RAProp: Ranking Tweets by Exploiting the Tweet/User/Web Ecosystem and Inter-Tweet Agreement

Srijith Ravikumar[†], **Kartik Talamadupula**[†], **Raju Balakrishnan**[§] and **Subbarao Kambhampati**[†]

[†]**Dept. of Computer Science and Engg.**
Arizona State University
Tempe AZ 85287
{sraviku2,krt,rao} @ asu.edu

[§]**Groupon, Inc.**
3101 Park Blvd
Palo Alto CA 94306
raju @ groupon.com

Abstract

The increasing popularity of Twitter renders improved trustworthiness and relevance assessment of tweets critical for search. However, given the limitations on the size of tweets, it is hard to extract measures for ranking from the tweets' content alone. We propose a method of ranking tweets by generating a *Feature Score* for each tweet that is based not just on content, but also additional information from the Twitter ecosystem that consists of users, tweets, and the webpages that the tweets link to. The Feature Score is propagated over an *agreement graph* based on tweets' content similarity. The propagated Feature Score that is sensitive to content popularity and trustworthiness is used to rank the tweets for a query. An evaluation of our method on 16 million tweets from the TREC 2011 Microblog Dataset shows that it doubles the precision over the baseline Twitter Search, and outperforms the best-performing method on the TREC 2011 Microblog dataset.

1 Introduction

Twitter, the popular microblogging service, is increasingly being looked upon as a source of the latest news and trends. The open nature of the platform, as well as the lack of restrictions on who can post information on it, leads to fast dissemination of the latest information. This open nature, however, proves to be a double-edged sword and leaves Twitter extremely vulnerable to the propagation of false information from profit-seeking and malicious users (*cf.* New York Times; The Economist). Unfortunately, Twitter's native search does not seem to consider the possibility of users crafting malicious tweets, and instead only considers the presence of query keywords, number of retweet instances, and recency of tweets (Twitter Support). This takes very little note of *content-based similarity* and hence content-centric popularity. Twitter tries to address this issue by filtering out spam tweets (Twitter Official Blog). However, while tweets that Twitter identifies as spam may well be untrustworthy, it cannot be assumed that tweets not marked as spam are all trustworthy. Moreover, providing correct and relevant information often requires more than just removal of spam (*cf.* New York Times). The popularity and credibility of a user alone may not be the solution to filtering out untrustworthy tweets; such accounts have been used to spread hoaxes (Twitter Account Hack). Even when tweets are not maliciously manipulated, they may end up being incorrect content-wise; hence more ways are needed to quantify and measure the content-based popularity along with the trustworthiness.

2 *RAProp*: Our Method

In order to measure the trust and popularity of a tweet in real-time (Feature Score), we use the following features from the Twitter ecosystem: tweet content, user, linked web pages, and relationships between these (via that tweet). Ranking based on Feature Score is heavily biased towards popular users and tweets. Tweets that pertain to the most popular topic may be assumed to be more relevant to the query (given that the recency of a tweet is important on Twitter). We use the pair-wise tweet agreement as endorsement on topics, and this endorsement is used to find the popularity of a topic. We combine these orthogonal measures of trustworthiness and popularity of a tweet (Feature Score) and the content based popularity (Agreement) by propagating the Feature Score over the "Agreement Graph" (see Figure 1). The results are then ranked based on this propagated Feature Score.

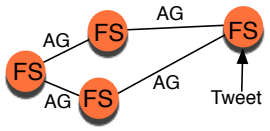

Figure 1: *Propagation of Feature Sores(FS) over Agreement Graph (AG).*

Feature Score: In order to compute the Feature Score for a tweet, we model the Twitter ecosystem as a graph consisting of: (i) user layer; (ii) tweet layer; and (iii) webpage layer. Features from the user layer are transmitted to the tweet layer by using the "tweeted-by" links; features from the web layer are transmitted to the tweet layer by the URLs in tweets. The user layer considers the follower count; friends count; whether that user (profile) is verified; the time since the profile was created; and the total number of statuses (tweets) posted by that user. We use the PageRank of the URL mentioned in the tweet as the score in the web layer. The tweet itself has specific features that may be used as part of the tweet layer: whether the tweet is a re-tweet; the number of hashtags; the length of the tweet; whether the tweet mentions a user; the number of favorites received; the number of re-tweets received; whether the tweet contains

affect; and the TF-IDF similarity of the tweet to the query, which is weighted by the proximity of the query keywords in the tweet. The user, tweet and the web features are used by a random-forest learner to compute the Feature Score of a tweet. We train the learner on 5% of the TREC gold standard, which we use.

Agreement: To keep the cost of agreement computation low, we use a modified version of TF-IDF similarity. We compute TF-IDF similarity on the stop-word removed and stemmed result set, R. We compute the IDF value of the TF-IDF similarity for each result set R_{Qi} separately. This ensures that the IDF value of the query term as well as other common words in the result set is negligible in the similarity computation, and guarantees that the agreement computation is not affected by this. Due to the sparsity of verbs and other stop-words in tweets, the IDF of some verbs is much higher than nouns and adverbs. Hence, we weight each part of speech used in the TF-IDF computation differently, such that parts that are important for agreement on Twitter are weighted higher. Instead of using $L2$ normalization, we use the highest TF value among the two documents being compared. This penalizes tweets with repeated content.

Ranking: In order to rank a set of tweets for a given query, we pick an initial set $R_{Q'}$ that is then filtered to remove retweets and replies as they are considered to be irrelevant to the query by the gold standard. We add more terms to the query Q' to get an expanded query, Q. The expansion terms are selected from $R_{Q'}$ by picking the top-5 nouns based on TF-IDF score. The set of the top-N tweets returned for the expanded query becomes the result set, R_Q. We compute the Feature Score and pairwise agreement for all the tweets in the result set R_Q as described above. We then propagate the Feature Score over the agreement graph once, to get the propagated Feature Score. The tweets are finally ranked based on this score.

3 Evaluation

Dataset: For our evaluation, we used the TREC 2011 Microblog Dataset and gold standard (NIST). Our experiments were conducted on the 49 queries that are provided along with this dataset. We used the Pagerank API in order to collect the PageRank of all the web URLs mentioned in the tweets in this set.

Experimental Setup: Using the set of returned tweets R_Q, we evaluate each of the ranking methods. Since our dataset is offline (due to the use of the TREC dataset and the gold standard as described above), we have no direct way of running a Twitter search over that dataset. We thus simulate Twitter search (*TS*) on our dataset by sorting a copy of R_Q in reverse chronological order (i.e., latest first).

Results: We compared the top-K Precision at 5, 10, 20, 30 and mean average precision (MAP) of *RAProp* with the simulation of Twitter's native ranking, and the current state of the art (USC/ISI). Since not all relevant tweets from the dataset for the query were part of the gold standard, we ignored those tweets that are not part of the gold standard while computing the precision value. We picked the N most recent tweets that contained one or more of the query keywords. As seen in Figure 2, *RAProp* shows significant improvement over Twitter Search (*TS*). We also show better precision values than the current state of the art method on this dataset, USC/ISI (Metzler and Cai 2011). Since in most cases there were less than K relevant documents in R (maximum achievable $P@30$ is 0.684), the precision values are expected to drop as the value of K increases. However, *RAProp* maintains its dominance over both the methods. Additionally, the MAP values show that *RAProp* is able to place relevant results higher than the other methods. These results thus confirm our claim that *RAProp* is an effective ranking method for Twitter search (see (Ravikumar 2013) for additional details and discussion.)

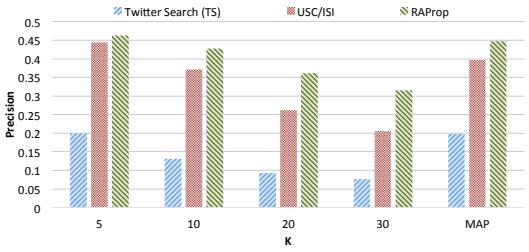

Figure 2: *Comparison of RAProp against the baseline (Twitter Search) and current state of the art, USC/ISI.*

4 Related Work

Although ranking tweets has received attention recently (c.f. NIST; Metzler and Cai), much of it is focused only on relevance. Most such approaches need background information on the query term which is usually not available for currently hot topics. There are also multiple approaches (Duan et al. 2010; Jiang et al. 2012) that try to rank tweets based on specific user features; *RAProp* complements these by adding trustworthiness of the tweets to the ranking algorithm. Credibility analysis of Twitter stories has been attempted by Castillo et al. (Castillo, Mendoza, and Poblete 2011); our problem differs in having to assess the credibility of individual tweets. Finding relevant and trustworthy results based on implicit and explicit network structures has been considered previously by (Balakrishnan and Kambhampati 2011).

5 Conclusion

In this paper, we proposed *RAProp*, a microblog ranking mechanism for Twitter that considers the relevance and trustworthiness of the underlying content, in order to filter out irrelevant results and spam. *RAProp* works by propagating a computed Feature Score for each tweet and propagating that over a graph that represents content-based agreement between tweets, thus leveraging the collective intelligence embedded in tweets. Our detailed experiments (Ravikumar 2013) on a large TREC dataset showed that *RAProp* improves the precision of the returned results significantly over Twitter's own search, and beats the existing state of the art method consistently.

Acknowledgements: This research is supported in part by ONR grants N000140910032 and N00014-13-1-0176, NSF grant IIS201330813 and a Google Research Award.

References

[Balakrishnan and Kambhampati 2011] Balakrishnan, R., and Kambhampati, S. 2011. Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of WWW*.

[Castillo, Mendoza, and Poblete 2011] Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of WWW*.

[Duan et al. 2010] Duan, Y.; Jiang, L.; Qin, T.; Zhou, M.; and Shum, H. 2010. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 295–303. Association for Computational Linguistics.

[Jiang et al. 2012] Jiang, J.; Hidayah, L.; Elsayed, T.; and Ramadan, H. 2012. Best of kaust at trec-2011: Building effective search in twitter. In *Proceedings of the 20th Text REtrieval Conference (TREC 2011)*.

[Metzler and Cai 2011] Metzler, D., and Cai, C. 2011. USC/ISI at TREC 2011: Microblog Track. In *Proceedings of the Text REtrieval Conference (TREC 2011)*.

[New York Times] Twitter death hoaxes, alive and sadly, well. http://nyti.ms/10qVW9j.

[NIST] TREC 2011 Microblog Track. http://trec.nist.gov/data/tweets/.

[Ravikumar 2013] Ravikumar, S. 2013. Raprop: Ranking tweets by exploiting the tweet/user/web ecosystem. Master's thesis, ARIZONA STATE UNIVERSITY.

[The Economist] Zombie followers and fake re-tweets. http://www.economist.com/node/21550333.

[Twitter Account Hack] In hacking, A.P. Twitter feed sends false report of explosions. http://nyti.ms/15EyoVV.

[Twitter Official Blog] State of Twitter Spam. http://bit.ly/d5PLDO.

[Twitter Support] About top search results. http://bit.ly/IYssaa.