Cognos: Crowdsourcing search for topic experts in microblogs, 2012.
[6] X. Liu, S. Zhang, F. Wei, and M. Zhou, Recognizing named entities in tweets, 2011.
[7] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi, Inferring who-is-who in the Twitter social network, 2012.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

[8] Ahmad Kardan , Amin Omidvar, Farzad Farahmandnia, Expert Finding on Social Network with Link Analysis Approach.
[9] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, Giuliano Vesci, Choosing the Right Crowd: Expert Finding in Social Networks.
[10] Jing Zhang, Jie Tang, and Juanzi Li, Expert Finding in A Social Network, 2007.
[11] Christopher S. Campbell Paul P. Maglio Alex Cozzi Byron Dom,Expertise Identification using Email Communications, 2003.
[12] Kevin R.Canini ,Bongwon Suh, Peter Pirolli, Finding Relevant Sources in Twitter Based on Content and Social Structure.