Topical Influence on Twitter: A Feature Construction Approach

Menno Luiten Walter A. Kosters Frank W. Takes

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

Abstract

In this paper we discuss the task of discovering topical influence within the online social network TWITTER. The main goal of this research is to discover who the influential users are with respect to a certain given topic. For this research we have sampled a portion of the TWITTER social graph, from which we have distilled topics and topical activity, and constructed a set of diverse features which we believe are useful in capturing the concept of topical influence. We will use several correlation and classification techniques to determine which features perform best with respect to the TWITTER network. Our findings support the claim that only looking at simple popularity features such as the number of followers is not enough to capture the concept of topical influence. It appears that more intricate features are required.

1 Introduction

The amount of information that is publicly available through the internet has drastically increased since the introduction of Web 2.0 [1]. Especially through online social networks [6], it has become extremely easy for users to share facts, opinions and news on any possible topic. When searching for information or news, we are confronted with a large number of information sources, from which we have to select what we believe to be correct and relevant content. Whereas before selecting sources of information was a matter of selecting certain websites, nowadays it is also a matter of selecting the correct users in a social network.

Within the online social network TWITTER [19], it is possible to follow users that are believed to produce relevant content. Such a user does not necessarily produce content which is relevant in general, but is more often only producing relevant content within a certain specific field of expertise. For example, Larry Page may be considered *influential* on the topic of internet search, but not on golf, whereas the opposite may hold for Tiger Woods. Selecting relevant users to follow on TWITTER is thus a matter of selecting users that produce relevant content on a certain *topic* (though we may ultimately be interested in multiple topics).

In this paper we will define features that can help us to determine who the influential (or authoritive) users on a certain topic are. We do this by analyzing the TWITTER social network, where we consider both the history of posted messages as well as a user's position in the social graph. Our goal is to better understand the concept of influence and to derive which characteristic features of users play a role when determining influence. In order to verify the performance of (combinations of) our features, we assume a definition of influence based on the sales funnel [3], as used by internet marketers. In this setting, a user is *influential* within the network if the links within the messages of a user are clicked on a lot by other users. As a second verification approach we consider the number of times a message has been "retweeted" by other users.

The motivation for doing this research is clear: it can help us to determine who we should definitely follow on TWITTER if we are interested in a certain topic. Also, having a list of influential users on a certain subject may be helpful to introduce new TWITTER users to build their list of people to follow based on a supplied list of interests. Additionally, it may help advertisers to select influential users who are likely able to successfully promote the advertiser's products or services. In this paper we will restrict ourselves to finding long-term authorities on a certain topic, as we will analyze multiple months of TWITTER messages.

The rest of this paper is organized as follows. First, we discuss some definitions, notations and assumptions in Section 2. After discussing related work in Section 3, we describe our sampling approach in Section 4. Next we consider a set of features for determining topical influence in Section 5, which we first filter based on effectiveness, and then apply to the TWITTER network in Section 6. Section 7 concludes.

2 Preliminaries

In this section we will first describe some concepts with respect to the TWITTER graph, after which we describe our main problem statement.

2.1 Twitter

We will be using the online social network graph $G(V, E)$ from TWITTER as the main dataset for our research. The edges (or links) E between the users (or nodes) V within the TWITTER social graph are, contrary to many other social networks, *directed*. When a user creates a link, a task which is commonly referred to as *following*, then this user can see all messages posted by the user to whom he created a link. This construct allows us to more accurately capture the real-life concept of influence as compared to a network consisting of only undirected links where it is not clear who is interested in whom. We use O_x to denote the outlinks, i.e., the set of users followed by user $x \in V$, and similarly we use I_x to denote the set of users that follow user x, representing x's inlinks.

Besides following, we will also mention several other concepts common to the TWITTER network. *Tweeting* is essentially posting a short 140-character message, referred to as a *tweet*. This message is not only visible on the profile of the originating user, but also in the *feed* of each user that follows this user. The set M_x denotes the set of messages sent by user x. A user's feed shows all messages posted by followed users. By *retweeting* we refer to a message being repeated by another user, allowing content to spread through the TWITTER network. We define the set R_m as the set of retweets of a message m. Retweeting happens for example because a user finds a message interesting and worth sharing with his followers. Referring to another user is called *mentioning*, denoted within a tweet by the symbol `@`, basically allowing users to direct messages to each other and have a conversation via TWITTER. In order to stress that a message is about a certain subject, so-called hashtags, denoted by the symbol `#`, are used. An example tweet, by user AEinstein, directed at `IsaacNewton` (a mention), asking about user `Apple` (a mention) with respect to the subject `#computerscience` (a hashtag), retweeted by user ScienceAcademy, is shown below:

> **AEinstein**
> `@IsaacNewton what do you think of the new @Apple`
> `product? http://bit.ly/12345 #computerscience`
> Retweeted by **ScienceAcademy**

2.2 Problem Statement

Our research focuses on the issue of determining *topical influence*. Influence, as defined by the Webster dictionary, is "the power or capacity of causing an effect in indirect or intangible ways". In our case, we will try to detect this capacity not on a global scale, but with respect to a certain topic:

Problem Statement. *Given a social graph of users, their connections, and posted messages, which user is most* ***influential*** *on a certain given* ***topic****?*

We try to answer this question by defining features which we believe describe the concept of influence. The question is then how we can measure whether or not our features are successful, which depends on our definition of when someone is influential.

Trivial ways of measuring global influence include looking at the total number of followers, or a user's position within the social graph. Furthermore, commercial websites such as Klout [13] develop metrics that have been suggested as measures of influence on TWITTER. Our definition of influence is based on the idea of the *sales funnel* [3], as used in internet marketing. This process, schematically outlined in Figure 1, traditionally describes the process of a visitor of a website from the moment he enters the website until a sale or some other action is completed. In our illustrated version of the sales funnel, social media is added prior to the visitor entering the website. The motivation for using the sales funnel is that one of the major questions in social media marketing searches for the strategy that most influences the *sales* of a company. It should however be noted that social media exposure also has strong advantages outside of the sales funnel such as brand exposure, creation of goodwill, community building and more. Our approach does not explicitly measure these benefits. We consider links in TWITTER messages as potential entry points to the sales funnel, and base our definition of influence on the number of incoming visitors in the sales funnel. We will thus consider the number of clicks on links present in TWITTER messages as a way of validating influence. As

a second validation measure of influence, we consider the number of retweets. Thus our two validation measures for determining the quality of our features, and therewith our definitions of influence, are:

Definition 1. *Influence within an online social network is the ability to generate clicks on posted links.*

Definition 2. *Influence within an online social network is the ability to generate retweets of posted messages.*

The relative value of these definitions can be inferred from their relative position in the sales funnel: the clicks of Definition 1 are closer to the end of the funnel than the retweets of Definition 2. Since we would ideally measure the effect on the end of the funnel, we value generated clicks over generated retweets.

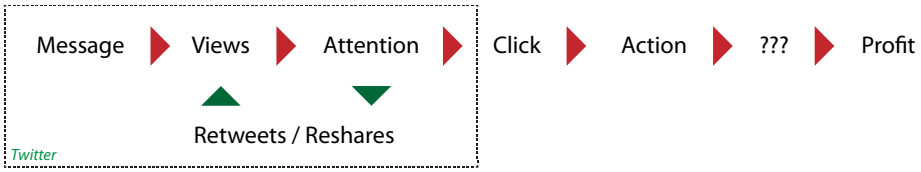

Figure 1: Schematic representation of the sales funnel.

3 Related Work

Trying to find the central nodes in a large network is a challenging task in the domain of data mining. Perhaps most notable is the work of Page and Brin, who introduced the well-known PageRank [14] measure for determining which nodes play an important role within the graph formed by the world wide web. Similar studies to find influential nodes have been done for social networks such as Flickr [7]. Unfortunately, the traditional PageRank-inspired measures only consider global influence, and do not take into account any topical information. Haveliwala introduced a topic-sensitive PageRank metric [10], which is applied to TWITTER through an algorithm called TwitterRank [20]. Here it is assumed that the influence of a user is a combination of the influence of his or her neighbors, and the relative amount of content of these neighbors.

With regard to influence measures, Cha et al. [5] empirically investigate the relation between common measures in influence on social media. However, when they test *topical* influence, they only take a small subset of users that have talked about all their defined topics. They find a strong correlation between topics, but this could be due to a selection bias towards generic TWITTER users, who have a tendency to talk about general topics. In other research on influence on TWITTER, specifically [2, 5, 10, 17], it is found that the traditional measures of follower counts and PageRank, while being good measures for popularity, are not as good at predicting influence when it is interpreted as the ability to engage one's audience. Romero et al. [17] include click data in their analysis, and find a weak correlation between clicks and popularity. It is found that retweets are mostly caused by a large group of "less-connected" users, instead of particular popular TWITTER users. Wu et al. [22] have a similar conclusion and suggest that the sociological theory of two-step communication flow [11] is still valid for electronic word-of-mouth networks.

4 Datasets

In this section we describe our sampling method for obtaining a large TWITTER dataset for our study of topical influence. The main approach to mine the TWITTER graph is based on the Forest Fire algorithm, which was found to be a reliable method of large graph sampling by Leskovec et al. [15]. This algorithm starts by randomly selecting a user in the graph (using a numeric identifier), and retrieves all of his or her connections and profile attributes. Next, it randomly selects ("burns") X random edges with probability 0.6, and recursively applies this step to these newly selected nodes. When the algorithm encounters an empty queue ("burns out"), it again selects a random user and repeats the process, until the required sample size is satisfied. We have used the TWITTER's REST API for our crawling activities. We ran the crawler several times for varying amounts of time, ultimately resulting in a data sample of over $30{,}000$ TWITTER users. Some indicative metrics on the size and shape of the data sample are shown in Table 1.

To be able to use the TWITTER graph in a topical context, we also retrieved up to $1{,}200$ tweets for each of the users in our sample, creating the set M_x of messages for each user x. These messages were then analyzed in order to define to which topic(s) they belong, allowing us to define M_{tx}, the set of messages

Property	Value
Nodes	$31,891$
Edges	$584,661$
Average Degree	18.3
Modularity	0.471
Density	0.001
Clustering Coefficient	0.068
Diameter	13
Average Path Length	4.03

Table 1: TWITTER dataset characteristics.

Topic	Keywords
Politics	democratic, republican, democrats, presidential, political, election, republicans, government, federal, constitution, executive, senators, elected, congressional, representatives, politics, presidents, perry, obama, biden, gingrich, romney, santorum
Tech	web, internet, www, html, computer, data, software, online, browser, oss, opensource, programmer, programming, developer, code, coding, java, c, c#, c++, php, visual basic, python, objective-c, perl, javascript, sql, ruby, haskell, perl, actionscript

Table 2: Topics along with associated keywords.

by user x on topic t. Taking into account a cut-off value of $|M_{tx}|/|M_x| \geq 0.005$ to disregard users who accidentally talked about a topic, we have generated sets $V_t \subseteq V$ of TWITTER users talking about topic t.

Our requirement for the definition of the topics was that they should be representative for a certain interest or a certain target group (e.g., politics, movies, technology, science). Using an automated topic distillation algorithm in the form of Latent Dirichlet Allocation (LDA) [4], we had no success in generating topics that complied with this requirement, as the topics more closely resembled random bags of words without a discernible theme. A similar undesired outcome was observed by [20]. Instead, we used a more simple technique of keyword matching, in which the keywords are based on the term frequency of a manually selected collection of Wikipedia articles surrounding a subject (e.g., American Politics, or Internet Technology). We emperically evaluated this distillation method to generate a more descriptive and complete set of topics compared to LDA and hashtag filtering. Throughout this paper, we use two topics in particular, namely "Politics" and "Tech", resulting in topic graphs of respectively $1,815$ and $3,109$ TWITTER users. Some keywords related to these topics are shown in Table 2.

In order to ultimately verify the influence of a user, we also gathered click data, as the number of clicks is going to serve as a measure of influence. We do this by unfolding `t.co` links that are present in TWITTER messages, and request click analytics from the ones that resolve to a `bit.ly` URL (see the example in Section 2.1). This way, we are able to retrieve a number of clicks for each link in a TWITTER message.

5 Features

In this section we will describe a list of features which we consider relevant with respect to topical influence, categorized based on the type of information that they use.

5.1 Graph-based features

Graph-based features solely consider the structure of the social graph, and are thus related only to a user x:

- The number of followers $|I_x|$ and the number of followed people $|O_x|$.
- PageRank $pr(x)$: the most prominent measure of importance on the web [14] .
- HITS authority $a(x)$ and hub $h(x)$ scores: an alternative measure of importance, also originally intented for the web [12].
- The 2-neighborhood $|N_2(x)|$: the size of the set of nodes at distance 2 from user x, extending the measure of followers by one step by counting the number of followers of followers.

5.2 Content-based features

Content-based features look at the message content, and are related to a user x and a topic t. We distinguish:

- The number of tweets by a user on a topic $|M_{tx}|$, describing the *activity* by user x on a topic t.
- Topical ratio $r(t, x) = |M_{tx}|/|M_x|$: the *relative* amount of activity of user x on a topic t, eliminating the effect of message (in)frequency.

- Term frequency-inverse document frequency $tfidf(t, x)$: similar to the topical ratio, but also considering the frequency of a keyword with respect to a certain topic.
- Number of mentions $m(x) = \sum_{v \in V} |\{m \in M_v : x \text{ is mentioned in } m\}|$. The number of times user x has been mentioned in the messages of other users can be an indication of popularity.
- Number of retweets $rt(x) = (1/|M_x|) * \sum_{m \in M_x} |R_m|$: this might indicate that a user or his content is popular.

5.3 Combined features

Considering both graph-based and content-based features, for a user x and a topic t, we can distinguish:

- Topic-sensitive PageRank $tpr(t, x)$: a PageRank measure that takes into account the topical ratio of the users [10].
- PageRank of a user x using only V_t, the set of users that talk about topic t, denoted $pr(t, x)$. This feature may indicate influence in a certain (topical) subset of users.
- Followers in the topic graph $ti(t, x) = |I_x \cap V_t|/|I_x|$: a high number of followers that also use the topic can indicate a topical clustering.
- Friends in the topic graph $to(t, x) = |O_x \cap V_t|/|O_x|$.
- Topical ratio of followers $fr(t, x) = (1/|I_x|) * \sum_{y \in I_x} r(t, y)$: the use of the topic by a user's followers can be indicative of a topical cluster.
- Average number of topical retweets $rt(t, x) = (1/|M_{tx}|) * \sum_{m \in M_{tx}} |R_m|$. This feature might not only indicate popularity on the topic, but also content value within the topical subset of users.

A more elaborate description of the features that we used can be found in [16].

5.4 Filtering

To determine which of the features are most relevant, we performed filtering by using Principal Component Analysis (PCA) [9] and Correlation-based Feature Selection (CfsSubsetEval) [8] from the popular data-mining software suite Weka [21]. These algorithms are designed to experiment with the feature space in order to to extract the features that explain variance of the features within the dataset.

The PCA approach showed that the strongest component that was found across topic graphs consisted of *popularity features* such as HITS authority score, PageRank, the number of followers and the neighborhood size. This indicates that a large part of the variance of the features might be explained by differences in popularity. A component that was less strong, yet still significant was a component that consisted mostly of topical ratio of followers, ratio of followers in topic graph, topical retweets, topic-sensitive PageRank, etc., which we will refer to as the *topical features*. We believe this component can be interpreted to be related to the topicality of the followers of the TWITTER user.

CfsSubsetEval, contrary to PCA, recognizes a *target variable* and attempts to find a subset of features of which the composite is highly correlated with the target feature, yet uncorrelated between the selected features themselves. When targeting the number of clicks, we found that the most important features are HITS hub score, ratio of followers in topic graph, topical ratio of followers and topical retweets, as can be seen in Table 3. In this table, *merit* denotes a heuristic of the (Pearson) correlation coefficient of the subset with the target variable. This indicates a certain importance of use of topicality by both the user and the followers of the user. Interestingly, popularity measures such as followers and PageRank are only found when the number of topical retweets $rt(t, x)$ is used as target, but not when the number of clicks $c(t, x)$ is used. Also, during our experiments, we noticed that removing the feature of average topical retweets resulted in a significant decrease in the correlation of the subset with the target feature of average clicks.

6 Experiments

Using the features found as a result of the filtering process in Section 5, we have tried to find classifiers that can explain the target features using the relevant features. As the source features for the classifiers we have used the two components, popularity features and topical features, as found in the PCA step from Section 5.4. We also used the relevant features found by CfsSubsetEval, namely:

Topic	Target	Merit	Selected attributes
Politics	$c(t,x)$	0.745	$ti(t,x)$ $h(x)$, $fr(t,x)$, $rt(t,x)$
Politics	$rt(t,x)$	0.360	$pr(x)$, $ti(t,x)$
Tech	$c(t,x)$	0.458	$h(x)$, $fr(t,x)$, $rt(t,x)$
Tech	$rt(t,x)$	0.454	$a(t,x)$, $pr(x)$, $ti(t,x)$, $rt(x)$

Table 3: Results of CfsSubsetEval on topics.

- Popularity features: authority score $a(x)$, hub score $h(x)$, global PageRank $pr(x)$, average number of retweets $rt(x)$ and average number of mentions $m(x)$.
- Topical features: ratio of topical followers $ti(t,x)$, follower ratio $fr(t,x)$ and average number of topical retweets $rt(t,x)$.

6.1 Classification

Now that we have extracted the relevant features, we are ready to start our process of classifying the target attribute in a way that can explain or even predict who the influential TWITTER users are. We will do this by classification of our two target attributes $c(t,x)$, the number of clicks on posted links, and $rt(t,x)$, the number of retweets as defined in Section 2.2. Our goal is to find a classifier that is not only accurate, but also easily interpretable and understandable. As a first step we have looked at naive Bayes classifiers and C4.5 decision trees [21]. We have discretized the number of clicks into four distinct categories (class 0 through 3, from no clicks at all, to a large number of clicks) and have used Cohen's kappa κ [9] as a measure of accuracy of the classifier. When the classifier finds (combinations of) features representative for certain classes of clicks, we can investigate the role of topicality of those features and interpret the classifier.

We trained classifiers on several topic graphs; the result of one topic can be seen in Table 4 ($\kappa = 0.4465$) and Table 5 ($\kappa = 0.238$). We noticed that only a few attributes have an increasing mean towards the higher classes of clicks, most prominently being average topical retweets, whereas most topical attributes have erratic, constant or even decreasing influence.

Because we suspected that even the filtered features were too detailed for the classification, we finally used a genetic algorithm [18] to find a combination of features that optimizes the kappa metric. While this approach may seem similar to PCA, it differs as it allows feature elimination in the classification attributes and uses the target variable for the accuracy of the model, simplifying the approach as a whole. We chose to use at least two features as a result of earlier findings from the PCA step, where we found a popularity and a topical feature set.

We again used the relevant features from Section 5 and combined them using linear weighting to generate the composite attributes, using 10-fold cross-validation to train and test the generated combinations. We experimented with adding attributes until the classifiers no longer improved their accuracy, which can be seen in Figure 2 and Figure 3. It can be observed that we only need to use a very limited number of attributes to optimize the classification of the model ($\kappa = 0.663$). Interestingly, it turned out that the features that were used by the algorithm consistently were various popularity features (mentions, retweets, HITS, PageRank), but only one topical feature, namely the number of topical retweets. It turned out that excluding

Attribute	Class 0	Class 1	Class 2	Class 3
$fr(t,x)$	0.0129	0.0179	0.0201	0.0071
$a(x)$	0.0001	0.0002	0.0008	0.0019
$h(x)$	0.0001	0.0001	0.0002	0.0003
$pr(x)$	0.0001	0.0002	0.0005	0.0014
$ti(t,x)$	0.3489	0.3908	0.3829	0.2513
$rt(x)$	0.0069	0.0229	0.0831	0.2753
$m(x)$	0.0147	0.0540	0.1421	0.3464
$rt(t,x)$	0.3294	6.5952	35.649	144.43

Table 4: Mean attribute values from topic "Politics" of the naive Bayes classifier.

Component	Class 0	Class 1	Class 2	Class 3
Popularity	-0.992	0.185	2.497	6.557
Topical	0.334	0.788	0.800	-0.831

Table 5: Mean principal component values from topic "Politics" of the naive Bayes classifier.

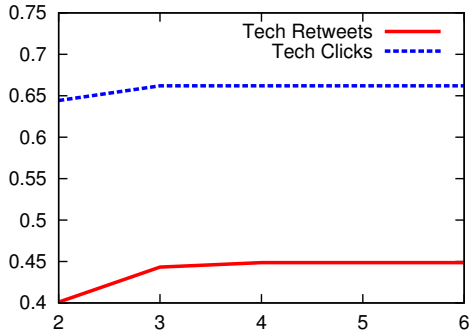

Figure 2: Kappa of best solution found (vertical axis) for increasing number of attributes (horizontal axis) on topic "Tech".

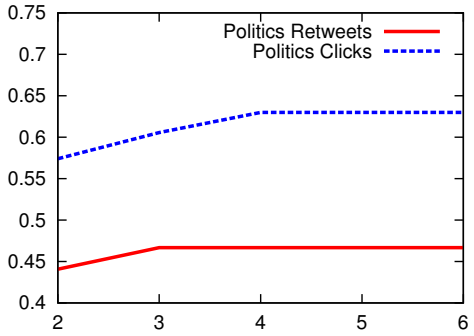

Figure 3: Kappa of best solution found (vertical axis) for increasing number of attributes (horizontal axis) on topic "Politics".

topical retweets had a significant impact on the accuracy of the model (maximum κ was reduced to 0.458), demonstrating the importance of topical retweets in the classification model.

6.2 Discussion

We believe that our observations are in accordance with earlier work. Bakshy et al. [2] found that the number of followers does not represent influence in the spreading of messages, and that large retweet cascades are originated mostly from many "less-connected" ordinary users. Our findings show that clicks correspond to high topical retweets, supporting the statement that popularity is only a secondary feature, whereas on-topic retweets is the most dominant primary feature. Romero et al. [17] state that influence is determined by activity of followers, instead of passive attributes such as number of followers. We confirm this observation by showing that topical retweets are an activity originating from followers, and not a passive metric such as the number of (topical) followers. Cha et al. [5] also suggest that number of followers are not the most important metric of influence in both a static as well as a changing environment. Instead they propose content value as a more superior metric. We believe topical retweets are an indication of content that fits well with the user's audience, which has been built over time, thus being a metric for both popularity, community and persistent content value.

7 Conclusion

Throughout this paper we have discussed various features that are useful in predicting topical influence on TWITTER. After a thorough investigation of which features contribute to predicting influence, we found two major classes of features: topical features and popularity features. Given our definition of influence based on the sales funnel, where the goal is to generate clicks on posted messages, the feature of topical retweets was found to be predominant in all classifiers. Apparently, when determining topical influence, it is most helpful to primarily investigate the interactions the user causes on his topical messages, especially regarding topical retweets. Our findings confirm earlier work which states that popularity features alone, such as the number of followers, are not sufficient to accurately capture the concept of influence.

In future work we would like to investigate if it possible to determine the extent to which a classification technique depends on the type of chosen topic. We are specifically interested in whether or not our approach works on short-term topics such as a specific soccer match or a local earthquake. Our current approach has been tested on various long-term topics and corresponding keywords, but it may very well be that when short-term topics are chosen, different classification techniques work better. We are also interested in how the influence of a user changes over time. Can we not only detect influential users, but also predict which user is going to become influential on a certain topic in the near future?

Acknowledgments

We thank Carlos Soares and Pedro Quelhas Brito at LIAAD, University of Porto, for their suggestions regarding this project. The third author is supported by the NWO COMPASS project (grant #612.065.92).

References

[1] A. Ankolekar, M. Krötzsch, T. Tran, and D. Vrandecic. The two cultures: Mashing up Web 2.0 and the semantic web. In *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pages 113–114, 2007.

[2] E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM '11)*, pages 65–74, 2011.

[3] T.E. Barry. The development of the hierarchy of effects: An historical perspective. *Current Issues & Research in Advertising*, 10(2):251–295, 1987.

[4] D. M. Blei, A. Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10)*, pages 10–17, 2010.

[6] M. Faloutsos, T. Karagiannis, and S. Moon. Online social networks. *IEEE Network*, 24(5):4–5, 2010.

[7] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pages 241–250, 2010.

[8] M. Hall. Correlation-based Feature Selection for Machine Learning. *PhD Thesis*, University of Waikato, 1998.

[9] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, second edition, 2009.

[10] T.H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:784–796, 2003.

[11] E. Katz and P. Lazarsfeld. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Free Press, 1955.

[12] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[13] Klout, Inc. Klout. `http://www.klout.com`, accessed June 1, 2012.

[14] A.N. Langville and C.D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[15] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pages 631–636, 2006.

[16] M. Luiten. Topical Influence on Twitter: A Feature Construction Approach. *Master Thesis*, Leiden University, 2012.

[17] D.M. Romero, W. Galuba, S. Asur, and B.A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*, pages 113–114, 2011.

[18] S.N. Sivanandam and S.N. Deepa. *Introduction to Genetic Algorithms*. Springer, 2007.

[19] Twitter, Inc. Twitter. `http://www.twitter.com`, accessed June 1, 2012.

[20] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential Twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pages 261–270, 2010.

[21] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, third edition, 2011.

[22] S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts. Who says what to whom on Twitter. In *Proceedings of the 20th International World Wide Web Conference (WWW '11)*, pages 705–714, 2011.