Who Will Tweet More? Finding Information Feeders in Twitter

Beibei Gu, Zhunchen Luo(✉), and Xin Wang

China Defense Science and Technology Information Center, Beijing, China
gubeiguying@sina.com, zhunchenluo@gmail.com, 20150101xl@sina.cn

Abstract. Twitter is an important source of information to users for its giant user group and rapid information diffusion but also made it hard to track topics in oceans of tweets. Such situation points the way to consider the task of finding **information feeders**, a finer-grained user group than domain experts. Information feeders refer to a crowd of topic tracers that share interests in a certain topic and provide related and follow-up information. In this study, we explore a wide range of features to find Twitter users who will tweet more about the topic after a time-point within a machine learning framework. The features are mainly extracted from the user's history tweets for that we believe user's tweet decision depends most on his history activities. We considered four feature families: **activeness, timeliness, interaction** and **user profile**. From our results, activeness in user's history data is most useful. Besides that, we concluded people who gain social influence and make quick response to the topic are more likely to post more topic-related tweets.

1 Introduction

Twitter, one of the most successful social media platforms with giant user groups and a cornucopia of information, has already become a major channel for content distribution where gathers first-hand information of most influential events and topics worldwide. In the meanwhile, information environment in Twitter is complex, where messages are in form of tweets within 140 characters, usually brief, massive and highly distributed, leading to data sparseness and redundancy for traditional information retrieval for a given topic. How to efficiently capture useful messages in an ocean of data is a hard question left to researchers. Here, we consider to find informing users to avoid some disadvantages.

Users are thought to be the center of releasing and distributing multi-sources information with the backup of their social networks. Out of interest or duty, some people will pay continuous attention to some certain topic and keep tweeting subsequent information as the topic continues and evolves. This kind of people usually have long term interests in topic-related fields. They probably have accumulated a certain amount of relevant knowledge and collected some reliable information sources, making themselves potential information providers of the topic.

C.-Y. Lin et al. (Eds.): NLPCC-ICCPOL 2016, LNAI 10102, pp. 437–448, 2016.
DOI: 10.1007/978-3-319-50496-4_36

We aim to identify people with the potential to keep releasing information about a topic. We call them *information feeders*. Obviously, rapidly identifying information feeders offers a new approach to keep track of topics directly from information sources and may avoid situations such as unpredictable subject terms caused by topic floating by means of keyword searching [9].

It is noteworthy that, different from **domain experts**, which usually means people with some expertise or experience about a certain subject, the concept of "information feeder" refers to a finer-grained user group, namely topic tracers, and especially emphasizes those who have a relatively high probability to give out further information on a specific topic. Online information explosion is simply too much for experts to allocate their finite attention for each and every topic within the domain, as a result, an expert does not necessarily keep track of a topic all the way, but an information feeder does. Information feeders around a topic unit are usually highly dynamic during the topic evolution, while domain experts are rather static. Instant recognition of the aforementioned type of users is of interest to information seekers like journalists and companies. This is a challenging task in face of various user characteristics and unpredictable changes over time.

In this paper, we explore the way to identify information feeders within the huge amount of Twitter users in conjunction with given topics. We formulate the task as a binary classification problem and apply a machine learning framework for predicting whether a user will tweet more about the topic, which relies on four feature families: activeness, timeliness, interaction and user profile. From our results, activeness in user's history data is thought to be most useful and that users with some social influence and quick response to the topic are more likely to continue to post topic-related tweets.

The main contributions of this paper can be summarized as follows. Firstly, to the best of our knowledge, this paper is the first to predict whether a user will continue to tweet more on certain topics and such users are so-called "information feeders" in this work. Moreover, this paper has presented a novel set of features and approaches for predicting information feeders. Finally, we build our own annotated data for the attributes concerned. All of the manually-annotated Twitter data sets developed in this work will be made available as a new shared resource to the research community.

2 Related Work

The public nature of Twitter and the cornucopia of users as well as information sources have made it a hot topic focused and lasted during recent years. Related work can be divided into following parts:

User Behaviour Analysis and Prediction. Efforts on users' behavior prediction mainly focus on retweeting, which is regarded as an important pattern in information propagates. Suh *et al.* [16] provided with a detailed and large-scale analysis of factors that have an impact on retweeting. The number of followers and friends showed much impact in their results. Boyd *et al.* [2] treated retweeting as a means of participating in a diffuse conversation, and presented a very in-depth study about retweeting in diverse ways through actually interviewing Twitter users on the reasons why and what they retweet most. Zaman *et al.* [20] trained a probabilistic collaborative filter model for predicting the spread of information via retweet in Twitter network. They found that the identity of the source of the tweet and retweeter were most important features for prediction. Artzi *et al.* [1] predicted the likelihood of a retweet through a discriminative model. Luo *et al.* [10] firstly brought up with a learning-to-rank framework to find out retweeters to a certain tweet, showing that the retweet history and the similarity between the content of the tweet and the posting times of followers are most effective for the task. In this paper, we make prediction on whether a user will continue to post messages related to a certain topic, including retweets.

Demographics in Twitter. User feature analysis is an important part in our method. A lot of achievements on latent attribute inference of Twitter users have been made, with recent work focusing on age [11], gender [15], user profile extraction [4,8], location [3,6], occupational class [13], political tendency [17], voting intention [7] and brand preferences [18], among which various research angles have been applied for different purposes. Our work builds on these findings to predict users that will tweet more on certain topics.

User Identification in Twitter. Twitter has collected all kinds of user types together, of which the defined information feeders can also be viewed as one. Diakopoulos *et al.* [5] is a related work for identifying credible sources. However, they aimed at getting access to information sources for journalists' reporting mission, while we intended to predict how many topic-related messages an information feeder will continue to provide for topic tracking. Zafarani *et al.* [19] developed a methodology that identifies malicious users with limited information. They made a detailed analysis of five general characteristics of malicious users and demonstrated that 10 bits of information can help a lot in the task.

3 Method

3.1 Task Description

In this paper, we present the task on automatically predicting whether a user will post more topic-related tweets. Given a topic T, we retrieve tweets and obtain initial user set U who have posted topic-related tweets from retrieval results. Our goal is to train a classification model R that predicts whether user u from U will continue to tweet about T.

The set of features we explore below is used in conjunction with a supervised machine learning framework providing models for binary classification. From the user information and their tweet data, we extracted features related to the prediction of information feeders. In the following, we describe our feature sets in more detail.

3.2 User Features

It is observed that decisions of a user can be explained better by his activity in the recent past, i.e., temporally local history [14]. A user's decision of tweeting more about topic T depends significantly on his temporal behavior. Thus the recent topic-related data of the user is considered to contain important information about his tweeting decision on topic T. **Activeness, timeliness** and **interaction** are three main aspects of the user's recent behavior characteristics that we analysis. We also believe that a user's basic **profile** indicates his general image on Twitter. Hence, we explore user features from these four dimensions. A summary of features shows in Table 1.

Table 1. Summary of features for information feeders

Feature family	Feature name	Description
Activeness	Count_Tw	Number of all tweets during the *period* (from the first topic-related tweet's posting time to the time t)
	Count_RelaledTw	Number of topic-related tweets posted by time t
	Ratio_RelatedTw	Ratio of topic-related tweets to all tweets during the *period*
	Ratio_RelatedOr	Ratio of original topic-related tweets to topic-related tweets
Timeliness	TD_Related	Time difference between the latest two topic-related tweets by time t, in seconds
	Response_Time	Time difference between the initial time of topic and the first topic-related tweet's posting time in seconds
Interaction	Ratio_Mt	Ratio of tweets with @username in topic-related tweets
	Ratio_Rt	Ratio of retweets in topic-related tweets
	Ratio_Fav	Ratio of favorites in topic-related tweets
Profile	Count_Fol	Number of user u's followers
	Count_Fri	Number of user u's friends
	Topic_Similarity	Similarity of user u's history tweets and the topic description

Activeness. Instinctively, an active user usually receives more information from all aspects and creates more tweets. The number of tweets posted in his recent past (i.e., the period from the beginning of topic T to the time when we collected the user data) indicates user u's recent activeness. We include the count of all tweets (**Count_Tw**) as well as topic-related tweets posted

(**Count_RelaledTw**) during that period as two features to measure user u's activeness on Twitter, especially on topic T. We also think the ratio of topic-related tweets (**Ratio_RelatedTw**) during the recent history describes the user's concentration on topic T.

Original tweets refer to those whose contents are edited by the user himself. Editing original tweets usually means new information, which requires to learn enough knowledge about topic T and form his own understanding. The ratio of original topic-related tweets (**Ratio_RelatedOr**) describes the user's tweet originality to some degree and can be regarded as an indicator of the user's activeness to T.

Timeliness. Information feeders are those who are willing to pay plenty of attention to topic T and keenly aware of the topic update. They are usually quick to keep up with a new topic with interest and provide fresh information about it whenever it has new evolution. So we regard user u's timeliness towards topic T as a measurement of u's interest in T.

Two features are selected to reflect user timeliness: **TD_Related** and **Response_Time**. The former describes the time difference between the latest two topic-related tweets, an expression of u's recent update frequency of information about T. **Response_Time** denotes how long it took u to post his first topic-related tweet from the start time of T. However, the initial time of a topic is usually hard to capture, so we replace it with the time of the earliest topic-related tweet in our dataset. Both the features are measured in seconds.

Interaction. Interaction in Twitter is a great motivation for users to get involved in information creation and diffusion. Mentions, giving likes and retweeting are three major mechanisms for user interaction. Posting tweets with mentions are meant to send information specifically to somebody which may possibly bring about a tweet stream between the users. Moreover, people usually give likes or retweet to show their agreement to the user's opinion or information. This can be seen as an encouragement for the user to post related tweets.

An information feeder is more likely to be encouraged by interaction with others. Thus we calculate the ratio of u's topic-related tweets with mention (**Ratio_Mt**) and the ratio of tweets got favorites (**Ratio_Rt**) or retweeted (**Ratio_Fav**) by others.

Profile. This feature family contains three features that can give an overview of the user's general image on Twitter, which are the numbers of the user's followers (**Count_Fol**) and friends (**Count_Fri**), and the similarity of the topic description and users' previous tweets (**Topic_Similarity**).

Count_Fol is a major factor of his influence and also a reflection of the quality of tweets. A regular information feeder may have gained his reputation and attracts a number of followers for his tweets. We include the feature **Count_Fri** for similar reason.

User's interest is another part of user's profile. If something has ever drawn one's attention, he is likely to be attracted for a second time when a new topic about it shows up. Take the topic *"Diesel gate of Volkswagen"* for example, if a user once tweeted about news about vehicles, which indicates he used to have interests in it, he is far more likely to be attracted by Volkswagen's emission cheating case and to post some messages about that than those who showed no interest in automotive news. It inspires us to calculate the similarity of the topic description and users' previous tweets (**Topic_Similarity**). When calculating the value of similarity, we filtered the top 100 high frequent words and the words which appear less than 5 times in our collected data [10].

4 Experiments

4.1 Data Preparation

To the best of our knowledge, there is no annotated dataset available, so we created labeled data required for this task. We document in detail our analytical method and the way we collected our data set. We randomly chose five topics of interest, including a live topic of the moment *#AlphaGo*, a gusty topic *#Turkey Ankara explosion*, a long-term topic *#American 2016 Presidential Election*, and two cooling topics *#NASA astronaut return to Earth* and *#Gravitational Waves*).

We searched for the hash tags of the topics and collected a significant number of topic-related tweets through the Twitter API for a whole day on March 17th, 2016. Hence, we got initial user set U. Then we filtered those who tweeted less than 500 tweets in total and whose tweeting frequency was beyond 30 and below 0.3 posts per day on average in order to reject inactive users and robots. About 200 users were randomly selected respectively for each topic from the filtered user set. 3200 recent tweets[1] posted by each user was crawled on March 27th, 2016. Our limitation to users' tweeting frequency makes sure that the crawled data covers all tweets posted from the beginning of our topics.

Two people involved in the manually annotating topic-related tweet process. The annotation process is applied with elicitation methods and take the starting time of each topic as well as their keywords as assistance. For each topic, any tweet with information related to the topic is labeled as "Related", and unwanted users such non-English users were rejected through judging by human experience. Final number of valid users in our dataset is 438, and 8,297 topic-related tweets were annotated. Table 2 displays the statistics of our data.

4.2 Data Description

The temporal distribution of the topic-related tweets for each topic is displayed in Fig. 1. In the pictures, we can see that the distributions of related tweets for the five topics respectively have different trending features. Although most topics follow the power law distribution with a peak near the beginning of the

[1] The maximum limitation of Twitter REST API is 3200 recent tweets per user.

Table 2. Data statistics of each topic

Number of valid users	438
Total number of tweets[a]	976,532
Number of topic-related tweets	8,297
Average number of topic-related tweets per user	18.90

[a] This number refers to the summation of tweets of all valid users we obtained from Twitter API.

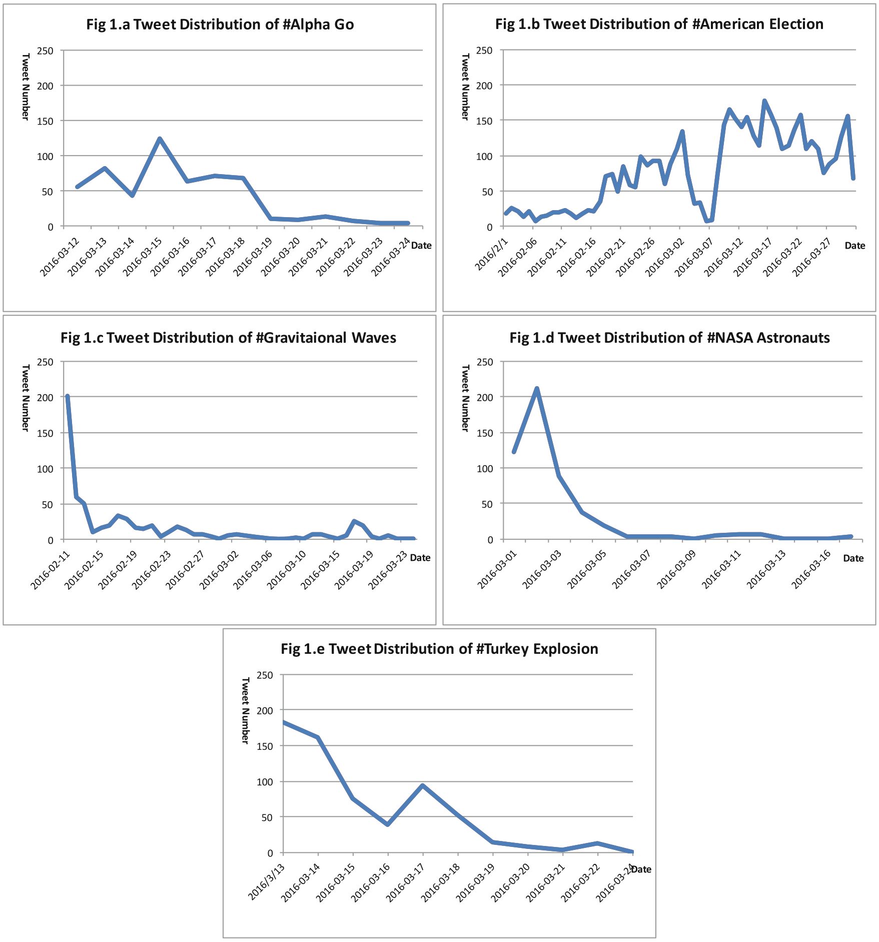

Fig. 1. From Fig. 1.a to 1.e, there are *(1) #AlphaGo, (2) #American 2016 Presidential Election, (3) #Gravitational Waves, (4) #NASA Astronauts Return to Earth* and *(5) #Turkey Ankara Explosion* related tweet distribution in sequence.

topic discussion time and then following a decrement, the duration and strength of each peak and the decay rate of each decrement have nothing in common with each other.

The long-term topic *American 2016 Presidential Election* shows a rather special distribution of topic-related tweets. From the distribution graph, tweets posted at the beginning of the preparation period of the elections are steady and rather sparse while a series of small peaks show up in sequence. It is totally different from the rest topics. The reason may be that there are plenty of movements during the elections, which make the subtopics and motivates bursts of tweeting.

The time we seleceted for predicting the user's next topic-related tweet stands right in different states of the five topics. We can assume, therefore, our method has general applicability for various types of topics.

4.3 Experiment Setting

We make "a user and a time-point" as a sample to predict the next topic-related tweet. For example, Twitter user u posted a tweet about topic T at time t. Our goal is to predict whether a will post another message about topic T after t.

In this section, we evaluated our dataset empirically using a SVM model with default parameters and a 5-fold cross validation was performed. Each validation has one fifth of dataset as testing set and the left as training set. Time t can be set as any day in our model while we take the day we harvested the users as the time t, namely March 17th. In our dataset, there are 184 positive instances and 254 negtive instances.

We evaluate the performance with two metrics as *accuracy* and *F-score*. *Accuracy* refers to the number of instances where the method correctly classified which user will continue to tweet after time t. *F-score* is standard in information retrieval where there is a similar imbalance between the relevant and non-relevant classes [12].

4.4 Results

To the best of our knowledge, our task is relatively new and we didn't find other methods for similary tasks. So we evaluate the effectiveness of our approach by devising one baseline method for comparison. When individuals are asked to guess whether a user will tweet more on a certain topic, they will probably review the users previous tweets for similar topic-related posts if not making random conjectures. Hence, we set our baseline as follows:

Baseline: Posted Ever. We consider that a user who has posted more than one topic-related tweet ever is an information feeder.

From our annotated data, we labeled the users by whether they posted topic-related data before March 17th as ground truth.

Comparison of Feature Families. In this part, we display our feature effectiveness by testing feature families along with the baseline method using SVM. As a baseline, we use a feature **PostedEver** indicating whether a user has

Table 3. Results for different feature families with SVM (Bold numbers denote the best).

Feature set	Accuracy	F-Score
PostedEver	0.5321	0.5655
Activeness	0.6134	**0.7434**
Timeliness	0.5878	0.6669
Interaction	0.5991	0.7215
Profile	0.5907	0.7315
PostedEver + Activeness	0.6179	0.7399
PostedEver + Timeliness	0.5920	0.6698
PostedEver + Interaction	0.5951	0.7112
PostedEver + Profile	0.5865	0.7251
Full	**0.6551**	0.7372

posted topic-related tweets ever with boolean value for modeling. Results are summarized in Table 3.

We can see that experiments with full feature set gained best performance, giving us a huge improvement in both accuracy and F-score over the baseline. *Activeness* features provided the highest F-score and a relatively good accuracy 0.6134. *Interaction* and *Profile* features showed an average level in all the metrics while *Timeliness* features had very poor F-score. Each feature family and their combination showed relatively great effectiveness for our task and overrode the baseline method.

Besides, all of our feature families improve the classification performance over the baseline method. The combinations of baseline and each feature family significantly improve the results when used with the baseline method in isolation.

Feature Analysis. We investigate whether our features can improve tweet prediction and are also interested in which features in particular are highly valued by our model. We combine each feature with baseline feature within our framework.

Table 4 shows the performance of each classification model. The features are ranked by F-score. We can see that all of our features improve the results with statistically significance.

All the four features of *Activeness* provide pretty good performance in testing models, ranking within the topic five, revealing that users' history information is helpful in our task, especially user activeness during the recent past. The result of **Ratio_RelatedOr** also proves that tweet originality is a strong indicator to user's interest in topic T which drives him to continue to tweet.

We also find that social features of a user perform well. **Count_Fol** brought about pretty good scores of accuracy and F-score, which means that user's influence may motivate him to tweet more.

Table 4. Performance of each classification model.

Feature set	Accuracy	F-Score
PostedEver	0.5321	0.5655
PostedEver + Ratio_RelatedTw	0.6218	0.7459
PostedEver + Count_RelatedTw	0.5942	0.7440
PostedEver + Count_Tw	0.5962	0.7428
PostedEver + Count_Fol	0.5907	0.7418
PostedEver + Ratio_RelatedOr	0.5872	0.7396
PostedEver + Response_Time	0.5869	0.7396
PostedEver + Ratio_Mt	0.5897	0.7352
PostedEver + Topic_Similarity	0.5865	0.7272
PostedEver + Ratio_Fav	0.5849	0.7252
PostedEver + Count_Fri	0.5820	0.7251
PostedEver + Ratio_Rt	0.5734	0.6749
PostedEver + TD_Related	0.5891	0.6659

Response_Time is another useful feature with a substantial improvement of about 5 points in accuracy and 17 points in F-score over the baseline. **Response_Time** stands for user's timeliness to topic T by measuring time it took the user to make response to a new topic. To a large extent, a user with little time's delay to keep up with a new topic is usually engaged in it and willing to tweet more.

The significant effectiveness of **Count_Fol** and **Response_Time** illustrates that user influence and timeliness on a certain topic are important indicators to whether a user will become an information feeder. Users with a range of followers and quick response to a topic are more likely to continue to pay attention to topic T and post more topic-related tweets.

5 Examples

Here are some examples showing the usefulness of our features.

ScottyFinch, a frequent Twitter user who provided a live report about the matches between AlphaGo and Lee Sedol from the staring time of the topic. 38 tweets related to Alphago were posted by March 17th (the time we collected our data), making up more than 30% in his tweet timeline. He shows a high possibility to keep tweeting on the AlphaGo topic. Our method predicted that *ScottyFinch* is an information feeder and actually he did tweeted a lot more after March 17th.

A counter-example is ***wildhare***, who posted 1,523 tweets in total during the topic *Gravitational Waves*'s discussion time, but only 10 retweets were about the topic. His first topic-related tweet was 4.5 hours later when the bursting news came out. *wildhare* showed no concentration or strong interest in the topic with low update rate. He posted no more tweets about *Gravitational Waves* and our method predicted so.

6 Conclusion and Future Work

In this paper, we studied the task of finding information feeders by predicting whether a user will tweet more about certain topics. This is a new task and our results benefit information seekers for acquiring topic-related information more efficiently and effectively via information feeders in Twitter, and also broaden ways to make better use of social media information.

We focus on users history tweet features for our predictive models, including users' activeness, timeliness and interaction features in the temporally local history, as well as user profile features. From the results, we find people who show plenty of concentration on information about T and active in the topic discussion are more likely to be information feeders.

Our approach is very flexible and allows for improvements on our current models by incorporating information such as users neighborhood status in Twitter as well as on other social media platforms. In the future we plan to apply new features to improve the performance of our predictive model and explore futher into topic specific tasks.

References

1. Artzi, Y., Pantel, P., Gamon, M.: Predicting responses to microblog posts. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 602–606. Association for Computational Linguistics (2012)
2. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: 2010 43rd Hawaii International Conference on System Sciences (HICSS), pp. 1–10. Institute of Electrical and Electronics Engineers (2010)
3. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating Twitter users. In: Proceedings of the 19th Association for Computing Machinery International Conference on Information and Knowledge Management, pp. 759–768. Association for Computing Machinery (2010)
4. Culotta, A., Ravi, N.K., Cutler, J.: Predicting the demographics of Twitter users from website traffic data. In: Proceedings of the International Conference on Web and Social Media (ICWSM). AAAI Press, Menlo Park (2015, in press)
5. Diakopoulos, N., De Choudhury, M., Naaman, M.: Finding and assessing social media information sources in the context of journalism. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2451–2460. Association for Computing Machinery (2012)
6. Jurgens, D.: That's what friends are for: inferring location in online social media platforms based on social relationships. ICWSM **13**, 273–282 (2013)
7. Lampos, V., Preotiuc-Pietro, D., Cohn, T.: A user-centric model of voting intention from social media. In: Association for Computational Linguistics, vol. 1, pp. 993–1003 (2013)

8. Li, J., Ritter, A., Hovy, E.: Weakly supervised user profile extraction from Twitter. In: Association for Computational Linguistics, Baltimore (2014)
9. Lin, J., Efron, M., Wang, Y., Sherman, G.: Overview of the TREC-2014 microblog track. Technical report, DTIC Document (2014)
10. Luo, Z., Osborne, M., Tang, J., Wang, T.: Who will retweet me?: finding retweeters in Twitter. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 869–872. Association for Computing Machinery (2013)
11. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: How old do you think i am?; a study of language and age in Twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. AAAI Press (2013)
12. Petrovic, S., Osborne, M., Lavrenko, V.: RT to win! Predicting message propagation in Twitter. In: International Conference on Weblogs Social Media (2011)
13. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content. In: Association for Computational Linguistics (2015)
14. Rangnani, S., Devi, V.S., Murty, M.N.: Autoregressive model for users retweeting profiles. In: Liu, T.Y., Scollon, C.N., Zhu, W. (eds.) SocInfo 2015. LNCS, vol. 9471, pp. 178–193. Springer International Publishing, Heidelberg (2015)
15. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in Twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents, pp. 37–44. Association for Computing Machinery (2010)
16. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 177–184. Institute of Electrical and Electronics Engineers (2010)
17. Volkova, S., Coppersmith, G., Van Durme, B.: Inferring user political preferences from streaming communications. In: Proceedings of Association for Computational Linguistics, pp. 186–196 (2014)
18. Yang, C., Pan, S., Mahmud, J., Yang, H., Srinivasan, P.: Using personal traits for brand preference prediction (2015)
19. Zafarani, R., Liu, H.: 10 bits of surprise: detecting malicious users with minimum information. In: Proceedings of the 24th Association for Computing Machinery International Conference on Information and Knowledge Management, pp. 423–431. Association for Computing Machinery (2015)
20. Zaman, T.R., Herbrich, R., Van Gael, J., Stern, D.: Predicting information spreading in Twitter. In: Workshop on Computational Social Science and the Wisdom of Crowds, NIPS, vol. 104, pp. 17599–601. Citeseer (2010)