Finding Influential Authors in Brand-Page Communities

Hemant Purohit[*1], Jitendra Ajmera[2], Sachindra Joshi[2], Ashish Verma[2], Amit Sheth[1]

[1]Kno.e.sis, Dept. of Computer Science and Engineering, Wright State University, USA

[2]IBM Research, New Delhi, India

{hemant,amit}@knoesis.org , {jajmera1,jsachind,vashish}@in.ibm.com

Abstract

Enterprises are increasingly using social media forums to engage with their customer online- a phenomenon known as *Social Customer Relation Management (Social CRM)*. In this context, it is important for an enterprise to identify "influential authors" and engage with them on a priority basis. We present a study towards finding influential authors on Twitter forums where an implicit network based on user interactions is created and analyzed. Furthermore, author profile features and user interaction features are combined in a decision tree classification model for finding influential authors. A novel objective evaluation criterion is used for evaluating various features and modeling techniques. We compare our methods with other approaches that use either only the formal connections or only the author profile features and show a significant improvement in the classification accuracy over these baselines as well as over using Klout score.

Introduction

Social CRM is the use of social media services, techniques and technology to enable organizations to engage with their customers. Many enterprises have set-up their official web-pages (referred here as *brand-pages*) over Twitter and similar social media forums. People subscribed to or interacting with these pages form a community, referred here as the *brand-page community*. In the case of Twitter, members of this community interact with each other by means of *tweeting*, *retweeting*, *mentioning*, *replying* etc.

While these brand pages provide a good platform for enterprises to engage with their customers, there are some problems associated with this mode of engagement. Many authors do not always have a serious intent and some authors use these pages only for spamming. Moreover, since the customer identities are rarely known, it is difficult to prioritize authors based on their business value, reachability, previous logs etc. In this scenario, author priority for whom to engage with can be decided based on the "influence" an author exerts on the brand-page community. Finding influential authors in Twitter brand-page communities is our focus here.

An obvious approach to find influencers would be to analyze the static structure for a follower-network (referred here as explicit network) of the community members (Weng et al. 2010, Cha et al. 2010, Easley and Kleinberg 2010). However, as shown in this paper, these brand-page communities are very different from other topical communities on social media forums in that the community members do not formally connect (make *follower-friends*) to each other in general. In our dataset, only 0.01% of the possible formal connections were actually observed among the brand page community members.

Given this sparse nature of the explicit network in brand page communities, we explore the interactions among the community members for the purpose of finding influential authors. We propose following two methods:

(i) In our first proposed method we use author interactions to induce an "implicit network" and analyze this network using popular network analysis algorithms such as HITS (Kleinberg 1999) and PageRank (Brin and Page 1998).

(ii) In the second method a decision tree classification framework (DT framework) is employed to investigate and evaluate different combination of interaction and author profile features.

Being a subjective concept, it is difficult to first define and then evaluate the influence of a community member. In this study, we propose an objective evaluation criterion for this purpose. This objective criterion exploits the formal connections, wherever they are present, for the purpose of evaluation. Specifically, for an author pair {X, Y}, if X follows Y but Y does not follow X then Y is regarded as more influential than X.

We also present and compare our results with the explicit network analysis. However, **the primary goal of our work is to exploit features that can be extracted even when there are no formal connections present among the author population**. The experiments presented here suggest that:

1.) Author profile features such as the in-degree and activity alone cannot explain the influence of an author in brand-page community.
2.) Within graph analysis methods, HITS authority scoring outperforms PageRank scoring for both the explicit network as well as the implicit network. These

* Part of this research was conducted during author's internship at IBM Research, India.

methods also outperform the author profile features as mentioned above.

3.) Within the DT framework, we observer better performance when we combine interaction features with the author profile features as compared to any of the features in isolation.

4.) Our approaches also outperform ranking based on Klout score web service for finding influencers in brand-page communities as per the evaluation metric presented above.

Related Work

Among the most relevant work, (Weng et al. 2010) proposed TwitterRank technique to find topic sensitive influential authors on Twitter. TwitterRank is a variation of PageRank in that the weight of an edge (transition probability) depends not only on the explicit network transition but also on the topical similarity between the two nodes (authors). Since this largely depends on the explicit network connections, this technique cannot be applied to the problem at hand. They evaluated this approach on a friend recommendation task. In another study, (Pal and Counts 2011) proposed an approach to find topical authorities. Several features are proposed in this paper, which can be extracted and used for clustering authors based on their similarity of activity behavior. (Cha et al. 2010) presented a large-scale study on Twitter, which contradicts some of the conclusions made in the two above-mentioned works. The study explores three features for influence: In-degree, retweets and mentions. They presented these three notable observations: 1) In-degree reveals very little about the influence of an author. 2) Most influential authors exert influence across a variety of topics. 3) Influence is not gained spontaneously and activity plays a major role, which suggests for consideration of author activity in influence analysis. (Ghosh and Lerman 2010) argued that influence not only depends on the structure of the network, but also on the details of the dynamic processes occurring on it and therefore models accounting for these dynamic processes are better able to predict influential users on Digg social network. Our approach considers insights from past work and creates an integrated model of author interaction and author profile features that can explain influence dynamics in sparsely connected brand-page communities.

Methodology

Problem Statement: We consider the problem of influencer finding as that of a classification problem:

Given a pair of authors {X,Y} in brand-page community, determine who is more influential of the two authors.

Note that an influencer ranking of all the authors in a brand-page community can be derived based on such pair-wise classification.

Data Collection: Tweets and user profile data for the analysis presented in this paper were obtained using Twitter's Streaming and REST API services (*https://dev.twitter.com/docs*). A four-step crawling mechanism was used as follows:

1.) All the tweets containing relevant keywords for the brand and corresponding authors with their profiles were collected using the *filter-track* method of Streaming API.

2.) When these tweets were part of a discussion (a thread), all the tweets up to the root-tweet and corresponding authors with their profiles were also collected using the REST API.

3.) All the other authors who were part of some interaction in any of the tweets collected so far were also collected with their profiles.

4.) Additionally, for all the authors thus collected, 200 most recent tweets were collected regardless of the brand context. This was done to get richer interaction features among the users.

We crawled such data from the brand-pages of two enterprises in two totally different domains, automobile and retail electronics. The set of keywords used for finding tweets relevant to a brand included terms such as *@enterprise_name, #enterprise_name, enterprise_name* and other brand related terms. The datasets associated with these two brand-pages are referred as dataset1 and dataset2. Table 1 shows various statistics associated with these datasets.

Table 1: Table shows statistics around the datasets. A non-reciprocal edge is a relation between two authors {X, Y} where X follows Y but Y does not follow X. The fraction of total possible follower edges being very small shows that the explicit network around the authors of interest (U) is indeed very sparse.

STATS	Dataset1	Dataset2
Crawl Duration (year 2011)	Nov 27 - Dec 17	Dec 27 - Dec 30
Number of tweets by users	2325650	3262834
Number of tweets related to brand	26899	37890
Number of authors or nodes (N)	13875	18713
Total follower-friend edges among N authors	100137	61076
Number of isolated Authors (no follower-friend connection)	5297	6285
Fraction of total possible follower edges	.05%	.017%
Number of non-reciprocal edges (E)	48604	45161

Table 1 shows higher activity for the dataset2 users as compared to the dataset1 users that is likely causing better performance in the case of dataset2 as shown later in Table 2. Table 1 also shows that the explicit network consisting of formal follower-friend connection is indeed sparse. There are only .05% and .017% of the total possible edges present in the data.

Evaluation Methodology: We use the *follower-friend* relationship among the authors population. Specifically, for a user pair {X, Y}, if X follows Y but Y does not follow X then we regard Y to be more influential than X. Since this information is not available for all the author-pairs, we only evaluate it on the pairs where it is available. We refer to this set of author pairs as evaluation set. Furthermore, we removed corporate authorized authors (e.g. characterized by *@enterprise_name*) for a fair analysis.

Approaches for finding influential authors

This section describes following two approaches for the purpose of finding influential authors:

1. Implicit network analysis
2. Decision tree classification model

Implicit Network Analysis: In this approach, we first create an implicit network by extracting interactions serving as implicit links between authors. We exploit three interaction properties of Twitter for this purpose: *Retweet, Reply* and *Mention*. A directed edge from user node X to user node Y is created iff: 1.) X *retweets* Y's tweet, or 2.) X *replies* to Y's tweet or 3.) X *mentions* Y in one of its tweets.

To penalize those *followers* who have a tendency to retweet (or reply to) everything or those who tend to follow a lot of friends, we apply a TF-IDF (term frequency – inverse document frequency) like normalization. The weight of a directed edge from author X to author Y (W_{XY}) is computed as follows:

$$W_{XY} = (RT_{XY}+RP_{XY}+MN_{XY}) \cdot \log(N/(A_{RT}+A_{RP}+A_{MN}))$$

Here, $A_{\{z\}}$ is the number of unique other users with whom X has interacted using property z where z can be retweet *(RT)*, reply *(RP)* or mention *(MN)*. N is the total number of users except X. RT_{XY} is the number of retweet interactions from X to Y and similarly for other properties. As explained earlier, the term within the logarithm will penalize those users who act on a lot of tweets from a lot of different authors.

Once the weights of the directed implicit network have been obtained in this manner, we apply popular link analysis algorithms, HITS and PageRank on the implicit network. Table 2 presents the classification accuracy for both these algorithms when applied to such network.

In this implicit network analysis presented above, we have only used the interaction links. However, as shown in Table 2, author profile features such as the number of followers for each user and the number of brand-related tweets written by a user in a given time-frame are also important features for characterizing the influence of a user. Since the author profile features involve activity (edges) and authors (nodes) outside the brand page community, it becomes inefficient to combine these features with the interaction links in the form of a network. Therefore, we used a decision tree based classification approach to investigate the combination of these features. This approach is presented in the next section.

Decision tree classification Model: We used the author profile features and the implicit network features in a Decision tree (DT) classification model. However, unlike PageRank and HITS approaches mentioned above, DT based classification is a supervised process. DTs have to be trained using some labeled data. For this purpose, 70% of the total number of non-reciprocal edges-set of explicit follower network, set (E) in Table 1, was used for training and 30% of the edges were used for evaluation. For a fair comparison, the classification accuracy for all the methods including the unsupervised PageRank and HITS are presented on the same 30% evaluation subset in Table 2. We investigated several combinations of the following features for the purpose of building and testing the DTs:

a.) User Interactions features: *Number of times a user got retweeted (N_{RT}), Number of times a user got replied (N_{RP}), Number of times a user got mentioned (N_{MN})*

b.) Author profile features: *Number of followers of a user on Twitter (N_f), Number of brand-related tweets of a user in a given timeframe (N_a)*

c.) Implicit network based features: *Authority and Hub score (result of implicit network analysis)*

Since the input to the DT model is a pair of authors, the features mentioned above were extracted for both the authors and used in the order of the authors in the input pair. Next section presents the classification accuracy for various combinations of these features.

Experiments and Result analysis

We perform implicit network analysis on the two datasets mentioned above as well as DT framework analysis and report the evaluation results for various combinations of features in this section. However, in order to compare the performance of the two proposed approaches, we first present and explain two baseline set-ups.

Author profile feature baseline: Motivated by some previous work on influencer finding, e.g. (Cha et al. 2010), we extract following three author profile features for each user in our author population: 1) Number of followers on Twitter 2.) Activity or number of tweets of a user in the brand-page community in a given timeframe 3.) Klout score (*http://klout.com*), a popular social network influence metric on the web, available through API. For all these features, if the feature value for a user X is higher than that of Y, then X is considered more influential than Y.

Explicit network analysis baseline: As explained earlier, the explicit network is a network formed by considering the formal follower-friend relationships in the brand-page community. It has already been shown that such network is very sparse in Table 1. We present the results of analyzing such network using the PageRank and HITS techniques (in exactly the same manner of analyzing implicit network) as baseline results. However, two points must be noted:

(1) This analysis can only be done on the formally connected edges and thus the recall of this method is going to be very poor.

(2) Since the evaluation criterion has been derived based on an explicit network property, the results of such analysis are going to be biased in favor of this baseline.

Result Analysis: Table 2 presents the classification accuracies for 30% of the non-reciprocal edge set (E). There are 14582 and 13549 such author-pairs to be evaluated in the two company datasets, respectively. As mentioned earlier, these test points do not contain any of

the brand-page representative authors. We observe the following insights from Table 2:

(1.) The *explicit network analysis baseline, especially the authority score, performs significantly better than the author profile features* baselines, especially for the second dataset.

(2.) The *activity feature does not correlate at all with the influence of authors*. This supports our observation that many users on such forums do not have serious intent and some use these forums for spamming only.

(3.) The authority score, a result of HITS analysis, performs better than the PageRank score, as observed in both the explicit and implicit networks. This suggests that the *HITS carefully exploits and discriminates between the out-links and in-links whereas PageRank does not*.

Table 2. Classification accuracy for baseline and proposed approaches

SET-UP / FEATURE	Dataset1	Dataset2
Author Profile Feature baseline		
Number of followers	79.8	83.5
Activity	36.3	16.7
Klout score	68.0	84.0
Explicit Network Analysis baseline		
PageRank	77.0	90.5
Authority score (HITS)	77.3	92.6
Hub score	70.2	82.6
Implicit Network Analysis		
PageRank	76.6	87.6
Authority score	80.0	91.4
Hub score	72.1	73.9
Decision-Tree based Analysis		
3 interaction features (N_{RT} , N_{RP} , N_{MN})	79.0	90.0
Only first 2 author profile features (N_f , N_a)	81.2	94.7
Interaction and author profile features (N_{RT} , N_{RP} , N_{MN} , N_f , N_a)	**85.1**	**94.7**
Interaction, author profile features and authority score (N_{RT} , N_{RP} , N_{MN} , N_f , N_a , authority_score)	85.0	94.7

(4.) *The implicit network analysis results are comparable to the explicit network analysis results*. In fact, the authority scoring in implicit network analysis works better than that in the explicit network, especially for dataset1.

(5.) The decision tree approaches clearly *gains from the combination of the interaction features and the author profile features*.

(6.) As a separate exercise to evaluate the goodness of the influence ranking (as opposed to classification accuracy), we compute the correlation coefficient between the influence rankings from implicit and explicit networks. The coefficients for the authority score (HITS output) are .82 and .63 for the two datasets, while for the PageRank score, they are .85 and .42. This further supports our argument that the *implicit network emulates the explicit network*.

Conclusion

This paper presented two approaches for finding influential authors in brand-page communities on Twitter. Since the follower network in these communities is too sparse to be analyzed by the standard network analysis approaches, we exploited interactions among the community members and author profile features for this purpose. An implicit network based on the interactions among authors was created and analyzed using graph analysis techniques. Furthermore, we combined author profile features with the interaction features in a decision tree classification model. A novel objective evaluation criterion was used to evaluate and compare different approaches and features. Our study suggests that user interactions can be used for analyzing influence in evolving communities where the follower network is very sparse, such as during emergency response. We plan to extend this work by integrating deeper content analysis of tweets such as the style, the vocabulary, sentiment and the type of tweets.

Authors would like to acknowledge colleagues at Kno.e.sis & IBM IRL, and reviewers for their useful comments.

References

Weng, J.; Lim, E.P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining (WSDM)*. 261-270.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring User Influence in Twitter : The Million Follower Fallacy. *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Pal, A.; and Counts, S. 2011. Identifying Topical Authorities in Microblogs. *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*, p. 45–54.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*. 65-74.

Easley D.; and Kleinberg, J. eds. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Ghosh, R.; and Lerman, K. 2010. Predicting Influential Users in Online Social Networks. *In Proceedings of KDD workshop on Social Network Analysis (SNAKDD)*.

Lee, C.; Kwak, H.; Park, H.; and Moon, S. 2010. Finding influentials from temporal order of information adoption in twitter. *In Proceedings of the 19th international conference on World wide web (WWW)*. 1137-1138.

Brin, S.; and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *Elsevier*, 30(1-7), 107-117.

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