EXPLORING TOPICAL EXPERTS IN TWITTER BY ELIMINATING SPAMMERS

Kishnaveni S R
M. Tech Scholar,
Department of Information Technology,
Government Engineering College, Barton Hill

Simi Krishna K R
Assistant Professor,
Department of Information Technology,
Government Engineering College, Barton Hill

Abstract— Social Networks are presently been utilized immensely everywhere throughout the world. They are dynamically being used as an information source. It has turned out to be exceptionally hard to recognize which among these data are rich ones. Thus, Expert finding has turned into a hotly debated issue with the flourishment of social networks. Expert finding task is aimed at recognizing persons with applicable ability or experience on a given topic. This paper focus on the issue of distinguishing topic experts in micro-blogging services like Twitter. Previous strategies can't be straightforwardly used to find out about topic specialists in Twitter. Some of the new methods use various relations between users and Twitter Lists for inferring user's expertise on a given topic. Such methodologies just focus on fractional relations and don't endeavour to isolate spammers from the set of genuine experts. Here, we find topical experts by considering query relevance, exploring particular relations among users and lists and discards spammers.

Keywords: Twitter, Expert finding, Spammer detection.

INTRODUCTION

Twitter is an online news and social networking service that enables users to send and get short messages called tweets (confined to 140 characters), empowering individuals to share and find topic of interest in real time. Registered users can post tweets, yet the individuals who are unregistered can just read them. Users get to twitter through its site interface, SMS or a cell phone application. Twitter utilizes a component called "following" using which any user can choose who he/she wants to follow and this requires no permission from the other.

Twitter can be considered as a repository of information in various domains. In order to gather information from twitter, a user must follow another user's tweet. At this point it is very important to identify who to follow. Users in twitter may have rich expertise on different topics. By identifying such experts others users can follow them in order to gather information on a particular topic of interest. If we are following the appropriate person, the gathered information will be trustworthy. Recognizing topical expert is the initial move towards finding legitimate data on a given topic. For many applications like opinion mining and Name Entity Recognition (NER), identifying topical experts is a pre-processing step. For instance, assessments drawn from beautician's tweets would support a cosmetic maker than those from regular users.

Conventional systems rely upon the assumption that tweets posted by the twitterers are related to their knowledge on a particular topic. However this won't not be valid, as users published tweets won't not be specifically identified with their aptitude. Since users may retweet a considerable measure of tweets contain the subject words. Therefore, the problem of expert finding in twitter is a tedious task. There exist several attempts for the Twitter expert finding problem. But most of them only make use of partial relations as well as user's bio and their published tweets and in view of this they are deducing an expertise score for each user. They don't consider the possibility of spammers while finding the candidate experts. There are several separate works for the spammer identification. To this end, this paper proposes a strategy that isolates the spammers and spontaneous bloggers from the specialists of a particular domain. For finding genuine experts we utilize all relations that exist among users and lists in twitter.

The relations that we are considering here are Follower Relation, User List Relation and List-List Relation. Follower relationship is a solid marker of likeness among users. In other words, a twitterer follows a friend since

he/she is keen in the topics the friend publishes in tweets, and the friend follows back because he/she finds they share similar topic interest. Utilizing Lists feature included in Twitter any user can group accounts that tweet on a topic that is important to her, and follow their aggregate tweets. Similarity between lists is explored to find the relevant list for a particular topic. For detecting spammers several user based and content based features are selected. The detailed description of the proposed method is presented in the following sections.

LITERATURE SURVEY

The paper, Twitter rank: Finding topic-sensitive influential twitterers [2], aims at finding influential users in micro blogging services. It introduces the phenomenon of homophily in twitter. A twitterer follows a friend when he/she is interested in the topics published by the friend through his/her tweets and the friend follows backs because he/she finds they have similar interest. This phenomenon is called homophily. Based on this phenomenon; an algorithm called twitter rank is proposed which is utilized to gauge the impact of users in the twitter community by considering the connection structure and topical similitude between the users. The method uses three major steps. Topic distillation, Relationship construction and Ranking. Topic distillation identifies the interested topics of users based on their published tweets. Then based on the topic specific relation between the users and their followers a network is constructed. Finally a ranking approach is used to retrieve a relevant list of users who is influenced on a particular topic.

The paper [3], proposes a method to identify and to rank users based on their relevancy to a given topic. A ranked list of relevant users for a given topic is created by combining a basic text search and an analysis of the social network structure. The procedure in the algorithm discovers a set of candidates who are relevant to the given topic. Using the Twitter API, a standard twitter search is executed and the output is a set of users who are associated with the particular topic. This user set is called voters. In the next step opinions of the voters are measured by observing who they follow. Considering this, a candidate users set is formed. It includes anyone who is followed by at least one of the voters. Then, retrieve two numbers fu and Fu, for each user u in the candidate set. fu denotes the number of voters who follow user u and Fu denotes the total number of twitter users who follow user u. With these two numbers, relevancy score for each candidate member is calculated. Then based on the score, ranking is performed.

The paper [4] uses propagation based approach that utilizes both local information and network information. First, Initialization step is performed which is followed by Propagation step. In the first step, it uses person local information to calculate an initial expert score for each person. Then top ranked persons are selected as candidates. A sub-graph is constructed based on the selected candidates. Second step, propagates one's expert score to the persons with whom he/she has relationships. The basic idea in the initialization stage is that if a man has wrote numerous reports on a theme or if the individual's name co-happens in ordinarily with the theme, at that point it is likely that he/she is a candidate expert on the subject. In Propagation step, the precision of expert finding is enhanced by making utilization of connections between people. The essential thought here is that if a man knows numerous specialists on a subject or if the individual's name co-happens in numerous times with other experts, at that point it is likely that he/she is a specialist on the topic.

The paper, Cognos: Crowd sourcing search for topic experts in micro-blogs [5] present Cognos, a system for finding topical experts in Twitter. Cognos exploits the Lists feature in Twitter. Using this List, any user can group Twitter accounts that tweet on a topic that is of interest to his/her, and follow their collective tweets. The twitter List gives information like list name, description and members. The key idea is to analyse the metadata of the Lists containing a user to infer the user's expertise score. Aside from the above research studies, there likewise exist a few services for recognizing topical specialists in Twitter. Recognizing the significance of hunting down specialists on specific themes, Twitter itself gives an official "who to follow" (WTF) benefit [7] where one can scan for specialists on a given theme. It is found[8] that Twitter WTF utilizes a few factors, for example, the profile data of users, their social connections, their level of engagement in Twitter to recognize topical specialists. All the current works just depend on discovering specialists on a given query, do not consider the instance of spammers. In this paper we extend the expert finding task by eliminating unwanted spammers and bloggers.

METHODOLOGY

The principle point of our approach is to discover top N pertinent specialists on a given query Q. The given query Q may contains a few terms which can be spoken to as $Q= \{t_1, t_2\ldots \ldots t_{|Q|}\}$, where |Q| is the quantity of terms. When we are running an inquiry, we need to discover applicable outcomes not just for the correct articulation we wrote on the pursuit bar, yet in addition for the other conceivable types of the words we utilized. For this, we apply some common language processing methods such as case folding, stemming, removal of stop words and tokenization. At that point we will get the topical words contained in the inquiry. Presently the assignment is to assess the likelihood of the query on a given user. That is, process the similitude between users distributed tweets and the query term. User query similarity of a user is calculated as the number of times the query terms used by the user divided by the total occurrences of the query word. If a user publishes more tweets about a particular topic word, that users will gangs the high closeness score (similarity score).

[1]The next step is to exploit the different relations among users and lists to assign an expertise score for users. For this let's first consider user-follower relations. If one user follow another in twitter, it means that the first user is interested/values the information published by the second. So user-follower relationship is very helpful in finding users topical expertise score. It has two effects. First, if a user is followed by a significantly expert user on a given theme, this user is more probable a specialist on that theme. Second, the more followers of a user are expert on a theme, the more probable that the user is a specialist on that theme. In order to calculate user-user (user-follower) similarity between two users u_i and u_j we use cosine similarity equation.

Next is user-list relation. We can investigate two sorts of user-list relations. A List is a curated assembling of Twitter accounts. We can make our own list or subscribe in to list made by others. The first is called Member Relation and the latter is called Subscriber Relation. From the Member Relation we can find that if a user is incorporate into numerous number of list identified with a given query, the user is probably going to be a specialist on that query term. On account of Subscriber relation, if the subscribers of a list are expert on a given topic, the more likely the subscribed list is significant to the topic. Here also we use cosine similarity to calculate similarity between a user and a list. While calculating cosine similarity between a user and a particular list, the number and similarity estimations of users incorporated in to that list is considered. In other words, the similarity value of a list is the sum of similarity values of users in that list and this value is used to find cosine similarity between a given user and the list. Finally comes List-List relation. If a list has high similarity to another list that is highly relevant to a topic, then the first list is said to be more relevant to the specified topic.

DEVELOPING EVALUATION METRICES

For further calculations we have to develop some metrics based on the above calculated values, which are[1]:

W_u : A n x n symmetric user-user/user-follower matrix. Each term in the matrix represents similarity between two users for a given topic. (Both users should follow each other).Here n is the number of users.

W_l : A m x m symmetric list-list matrix. Each term in the matrix represents similarity between two lists for a given topic. Here m is the number of lists.

W_{ml} : A n x m user-list matrix. Each term in the matrix denotes the similarity between a user and a list for a given topic.

We need to derive a regularization factor which gives similar positioning scores for similar users and similar lists by considering the previously mentioned three sorts of similarities. In order to calculate this term the following terms needs to be calculated [1].

M_u : n x n diagonal matrix
M^u_{ml} : n x n diagonal matrix
M_l : m x m diagonal matrix
M^l_{ml} : m x m diagonal matrix

The $(i,i)^{th}$ element of M_u is the sum of i^{th} row of W_u
The $(i,i)^{th}$ element of M_l is the sum of i^{th} row of W_l
The $(i,i)^{th}$ element of M^u_{ml} is the sum of i^{th} row of W_{ml}
The $(i,i)^{th}$ element of M^l_{ml} is the sum of i^{th} column of W_{ml}

$$Term_1 = \alpha_1 \sum_{i,j=1}^{n} w_{i,j}^{u} \left(\frac{f_i}{\sqrt{[M_u]_{ii}}} - \frac{f_j}{\sqrt{[M_u]_{jj}}} \right)^2$$

$$Term_2 = \alpha_2 \sum_{i,j=1}^{m} w_{i,j}^{l} \left(\frac{g_i}{\sqrt{[M_l]_{ii}}} - \frac{g_j}{\sqrt{[M_l]_{jj}}} \right)^2$$

$$\text{Term}_3 = \alpha_3 \sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}^{ml}\left(\frac{f_i}{\sqrt{[M_{ml}^{u}]_{ii}}} - \frac{g_j}{\sqrt{[M_{ml}^{l}]_{jj}}}\right)^2$$

Where $w_{i,j}^{u}$ is the similarity between user u_i and follower u_j, $w_{i,j}^{l}$similarity between two lists L_i an L_j and $w_{i,j}^{ml}$ similarity between a user u_i and his included list.f_i denotes the relevancy of a user u_i on given topic and g_i denotes the relevancy of a list L_i up on a given topic query. In addition, the fusing weight $\alpha_i \geq 0$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Now the regularization factor can be calculated as the sum of the three terms.Term_1 ensures that a user will get high score if the follower of that user is an expert on a given query. Also, if a user has many followers expert on a given topic, then the user will be assigned a high score.Term_2 has the same purpose for lists as Term_1.Term_3 mutually ranks users and lists. Finally the regularization factor is normalized.

[1]To encode Member Relation and Subscriber Relation we develop two matrices C and D respectively. Each element of C (c_{ij}) is computed as the reciprocal of number of lists containing a particular user. Similarly, each element of D (d_{ij}) is the reciprocal of number of subscribers for a particular list.

$$\text{Lterm}_1 = \sum_{i=1}^{n}(f_i - \sum_{j=1}^{m} c_{ij}\, g_j)^2$$

$$\text{Lterm}_2 = (1-\beta)\sum_{i=1}^{m}(g_i - \sum_{j=1}^{n} d_{ij}\, f_j)^2$$

$$\text{Loss term} = \text{Lterm}_1 + \text{Lterm}_2$$

β is in the range ($0 \leq \beta \leq 1$) which is used to trade off the two terms in loss term. Then we normalize the loss term.

In order to rank users based on their expertise level on a given topic, we need to develop a ranking framework. So based on the relations, the framework is the sum of regularization term and loss term. But we have to consider the direct relevance of query term on each user also. For this we have to add user query similarity value to the above sum.So the final score for each user will be

Last Score(u_i)=Query relevance value+ μ.regularization term+(1-μ)loss term. μ is used to trade off the contributions of regularization and loss term. We make the ranking of users based on this last score value.

CHALLENGES INVOLVED

The procedure for expert finding depicted above as per[1] expect that all users are genuine. By considering only the above factors we can't ensure that the last yield will be an arrangement of expert users. There is additionally a risk for Spammer users. Spammer user can deliberately make counterfeit profiles and can be a member of a list or can be a follower of a few specialists. Before performing the above calculations it is very needed to filter out those users who are not genuine. Thereby we can reduce the processing overhead and time. So the first procedure in our work is to perform spammer detection and the output of this module serves as the input for the above calculations.

ELIMINATING SPAMMERS

Attributes used for detecting Spammers are:

User based features which include features like number of followers, number of followings, followers/following ratio, age of account etc.

Content based features which include number of hash tags, number of URLs in tweets, number of mentions, retweets, etc.

Role of above mentioned features for spammer detection according to Twitter policy [9]:

- Numbers of followers-spammers have less number of followers.
- Numbers of followings-Spammers tend to follow a large number of users.
- Followers/Following Ratio- this proportion is under 1 for spammers
- Age of account- Spammers have generally new accounts so this feature has less value for spammers.
- No. of hashtages(#)- spammers tweet multiple unrelated updates to the most mentioned topics on Twitter using # to lure legitimate users to read their tweets.
- No. of URLs- spammer's tweets consist of large number of URLs of malicious sites.
- @mentions- spammers use maximum @usernames of unknown users in their tweets so as to avoid being detected.
- Retweets- Retweets are the replies to any tweet using @RT symbol and spammers use maximum @RT in their tweets.

In view of the above recognized highlights, we continue to utilize conventional classifiers, for example, Naives Bayesian and SVM to distinguish spammers. At that

point the gathered information (CSV organize) is trained utilizing these characterization calculations. At that point its precision is checked utilizing test information in CSV format. Now we can undoubtedly recognize spammers and non spammers in our entire user set. Our point is to recover the set of non spammers, which is required in our further calculations (calculations performed in evaluation matrices).

CONCLUSION AND FUTURE WORK

As Twitter rises as a well known platform for discovering genuine data on the Web, an essential research challenge lies in distinguishing specialists in particular topics. Conventional techniques are observed to be not successful for finding topical specialists and do not consider the spammer problem. In our work we are separating the genuine users by taking out spammers. In our expert computation, we just consider the non-spammer users and we can ensure that our final sorted list of users contain only genuine topical experts. The future extent of this paper is to utilize semantic relatedness between query terms and tweets so we can adequately compute the relevancy of the query term with the user.

REFERENCES

[1] Wei Wei, Gao Cong, Chunyan Miao, Feida Zhu, and Guohui Li, "Learning to Find Topic Experts inTwitter via Different Relations"

[2]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. 261–270.

[3]Kevin R.Canini ,Bongwon Suh, Peter Pirolli, "Finding Relevant Sources in Twitter Based on Content and Social Structure".

[4] Jing Zhang, Jie Tang, and Juanzi Li," Expert Finding in a Social Network".

[5]. S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K.Gummadi, "Cognos: Crowdsourcing search for topic experts In micro-blogs," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2012, pp. 575–590.

[6] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, ATC'11, Banff, Canada, Sept 2-4, 2011, IEEE.

[7] Twitter: Who to Follow. http://twitter.com/#!/who_to_follow.

[8] Twitter Improves "Who To Follow" Results & Gains Advanced Search Page. http://selnd.com/wtfdesc.

[9] http://help.twitter.com/forums/26257/entries/1831-The Twitter Rules

[10] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, Detecting Spammers on Twitter, CEAS 2010 Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference, July 2010, Washington, US.