Pests Hidden in Your Fans: An Effective Approach for Opinion Leader Discovery

Binyang Li, Kam-fai Wong, Lanjun Zhou, Zhongyu Wei, Jun Xu

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Key Laboratory of High Confidence Software Technologies
Ministry of Education, China
{byli,kfwong,ljzhou,zywei,jxu}@se.cuhk.edu.hk

Abstract.

With the development of Web 2.0, people would like to share opinions on the Web, which are very helpful for other users to make decisions. Especially, some users have more powerful influence to other members of a community, group, or society, and their advice, opinions, and views are more valuable. We call these people opinion leaders. The study of opinion leader discovery from the social media is meaningful because it could help users to understand influential user behavior, and trace vital information diffusion of an e-society, even on-line ecology. However, existing approaches focus on linkage-based methods without considering the *pests* who have relationship with the potential opinion leader but carrying opposite opinions. In an extreme case, an opinion leader might be mistakenly identified according to his richer relationships with the *pests*. In this paper, we start from explaining the definition of opinion leader, and take into consideration of the user profile and posts' opinions instead of using structural information (linkage) only. As such, those *pests* carrying opposite opinions could be gotten rid of from the social network, which could further improve the effectiveness of discovering opinion leaders. To evaluate the performance of our approach, we made experiments based on the Tweets data, and the results showed that our proposed approach could achieve 8% improvement compared with the linkage-based approach.

Keywords: Opinion leader, social network, Web 2.0, pests.

adfa, p. 1, 2011.

1 Introduction

With the development of Web 2.0, people would like to share opinions on the Web, which are very helpful for other users to make decision. Some of the users could always provide more valuable advice, and they gradually have more powerful influence to other members in the specific e-community, and further to the whole e-society. For example, *Gangnam Style* unusually drew the attention from the world in the past few months. Till November 2012, its MV uploaded in *Youtube* has been watched almost 8.4 hundred million times. How did *Gangnam Style* turn into a success? To trace backward, we found that opinion leaders gave most of the effort on this. At first, the click-through rate was poor. But then, some artists who have strong calling power, like *Britney Spears*, *Katy Perry*, promoted it keenly in Twitter. Influenced by them, artists from other areas shared it, as a result, the trend started to spread among the whole world. *Gangnam Style* became famous incredibly. It is clear that the opinion leader is the key of success.

The formal definition of opinion leader is that people who are influential members of a community, group, or society to whom others turn for advice, opinions, and views. A user is considered as an opinion leader when he involves the following factors[1]:

- Expression of values
- Professional competence
- Nature of his social network

Due to the commercial factors of opinion leader, the study of opinion leader focuses on mining the commercial value of a given opinion leader, such as product suggestion, advertising campaign, and so on, while ignore how to discover opinion leaders. According to our preliminary study, there exists some related work on opinion leader identification, e.g. Central Policy Unit.[1] Unfortunately, most of the opinion leader analysis utilizes manual methods for discovering opinion leaders, rather than carrying out automatically. Manual methods may be work in Web 1.0 era with limited portal websites, but hardly continue when entering big data.[2] Besides, the research on automatically opinion leader discovery takes only the nature of the social network into consideration, e.g. PageRank, HITS like models.[10][11] However, not all the relationships between users express supporting opinions. For instance, "*Sister Phoenix*"(凤姐) has a great amount of fans and relations on *Weibo*, but most of her fans oppose her. It is very often that the relationship between an opinion leader and a non-opinion leader carries the opposite opinion. Therefore, we argue that distinguish pests from your fans or replies will improve the performance of opinion leader identification.

In this paper, we target to identify the opinion leaders on social media. We discover an opinion leader by accounting for all the three factors in the definition of opinion leaders. We present a 2-step approach for discovering opinion leaders: we first generate a group of candidate/potential opinion leaders according to the expression of val-

[1] http://en.wikipedia.org/wiki/Opinion_leader

ues and professional competence by analyzing the profile and the post content of the users. Specifically, we propose two categories of features to (1) describe the user profile which will help detect the professional area and measure his/her competence; (2) analyze the content of his/her posts together with the replies and the comments which will help determine the value of the expression, e.g., get rid of the pests expressing opposite opinions to refine the social network. We then integrate the potential opinion leaders into a graph-based model to measure its nature of social networks.

In order to investigate the performance of our proposed method, we also conduct several experiments based on the real data, which was collected from Twitter about the UK General Election in 2010[3]. We investigate the contribution of different features, including content-based, user-based and linkage-based for identifying opinion leaders. A comparative experiment is conducted and the experimental results showed that our proposed approach outperform the state-of-the-art linkage based method.

The rest of this paper is organized as follows: we will review the related works in Section 2. In Section 3, we will present our 2-step approach for opinion leader discovery. We evaluate our approach in Section 4. Finally, we will conduct the conclusion and suggest future works in Section 5.

2 Related Work

Most of the previously work about opinion leader focused on how to utilize its commercial value, such as marketing research, product sampling, retailing/personal selling, and advertising.[4][5][6] For this kind of research, the opinion leaders were predetermined manually, and the data was not open to public.

Until the recent decade, with the explosion of information, automatically discovering opinion leader attracted more and more attentions.

Thomas, et al., proposed analytic hierarchy process (AHP) method, which is a structured technique for organizing and analyzing complex decisions.[7] According to each shortlisted user, assorted factors should be considered. Thus, the AHP provides a comprehensive and rational framework for quantifying its factors in order to relate those factors and evaluate the solutions, which mean the choices of opinion leaders in this case. An AHP hierarchy consists of an overall goal, a group of alternatives for reaching the goal, and a group of related factors. The factors can be further broken down into many levels as the problem requires.

Laclavik, et al., proposed a method for opinion leader identification based on the relationship network.[8] They suggested to determine the communication relationships between users by relationship mining methods. The extracted users and its relationship network formed a social network could be then represented as graphs. The resulting graph was analyzed by determining key figures for the position of single users and for the overall structure of the network. In this way, opinion leaders could be identified.

Centrality analysis approach was proposed to measure the degree of activist's connection with others.[9] The more the connections were, the more the influences of that activist on others. Opinion leaders could be listed out based on the degree of influence

and activeness. Social network analysis provided a number of key features which described the structure of the entire network. For analyzing opinions, the key figures density, connectivity, and closeness centralization were especially relevant. Density measured the connection of a network and is an indicator for communication within the network.

In summary, all of the above approaches only focus on analyzing the relationship between the users. They make use of the concept of tree structure or graph stricter to illustrate the influence brought by the opinion leaders. They concern the levels, width and size which can show the structural information of opinion leaders. However, structure-based methods only show the number of people replied or retweeted his or her posts but not going to consider the content of the tweets, i.e. opinions. This kind of content of a tweet can illustrate whether it supports the idea of the previous posts or whether they are talking about the same topic. If we find that the repliers do not agree with the viewpoints of the previous tweet, then this tweet cannot be counted as an effect of influence in his social network. Therefore, in this paper, we will account for the content-based features for opinion leader discovery.

3 Methodology

In this section, we will present our 2-step approach for opinion leader discovery. It is intuitively that the best way of identify the *website leader* is based on the analysis of relationships of websites where the link between two websites only carry "supporting" meaning. However, for opinion leader discovery from social media, the link between two users may indicate an "opposite" meaning or "none". In other words, the traditional meaning of the linkage cannot tell the whole story, and there exists pests hidden in the relations carry opposite opinions. Therefore, we suggest to identify the opinion leaders by considering "opinion" and propose a 2-step approach where a candidate opinion leader set is firstly generated by considering the factor of sentiment and then implemented into a graph-based model for final identification.

3.1 Potential Opinion Leader Generation

Recall that a user is considered as an opinion leader when he/she involves the following factors: expression of values, professional competence, and nature of its social network. We propose a number of features to describe the first two factors in this subsection and put them into SVM[12] to generate a candidate opinion leader set. Then we implement the results into graph-based model for analyzing the nature of its social network to discover opinion leaders.

- Professional Competence

Since our target is to discover opinion leader from social media, we take Twitter as an example for further description. We list some statistics related to the users' profession from his/her profile. This category of features represents the nature of user's professional competence shown in Table 1.

Table 1. The descriptions of user-based features.

Name	Description
Tweet Count	If the post contains more comments, it means that the scope of influence by the post increases. Also, when one user's posts are being commented for more times, it proves that the user might have a greater influence.
Follower Count	This simply counts the number of followers of each user in order to measure his credibility/authority.
Verified	Verification is currently used to establish authenticity of identities on Twitter, which could improve the authority of the user.
Retweet Count	The value of the user's professional degree can be reflected by the count of the retweeting the post.
Reply Count	The more number of reply, the higher attention from the other users have paid on the topic.
Retweet time range	To investigate the time of validity of a tweet.

- Value of Expression

In order to measure the expression of values, we analyze the content of the post together with its comments/replies. Accordingly, the category of content-based features is proposed to help us find out the viewpoint and argument of a comment as shown in Table 2.

Table 2. The descriptions of content-based features.

Name	Description
Pos. Ratio of Reply	The pos. ratio tells us the percentage of reply having the same attitude towards the same side is. The higher the ratio is, the more the people agree with his or her viewpoints.
Cons. Ratio of Reply	The cons. ratio tells us the percentage of reply having the opposite attitude towards the author's side is. The higher the ratio is, the more the people disagree with his or her viewpoints.
Sentiment Degree	It calculates the strength of the sentiment words within the tweet or reply.
Words Count	How many words are included in a tweet?
Positive Words Count	How many positive words are included in a tweet, which could help us to understand how strong its attitude is expressed?
Negative Words Count	How many negative words are included in a tweet, which could help us to understand how strong its attitude is expressed?
Hashtag	The Hashtag is used to mark keywords or topics in a Tweet.

To understand more about the attitude of users towards their opinion leaders, the reply's content is important. Does the user support or oppose the viewpoint of the opinion leader? In our method, we utilize *Sentiwordnet*[2] to help us analyze the opinion of the tweet. *Sentiwordnet* is an open source which consists of 17,370 negative words and 18,157 positive words. There is a score assigned to each individual sentiment word to indicate the positive or the negative strength of the sentiment word.[16] (In this paper, the supporting opinion and opposite opinion are also referred as positive opinion and negative opinion, respectively.)

For simplicity, we just measure the opinionatedness in two naive ways: calculating the scores of the sentiment words appearing in the tweet; counting the number of sentiment words. Then we sum up the score or the sentiment word count of a tweet to indicate the attitude (positive or negative).

We then put both categories of features into *Supporting Vector Machine* (SVM) to generate candidate opinion leaders.

3.2 Opinion Leader Identification

After we generate the candidate opinion leaders, we then integrate them into a graph-based model to rank and generate the final opinion leaders.

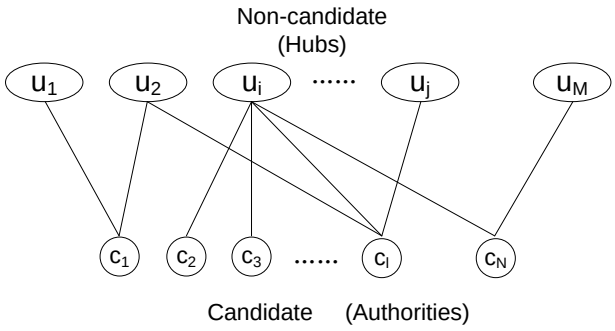

Fig. 1. Graph model for opinion leader identification.

Our proposed graph model is based on HITS algorithm, which distinguishes the users into hubs and authorities. An authority represents a candidate opinion leader, noted by *c*, while a hub indicates the non-candidate user, noted by *u*. For each individual u_i, it has links to many authorities. An authority c_j would have many hubs linking to it. The hub scores and authority scores are computed in an iterative way. Fig. 1 gives the graph model representation of the HITS model.

For our purpose, the non-candidate layer is considered as hubs and the candidate layer authorities. If a non-candidate user posts a reply or comment to support the opinion of candidate opinion leader, there will be an edge between them. In Fig. 1, we can see that the candidate opinion leader that has links from many non-candidate us-

[2] http://sentiwordnet.isti.cnr.it/

ers can be assigned a high weight to denote a strong social network. On the contrary, if a candidate opinion leader has few links from the Hubs, the score is low, which will result in a low ranking. Each edge is associated with a weight w_{ij} denoting the contribution of u_i to the candidate opinion leader c_j. The weight w_{ij} is computed by the contribution of non-candidate users.

Different from existing approaches, we consider sentiment factor for opinion leader discovery. We divide the links between users into supporting (positive) ones and opposite (negative) ones, and regard those positive links are valuable in its social network. Therefore, we filter out the pests from follower with links expressing opposite opinions more than supporting ones in the first step. We then compute the weight of the edge by only accounting for positive links.

For computation of the final scores, the initial scores of all candidate opinion leaders are set to *1/N*, and non-candidate are set to *1/M*. The above iterative steps are then used to compute the new scores until convergence. Usually the convergence of the iteration algorithm is achieved when the difference between the scores computed at two successive iterations for any nodes falls below a given threshold [13][14][17]. In our model, we use the authority scores as the total scores. The opinion leaders are then ranked based on the total scores.

4 Experiment

4.1 Experimental Setup

Dataset

The tweets data we used in this paper were collected using the Twitter Streaming API for 8 weeks leading to the UK General Election in 2010[3]. Search criteria specified include the mention of political parties such as Labour, Conservative, Tory1, etc., the mention of candidates such as Brown, Cameron, Clegg, etc., the use of the hash tags such as #election2010, #Labour etc., and the use of certain words such as election. The corpus contains around 919,662 unique tweets and 68,620 users. We also collected the following links of all the users.

According to the Twitter setting, data can only be revealed at most one-week quantity each time. Thus, data sample will be separated randomly into one-week sized in this case for afterward analysis in order to fit the circumstance, which data is extracted automatically in the future. Thus, the dataset for analysis contains around 44,391 unique tweets and 18,713 users.

Annotation

We have also done the clustering for the data before the annotation. There are 4 subgroups such as conservative party, labour party, liberal democrat party and others. For each subgroup, we have annotated the opinion leaders in the subgroup manually while two out of three members in our group agree that the user is opinion leader, which mean more than half of the group members agree, then the user is opinion leader. And it is the majority rule to identify the opinion leaders manually. The Kappa coefficient[18] indicating inter-annotator agreement was 0.8236 for the binary

classification. The conflict labels from the two annotators were resolved by a third annotator. Finally, there are 129 opinion leaders annotated in the training dataset.

In our experiment, the data is divided into five folds and four of them are training data and the one left is testing data. In order to investigate the performance of 2-step approach, we compare our proposed method with the linkage-based approach, which achieved the best run. Beside, we have proposed several categories of features to identify potential opinion leaders in Section 3, which are content-based and user-based mentioned above. In order to investigate the effectiveness of each category of features, we also tried different combination of feature sets.

Baseline

We choose linkage-based method as the baseline, which achieved best performance among linkage-based methods[15].

Metrics

We utilize *precision recall* and *f-value* as our evaluation metrics. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The recall is intuitively the ability of the classifier to find all the positive samples. The F-value can be interpreted as a weighted harmonic mean of the precision and recall. A measure reaches its best value at 1 and worst score at 0.[3]

4.2 Experimental Result

The overall performance of different approaches is shown in Table 3. We use SVM+C to denote Content-based feature, SVM+U means user-based feature, and SVM+ALL stands for adding them all.

The experimental results showed that the 2-step approach with our proposed features outperformed the baseline. Especially, when all the features were taken into consideration, the F-value was the highest which identified 106 opinion leaders and achieved around 8% improvement over the baseline. Moreover, there more than 10 fake opinion leaders identified by the baseline due to their well social network but mainly negative comments.

We further compared the features in different categories, the content-based features were more important than user-based features, which could identify 94 opinion leaders.

Table 3. Comparison between different approaches for opinion leader discovery.

Approach	Precision	Recall	F-value
Baseline	0.7322	0.6285	0.6763
SVM+C	0.8032	0.6345	0.7089
SVM+U	0.7740	0.6316	0.6955
SVM+ALL	**0.8180**	**0.6351**	**0.7150**

[3] http://scikit-learn.org/dev/modules/model_evaluation.html

We further investigate the performance of each individual feature of content-based category, and the results were shown in Table 4.

It is clear that SVM+ratio achieved the best run, which demonstrated that the feature of pos. or cons. ratio was the most effective way to decrease the impact of relationship carrying negative opinion comparing with the baseline. Besides, according to our analysis of the experimental results, one would like to express opinions to attack others during the Election rather than post replies for supporting. As a result, if a candidate opinion leader has high pos. ratio, it is probable to be an opinion leader.

Table 4. Comparison between different features for opinion leader discovery.

Approach	Precision	Recall	F-value
SVM+Pos./Cons. ratio of reply	**0.8089**	**0.6317**	**0.7094**
SVM+Sentiment word count	0.7461	0.6234	0.6792
SVM+Sentiment degree	0.7475	0.6267	0.6817
SVM+Word count	0.7418	0.6182	0.6685
SVM+Hashtag	0.7122	0.6085	0.6563

5 Conclusion and Future Work

5.1 Conclusion

This paper targets to identify opinion leaders on the social media. The main difference from traditional *website leader* is that the link of social network doesn't always mean supporting. A pest link is likely to exist with negative opinions between users. We, therefore, design a 2-step model by taking into consideration of the key factors of opinion leaders, the value of expression, the professional competence, and the nature of social networks.

Specifically, a candidate opinion leader set is generated by utilizing the user profile to detect the professional area and measure user's competence; by analyzing the content of user's posts to determine the value of the expression in the first step. In the second step, a HITS-like graph is constructed based on the potential opinion leaders to rank the opinion leaders.

In conclusion,

1. We propose a set of useful features to describe the key factors of opinion leaders, which is proved to be effective for candidate opinion leader generation;
2. A graph-based model is devised for ranking opinion leaders, which decreases the impact of links with negative opinions;
3. A 2-step approach for opinion leader identification is presented from the perspective of view of opinion leader definition;
4. Several experiments were conducted and the results showed the effectiveness of our proposed 2-step approach, which could achieve 8% improvement over the baseline.

5.2 Future Work

In the future, we will continue our research on opinion leader discovery in the following directions:

1. Develop a unified model for opinion leader discovery by considering information diffusion;
2. Implement the fine-grained opinion analysis into content analysis, e.g. opinion target identification[19];
3. Classify the comments into different categories so as to build up the relationship between comments.
4. Besides Tweets, we would like to move forward to other data from different languages of social media, like Weibo, My Space, etc.

Acknowledgments

This work is partially supported by National 863 program of China (Grant No. 2009AA01Z150), General Research Fund of Hong Kong Research Grants Council (Project No. 417112), and CUHK Direct Grants (No. 2050525). We also thank Xu Han and anonymous reviewers for their helpful comments.

References:

1. Lewis J. The Search for Coordination: The Case of the Central Policy Review Staff and Social Policy Planning, 1971–77[J]. Social Policy & Administration, 2011, 45(7): 770-787.
2. Asur S, Huberman B A. Predicting the future with social media[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010, 1: 492-499.
3. He Y, Saif H, Wei Z, et al. Quantising Opinons for Political Tweets Analysis[C]//Proceeding of the The eighth international conference on Language Resources and Evaluation (LREC)-In Submission. 2012.
4. Wang J C, Chen C L. An automated tool for managing interactions in virtual communities-using social network analysis approach[J]. Journal of Organizational Computing and Electronic Commerce, 2004, 14(1): 1-26.
5. R. van der Merwe & G. van Heerden (2009), "Finding and utilizing opinion leaders: Social networks and the power of relationships", Division of Industrial Marketing, eCommerce and Supply Chain Management, Luleå University of Technology, Luleå, Sweden
6. http://www.opinionleader.co.uk/
7. Saaty, Thomas L.; Peniwati, Kirti (2008). *Group Decision Making: Drawing out and Reconciling Differences*. Pittsburgh, Pennsylvania: RWS Publications. ISBN 978-1-888603-08-8.
8. Xiaofei Zhang and Dahai Dong. Way of Identifying the Opinion Leaders in Virtual Communities, July 2008
9. Laclavík M, Dlugolinský Š, Šeleng M, et al. Email analysis and information extraction for enterprise benefit[J]. Computing and informatics, 2012, 30(1): 57-87.

10. Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.
11. Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM, 46(5)*: 604-632.
12. Cortes C, Vapnik V. Support vector machine[J]. Machine learning, 1995, 20(3): 273-297.
13. Binyang Li, Lanjun Zhou, Shi Feng, Kam-Fai Wong. A Unified Graph Model for Sentence-based Opinion Retrieval. In Proceedings of ACL 2010.
14. Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In SIGIR '08: Proceedings of the 31th annual international ACM SIGIR conference on Research and development in information retrieval, pages 299-306. ACM.
15. Cho Y, Hwang J, Lee D. Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach[J]. Technological Forecasting and Social Change, 2012, 79(1): 97-106.
16. Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis[J]. Computational linguistics, 2011, 37(2): 267-307.
17. Erkan G, Radev D R. LexRank: Graph-based lexical centrality as salience in text summarization[J]. J. Artif. Intell. Res. (JAIR), 2004, 22: 457-479.
18. Jean Carletta. 1996. Assessing agreement on classi-fication tasks: the kappa statistic. Computational linguistics, 22(2):249–254.
19. Lanjun Zhou**,** Yunqing Xia, Binyang Li and Kam-Fai Wong, WIA-Opinmine System in NTCIR-8 MOAT Evaluation, in NTCIR-8 Workshop Meeting, 2010.