Topical Authority Propagation on Microblogs

Juan Hu
Computer Engineering
Department
Santa Clara University
Santa Clara, CA 95053, USA
jhu1@scu.edu

Yi Fang
Computer Engineering
Department
Santa Clara University
Santa Clara, CA 95053, USA
yfang@scu.edu

Archana Godavarthy
Computer Engineering
Department
Santa Clara University
Santa Clara, CA 95053, USA
agodavarthy@gmail.com

ABSTRACT

With a huge number of active users on microblogs, it becomes increasingly important to identify authoritative users on specific topics. This paper tackles the task of finding authorities on Twitter given any query topic. Although there exists much work on identifying influential users on Twitter, most of them focus on global authority regardless of the topic. We propose a novel Topical Authority Propagation (TAP) model by utilizing the fact that topical authority can be propagated through retweeting, i.e., if a user's tweet on a given topic is retweeted by a topical authority, that user is likely to be an authority on the topic as well. Topical relevance of candidate authorities can be seamlessly integrated into the model. Link analysis algorithms such as PageRank can then be utilized to characterize how topical authority is propagated through retweeting. We conduct a set of experiments on Twitter and demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation

Keywords

Topical authority, Twitter, Link analysis

1. INTRODUCTION

Microblogs such as Twitter have became a very important platform to publish and obtain real-time information. Compared with traditional medias such as newspapers and TV, news spreads very fast on Twitter primarily via tweeting and retweeting. In Twitter, a user A may follow another user B to get information from B's public tweets. If A find B's tweet interesting, he or she may retweet it which can then be seen by all of user A's followers. People often want to obtain relevant and reliable information on specific topics by following authoritative users or so called topical authorities. It is very challenging to identify such authorities on Twitter given a specific topic, due to the huge number of users, incomplete user profiles, and the limited length of a tweet. Furthermore, there is a great variety of Twitter users ranging from popular celebrities to experts in certain fields such as scientists and well-known tennis players, to ordinary users who are passionate about certain topics and constantly producing useful real-time information on that specific topic, and also many other users who just post messages about their personal life. This notable diversity results in the great difficulty in finding topical authorities.

This paper tackles the task of finding authorities on Twitter given any query topic. Although there exists much work on identifying influential users on Twitter, most of them focus on global authority regardless of specific topics. We propose a novel Topical Authority Propagation (TAP) model by utilizing the fact that topical authority can be propagated through retweeting, i.e., if a user's tweet on a given topic is retweeted by a topical authority, that user is likely to be an authority on the topic as well. Unlike the prior work that builds a follower-followee graph, we build the TAP graph based on the retweeting activities where each link is from the user who retweeted the tweet to the user who posted the tweet. Topical relevance of the candidate authorities can then be seamlessly integrated into the model. Specifically, the weight of the link is determined by the content relevance of the tweet text with respect to the specific topic. A related tweet retweeted by a topical authority is a stronger indicator of topical authority of the tweet author compared to an irrelevant tweet retweeted by the topical authority. Link analysis algorithms such as PageRank [2] can then be utilized to characterize how topical authority is propagated through retweeting. TAP integrates user authority and topical relevance into a single unified model by utilizing one of the most popular user activities on microblogs: retweeting. Moreover, TAP can be easily extended to consider other microblogging features such as mention, reply, and user profile. We conduct a set of preliminary experiments on Twitter and demonstrate the effectiveness of the proposed approach.

2. RELATED WORK

While much efforts have been made to quantify user's overall influence on Twitter [1, 7, 3, 5], very few of them measures user authority given a specific topic. KDD Cup

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2507891.

launched a track in 2012 to predict which users one user might follow in Tencent Weibo, a Chinese microbloging site. This is essentially a recommendation task and did not target on finding authoritative users for a particular topic.

To the best of our knowledge, there exists only three works on identifying topical experts on microblogs [6, 8, 4]. Pal et. al. [6] proposed probabilistic clustering over a set of 15 features extracted from Twitter graph and retrieved a list of top authors by a Gaussian Ranking algorithm. The TwitterRank proposed in [8] constructs the user graph based on *following* behavior where edge weight is decided by topical similarity between two users. While both approaches involve network-based calculations like PageRank, our proposed approach differs in the following important ways. First, the link structure in our social graph is based on the retweeting activities instead of following. While the number of retweets is utilized in some prior work, none of them constructed the link graph based on retweeting. In fact, the retweeting activities are the main reason that information spreads so fast on Twitter. If a topical authority retweete a user's tweet, this user is likely to be an authority for this topic as well. Secondly, our proposed approach considers the topical relevance of the retweeted content. If a tweet is not very related to the topic, it should not be a good indicator of topical authority even though it might be retweeted many times.

Given the importance of identifying topical authorities, Twitter has officially launched a service called *Who To Follow* for users to search a topical expert by utilizing signals from user profiles and their followers. However, the performance of this service is not satisfactory according to the study by [4]. The Cognos [4] addressed the problem of identifying topical experts by mining the Twitter List information, where one user's expertise on a specific topic is inferred by annotation from other users' lists. However, only relying on the list information could be risky. Lists on Twitter are much less popular than retweeting. This is probably due to the nature of Twitter, which is not a professional social network site. Users on Twitter are reluctant to build lists and provide detailed descriptions under each list. In addition, some list information may be private and thus the information is unavailable for Cognos. In fact, some other popular microblogging websites such as Sina Weibo currently do not provide the list feature. In contrast, our proposed approach only relies on retweets which are pervasive on all the major microblogging sites.

3. TOPICAL AUTHORITY PROPAGATION VIA RETWEETING

On microblogging websites such as Twitter, a user obtains information not only directly from the tweets posted by people they follow but also from retweets. Retweet is represented by RT together with "@" followed by user screen name which describes the source of this retweet. If a tweet is found to be interesting or valuable, the user may want to share it with their friends through retweeting. Information conveyed through retweets is often reliable and useful. It is interesting to note how retweets can reach users who might be far beyond the social network of the original user. Specifically, retweet starts from the friends of the original author of the tweet and then it is evaluated and consumed by these friends. If the tweet is informative, these friends retweet the tweet immediately so the tweet reaches friends of friends.

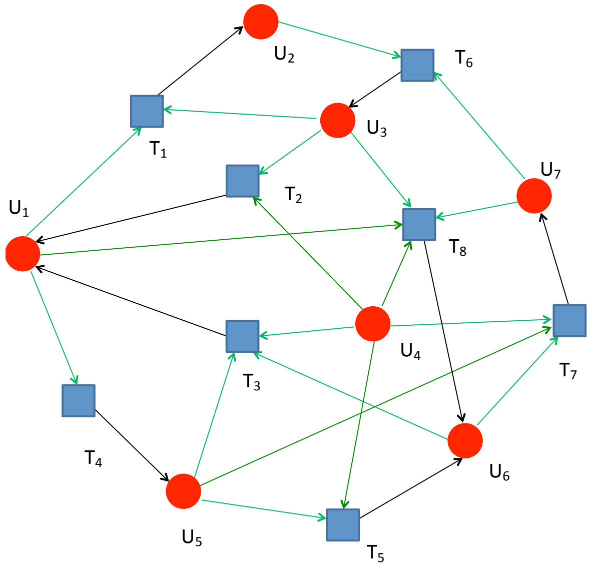

Figure 1: An example of the Topical Authority Propagation (TAP) graph. U denotes user node (red circle) and T denote tweet node (blue rectangle)

As a tweet is short and easy to understand, the review is quickly done and as a result, informative retweets spread rapidly over the Twitter network. The power of Twitter as a social media largely relies on the function of retweeting.

In this paper, we exploit retweeting activities to rank topical authorities on Twitter. Being retweeted by multiple people indicates the quality of the tweet. If one tweet is related to a specific topic and this tweet has been retweeted many times, the author of the tweet can be recognized as an authoritative source for this specific topic since his or her tweet has been examined by the collective wisdom of the crowd. Furthermore, if the tweet is retweeted by many authoritative users on this topic, it reveals that the author is very likely an authority on the topic as well. Since users can be linked to each other through the retweeting relation, the topical authority can then be propagated through the retweeting activities. It is important to note that the topical relevance of the tweet should be crucial in identifying the topical authority.

We formalize the retweeting activities in the Topical Authority Propagation (TAP) graph. Figure 1 shows a sample of the TAP graph. There are two types of nodes in the graph: the user node and the tweet node, denoted by red circle and blue rectangle respectively. Each user node represents one particular user and each tweet node represents one particular tweet. In addition, there are two types of links in the graph: 1) the link (in black) from tweet node to user node denotes the ownership of the tweet, i.e., the tweet is posted by this user; 2) the link (in green) from user node (retweeter) to tweet node represents the retweeting relation, i.e., the user retweets the tweet.

Each link in the TAP graph has a weight which reflects how much authority is propagated from source node to destination node. The weights for the two types of links should be computed. In order to encode topical relevance, the weight of the link from tweet node to user node is determined by the topical relevance score of the tweet with respect to the query topic. For example, in Figure 1, user U_1 has posted two tweets T_1 and T_2. If T_1 and T_2 have relevance scores of 10 and 5 respectively, the weight from T_1 to U_1 would be twice as the weight of the link from T_2 to U_1. The topical authority will be propagated to the author through the link in proportional to the weight. For the link from user node to tweet node, the weight is one divided by the total number of tweets the user has retweeted. The intuition is similar to that in the PageRank algorithm: if a user retweets too many tweets, the support from this user should be discounted.

Based on the link structure demonstrated in Figure 1, we can use link analysis algorithms such as PageRank or HITS to compute the authority scores of all the nodes. In fact, we notice that the link graph derived from retweeting activities is a special one. Because each tweet has a unique user, the link from a particular tweet node always points to a unique user node. Consequently, we can combine the two links from the retweeter to the tweet and the tweet to the author into one single link which is from the retweeter to the author. The weight of this single link is determined by the content relevance score of the tweet with respect to the query topic, which can be computed by any text retrieval model such as vector space model and language modeling. The weight is then normalized by the total relevance scores of all the outlinks from the retweeter. This normalization discounts the links coming from the retweeters who retweets too many tweets. The intuition of the proposed approach can be summarized as: a user is authoritative on a given topic if many authoritative users of the topic have exclusively retweeted this user's many tweets on the topic.

Formally, we can define the topical authority $TA(i)$ of user i as:

$$TA(i) = \sum_{j \in ret(i)} w_{ji} \times TA(j) \tag{1}$$

where $ret(i)$ is the set of user nodes that retweet any tweet of user i. w_{ji} is the edge weight of the link from user j to user i, which is computed by

$$w_{ji} = \frac{s_{ji}}{\sum_{k=1}^{n_j} s_{jk}}$$

where s_{ji} is the topical relevance score of the tweet of user i that user j has retweeted, which can be computed by any text retrieval model. n_j is the total number of outlinks from j. As we can see, w_{ji} is normalized and discounted by the total topical relevance score from j. Based on Eqn. (1), we can adapt the PageRank algorithm to compute the topical authority for all the user nodes. The topical authority is initialized as $\frac{1}{N}$ where N is the total number of users. Eqn. (1) is then repeatedly applied until $TA(i)$ converges for all i.

In Twitter, some users may never retweet. These users will form dead-ends in the TAP graph because there is no outlink from them. If we apply Eqn.(1) to such a graph, these nodes will absorb all the topical authority. Similar to PageRank [2], we can introduce the teleport operation as shown in Eqn. (2) so that we are never stuck at the dead-ends.

$$TA(i) = \alpha \sum_{j \in ret(i)} w_{ji} \times TA(j) + \frac{1-\alpha}{N} \tag{2}$$

where α is the damping factor which is generally set around 0.85 [2].

4. EXPERIMENTS

4.1 Experimental Setup

To evaluate the proposed approach, we built a testbed by crawling public tweets from Twitter Streaming API [1] over the period between May 16th and May 22nd, 2013. In the seven days, we collected 30,429,000 tweets in total. Since our proposed method is based on retweets, we then formed the retweet corpus. In the retweet object from Twitter API, the retweet structure consisted of not only the retweet information such as retweeter id, screen name, and retweet text, but also the full set of the original tweet information in the field of *retweeted_status*. We utilized this information to extract the retweets by pattern matching the keyword *retweeted_status*. An alternative is to match the RT @ symbol, but we found it returned some alias and was not very effective. Thus, we used *retweeted_status* and built the retweet corpus which consisted of 6,246,318 retweets with 8,759,537 users. We then parsed the retweet corpus and created the TAP graph for each query topic.

In our proposed Topical Authority Propagation (TAP) method, s_{ji} is the topical relevance score of the tweet of user i that user j has retweeted. This score can be computed by any retrieval model such as vector space model and language modeling. In the experiments, s_{ji} was calculated by the Okapi BM25 ranking function. We chose a competitive content-based baseline for comparison. The baseline method measures topical relevance of the user: rank candidates according to the relevance between the user's retweets and the query. Similarly, Okapi BM25 was used to measure the topical relevance. This baseline only considers content based relevance and does not take authority into account. In the preprocessing, a standard list of stopwords were removed from the corpus and no stemming was applied.

It is worth noting that our dataset did not cover all the tweets during the period because Twitter Public Streaming API only provides a small random sample of all the tweets. Consequently, our testbed may not include all the topical authorities given a specific topic. Therefore, it is not appropriate to directly compare our results with the prior work such as [6, 4] that utilized the full data of Twitter. In addition, some of the query topics in the prior work [6] were trending queries such as *oil spill* which was breaking news three years ago, while people rarely discussed these topics recently. As a result, we can hardly get any informative tweets on these topics and it is also less valuable to study such topics as they are outdated. In the evaluation, we choose topics such as *IRS* which is trending during the period when we collected the data. We also use some stable topics such as *music*, since people discuss them all the time. We used five query topics: *Tornado*, *Media*, *Obama*, *Music*, *IRS* in the experiments. The choice of the query topics is also similar to [6] where three query topics were used in their evaluation. The evaluation metrics in the experiments are Precision at 5 (P@5) and Precision at 10 (P@10).

4.2 Comparison with the Baseline Method

Table 1 shows P@5 and P@10 of the BM25 baseline and our proposed TAP method on the five query topics. As we can see, TAP yielded consistently better performance than BM25 did across all the queries on both metrics. For

[1] https://dev.twitter.com/docs/streaming-apis

Table 1: Topical Authority Propagation (TAP) vs Okapi BM25 for the five query topics

	P@5	P@10
Topic 1: *Tornado*		
BM25	0.6	0.4
TAP	1	1
Topic 2: *Media*		
BM25	0.8	0.6
TAP	1	1
Topic 3: *Obama*		
BM25	0.8	0.8
TAP	1	1
Topic 4: *Music*		
BM25	1	0.7
TAP	1	1
Topic 5: *IRS*		
BM25	0.8	0.7
TAP	1	0.9

topic $Tornado$, all the top 5 users identified by TAP are true topical authorities while BM25 only returns 3. The improvement is more substantial for top 10 results for which TAP achieved P@10 of 1 and BM25 yielded 0.4. All the top 10 users returned by TAP are topical authorities while only 4 from BM25 are. We can see the similar pattern for the other four query topics. By considering authority propagation via retweets, TAP can bring substantial improvement over the baseline which only utilizes the topical relevance. In the near future, we will conduct a more comprehensive set of experiments to further validate the TAP approach.

4.3 Impact of the Amount of Data Available

One of the major challenges that hindered our thorough analysis is the lack of full access to all the tweets. Many tweets on trending topics such as IRS may not be included in our corpus. To quantitatively investigate the effect of the amount of available data on our method, we evaluated the performance of TAP for the same five queries by varying the amount of data. We formed seven corpora by extracting data from various durations, i.e., the n-day corpus where n is the number of days during which the tweets were crawled where n is ranged from 1 to 7. Specifically, the 1-day corpus consists of the tweets we crawled on May 16th and 5-day corpus includes the tweets from May 16th to May 20th. The results shown in the previous section are from the 7-day corpus which is the tweet data we crawled during the whole week.

For each corpus, we computed the averages of P@5 and P@10 across the five query topics. The experimental results are shown in Figure 2. The blue solid line represents averaged P@5 and the red dash line represents averaged P@10. As we can see, the precisions increase when more data are available. The reason may lie in the fact that with more tweets, retweets, and users, the TAP graph becomes dense and thus authority propagation is further enhanced. It is expected that with full access to tweets, TAP could be a very effective approach to identifying topical authorities.

5. CONCLUSIONS AND FUTURE WORK

This paper addresses the task of finding authorities on Twitter given any query topic. We propose a novel Topical Authority Propagation (TAP) model by utilizing the fact that topical authority can be propagated through retweeting, i.e., if a user's tweet is retweeted by a topical authority, that user is likely to be an authority on the topic as well. Topical relevance of the candidate authorities can be seamlessly integrated into the model. Link analysis algorithms such as PageRank can then be utilized to characterize how topical authority is propagated through retweeting. We conduct experiments on the testbed from the retweet corpus crawled from Twitter API and demonstrate the strength of the proposed approach by comparing with the baseline that uses topical relevance only. We also show the effect of various amount of data available on the model performance. In the future, we will conduct a more comprehensive set of experiments. We will also extend the proposed model to incorporate other available information such as mention, reply, and user profile.

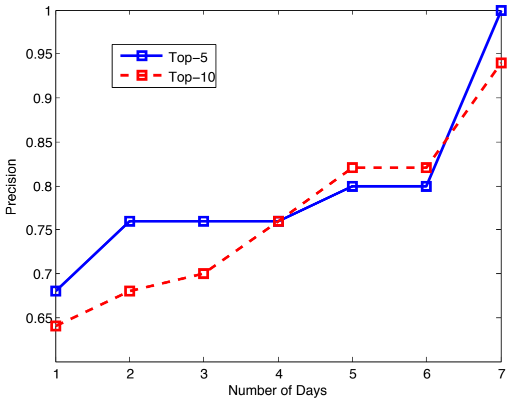

Figure 2: The impact of the amount of available data on Precisions

6. REFERENCES

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, pages 65–74, 2011.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, pages 8–15, 2010.

[4] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*, pages 575–590, 2012.

[5] C. Lee, H. Kwak, H. Park, and S. Moon. Finding influentials based on the temporal order of information adoption in twitter. In *WWW*, pages 1137–1138, 2010.

[6] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, pages 45–54, 2011.

[7] 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.

[8] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.