# Identification of Opinion Leaders Using Text Mining Technique in Virtual Community

**Chihli Hung**
Department of Information Management
Chung Yuan Christian University
Taiwan 32023, R.O.C.
chihli@cycu.edu.tw

**Pei-Wen Yeh**
Department of Information Management
Chung Yuan Christian University
Taiwan 32023, R.O.C.
mogufly@gmail.com

## Abstract

Word of mouth (WOM) affects the buying behavior of information receivers stronger than advertisements. Opinion leaders further affect others in a specific domain through their new information, ideas and opinions. Identification of opinion leaders has become one of the most important tasks in the field of WOM mining. Existing work to find opinion leaders is based mainly on quantitative approaches, such as social network analysis and involvement. Opinion leaders often post knowledgeable and useful documents. Thus, the contents of WOM are useful to mine opinion leaders as well. This research proposes a text mining-based approach to evaluate features of expertise, novelty and richness of information from contents of posts for identification of opinion leaders. According to experiments in a real-world bulletin board data set, this proposed approach demonstrates high potential in identifying opinion leaders.

## 1 Introduction

This research identifies opinion leaders using the technique of text mining, since the opinion leaders affect other members via word of mouth (WOM) on social networks. WOM defined by Arndt (1967) is an oral person-to-person communication means between an information receiver and a sender, who exchange the experiences of a brand, a product or a service based on a non-commercial purpose. Internet provides human beings with a new way of communication. Thus, WOM influences the consumers more quickly, broadly, widely, significantly and consumers are further influenced by other consumers without any geographic limitation (Flynn et al., 1996).

Nowadays, making buying decisions based on WOM becomes one of collective decision-making strategies. It is nature that all kinds of human groups have opinion leaders, explicitly or implicitly (Zhou et al., 2009). Opinion leaders usually have a stronger influence on other members through their new information, ideas and representative opinions (Song et al., 2007). Thus, how to identify opinion leaders has increasingly attracted the attention of both practitioners and researchers.

As opinion leadership is relationships between members in a society, many existing opinion leader identification tasks define opinion leaders by analyzing the entire opinion network in a specific domain, based on the technique of social network analysis (SNA) (Kim, 2007; Kim and Han, 2009). This technique depends on relationship between initial publishers and followers. A member with the greatest value of network centrality is considered as an opinion leader in this network (Kim, 2007).

However, a junk post does not present useful information. A WOM with new ideas is more interesting. A spam link usually wastes readers' time. A long post is generally more useful than a short one (Agarwal et al., 2008). A focused document is more significant than a vague one. That is, different documents may contain different influences on readers due to their quality of WOM. WOM documents per se can also be a major indicator for recognizing opinion leaders. However, such quantitative approaches, i.e. number-based or

SNA-based methods, ignore quality of WOM and only include quantitative contributions of WOM.

Expertise, novelty, and richness of information are three important features of opinion leaders, which are obtained from WOM documents (Kim and Han, 2009). Thus, this research proposes a text mining-based approach in order to identify opinion leaders in a real-world bulletin board system.

Besides this section, this paper is organized as follows. Section 2 gives an overview of features of opinion leaders. Section 3 describes the proposed text mining approach to identify opinion leaders. Section 4 describes the data set, experiment design and results. Finally, a conclusion and further research work are given in Section 5.

## 2 Features of Opinion Leaders

The term "opinion leader", proposed by Katz and Lazarsfeld (1957), comes from the concept of communication. Based on their research, the influence of an advertising campaign for political election is lesser than that of opinion leaders. This is similar to findings in product and service markets. Although advertising may increase recognition of products or services, word of mouth disseminated via personal relations in social networks has a greater influence on consumer decisions (Arndt, 1967; Khammash and Griffiths, 2011). Thus, it is important to identify the characteristics of opinion leaders.

According to the work of Myers and Robertson (1972), opinion leaders may have the following seven characteristics. Firstly, opinion leadership in a specific topic is positively related to the quantity of output of the leader who talks, knows and is interested in the same topic. Secondly, people who influence others are themselves influenced by others in the same topic. Thirdly, opinion leaders usually have more innovative ideas in the topic. Fourthly and fifthly, opinion leadership is positively related to overall leadership and an individual's social leadership. Sixthly, opinion leaders usually know more about demographic variables in the topic. Finally, opinion leaders are domain dependent. Thus, an opinion leader influences others in a specific topic in a social network. He or she knows more about this topic and publishes more new information.

Opinion leaders usually play a central role in a social network. The characteristics of typical network hubs usually contain six aspects, which are ahead in adoption, connected, travelers, information-hungry, vocal, and exposed to media more than others (Rosen, 2002). Ahead in adoption means that network hubs may not be the first to adopt new products but they are usually ahead of the rest in the network. Connected means that network hubs play an influential role in a network, such as an information broker among various different groups. Traveler means that network hubs usually love to travel in order to obtain new ideas from other groups. Information-hungry means that network hubs are expected to provide answers to others in their group, so they pursue lots of facts. Vocal means that network hubs love to share their opinions with others and get responses from their audience. Exposed to media means that network hubs open themselves to more communication from mass media, and especially to print media. Thus, a network hub or an opinion leader is not only an influential node but also a novelty early adopter, generator or spreader. An opinion leader has rich expertise in a specific topic and loves to be involved in group activities.

As members in a social network influence each other, degree centrality of members and involvement in activities are useful to identify opinion leaders (Kim and Han, 2009). Inspired by the PageRank technique, which is based on the link structure (Page et al., 1998), OpinionRank is proposed by Zhou et al. (2009) to rank members in a network. Jiang et al. (2013) proposed an extended version of PageRank based on the sentiment analysis and MapReduce. Agarwal et al. (2008) identified influential bloggers through four aspects, which are recognition, activity generation, novelty and eloquence. An influential blog is recognized by others when this blog has a lot of in-links. The feature of activity generation is measured by how many comments a post receives and the number of posts it initiates. Novelty means novel ideas, which may attract many in-links from the blogs of others. Finally, the feature of eloquence is evaluated by the length of post. A lengthy post is treated as an influential post.

Li and Du (2011) determined the expertise of authors and readers according to the similarity between their posts and the pre-built term ontology. However both features of information novelty and influential position are dependent on linkage relationships between blogs. We propose a novel

text mining-based approach and compare it with several quantitative approaches.

## 3 Quality Approach-Text Mining

Contents of word of mouth contain lots of useful information, which has high relationships with important features of opinion leaders. Opinion leaders usually provide knowledgeable and novel information in their posts (Rosen, 2002; Song et al., 2007). An influential post is often eloquent (Keller and Berry, 2003). Thus, expertise, novelty, and richness of information are important characteristics of opinion leaders.

### 3.1 Preprocessing

This research uses a traditional Chinese text mining process, including Chinese word segmenting, part-of-speech filtering and removal of stop words for the data set of documents. As a single Chinese character is very ambiguous, segmenting Chinese documents into proper Chinese words is necessary (He and Chen, 2008). This research uses the CKIP service (http://ckipsvr.iis.sinica.edu.tw/) to segment Chinese documents into proper Chinese words and their suitable part-of-speech tags. Based on these processes, 85 words are organized into controlled vocabularies as this approach is efficient to capture the main concepts of document (Gray et al., 2009).

### 3.2 Expertise

This can be evaluated by comparing their posts with the controlled vocabulary base (Li and Du, 2011). For member *i*, words are collected from his or her posted documents and member vector *i* is represented as $f_i=(w_1, w_2, \ldots w_j, \ldots, w_N)$, where $w_j$ denotes the frequency of word *j* used in the posted documents of user *i*. *N* denotes the number of words in the controlled vocabulary. We then normalize the member vector by his or her maximum frequency of any significant word. The degree of expertise can be calculated by the Euclidean norm as show in (1).

$$\exp_i = \left\| \frac{f_i}{m_i} \right\|, \quad (1)$$

where $\|\bullet\|$ is Euclidean norm.

### 3.3 Novelty

We utilize Google trends service (http://www.google.com/trends) to obtain the first-search time tag for significant words in documents. Thus, each significant word has its specific time tag taken from the Google search repository. For example, the first-search time tag for the search term, Nokia N81, is 2007 and for Nokia Windows Phone 8 is 2011. We define three degrees of novelty evaluated by the interval between the first-search year of significant words and the collected year of our targeted document set, i.e. 2010. This significant word belongs to normal novelty if the interval is equal to two years. A significant word with an interval of less than two years belongs to high novelty and one with an interval greater than two years belongs to low novelty. We then summarize all novelty values based on significant words used by a member in a social network. The equation of novelty for a member is shown in (2).

$$nov_i = \frac{e_h + 0.66 \times e_m + 0.33 \times e_l}{e_h + e_m + e_l}, \quad (2)$$

where $e_h$ , $e_m$ and $e_l$ is the number of words that belong to the groups of high, normal and low novelty, respectively.

### 3.4 Richness of Information

In general, a long document suggests some useful information to the users (Agarwal et al., 2008). Thus, richness of information of posts can be used for the identification of opinion leaders. We use both textual information and multimedia information to represent the richness of information as (3).

$$ric = d + g, \quad (3)$$

where *d* is the total number of significant words that the user uses in his or her posts and *g* is the total number of multimedia objects that the user posts.

### 3.5 Integrated Text Mining Model

Finally, we integrate expertise, novelty and richness of information from the content of posted documents. As each feature has its own

distribution and range, we normalize each feature to a value between 0 and 1. Thus, the weights of opinion leaders based on the quality of posts become the average of these three features as (4).

$$ITM = \frac{Norm(nov) + Norm(exp) + Norm(ric)}{3}. \quad (4)$$

## 4 Experiments

### 4.1 Data Set

Due to lack of available benchmark data set, we crawl WOM documents from the Mobile01 bulletin board system (http://www.mobile01.com/), which is one of the most popular online discussion forums in Taiwan. This bulletin board system allows its members to contribute their opinions free of charge and its contents are available to the public. A bulletin board system generally has an organized structure of topics. This organized structure provides people who are interested in the same or similar topics with an online discussion forum that forms a social network. Finding opinion leaders on bulletin boards is important since they contain a lot of availably focused WOM. In our initial experiments, we collected 1537 documents, which were initiated by 1064 members and attracted 9192 followers, who posted 19611 opinions on those initial posts. In this data set, the total number of participants is 9460.

### 4.2 Comparison

As we use real-world data, which has no ground truth about opinion leaders, a user centered evaluation approach should be used to compare the difference between models (Kritikopoulos et al., 2006). In our research, there are 9460 members in this virtual community. We suppose that ten of them have a high possibility of being opinion leaders.

As identification of opinion leaders is treated to be one of important tasks of social network analysis (SNA), we compare the proposed model (i.e. ITM) with three famous SNA approaches, which are degree centrality (DEG), closeness centrality (CLO), betweenness centrality (BET). Involvement (INV) is an important characteristic of opinion leaders (Kim and Han, 2009). The number of documents that a member initiates plus the number of derivative documents by other members is treated as involvement.

Thus, we have one qualitative model, i.e. ITM, and four quantitative models, i.e. DEG, CLO, BET and INV. We put top ten rankings from each model in a pool of potential opinion leaders. Duplicate members are removed and 25 members are left. We request 20 human testers, which have used and are familiar with Mobile01.

In our questionnaire, quantitative information is provided such as the number of documents that the potential opinion leaders initiate and the number of derivative documents that are posted by other members. For the qualitative information, a maximum of three documents from each member are provided randomly to the testers. The top 10 rankings are also considered as opinion leaders based on human judgment.

### 4.3 Results

We suppose that ten of 9460 members are considered as opinion leaders. We collect top 10 ranking members from each models and remove duplicates. We request 20 human testers to identify 10 opinion leaders from 25 potential opinion leaders obtained from five models. According to experiment results in Table 1, the proposed model outperforms others. This presents the significance of documents per se. Even INV is a very simple approach but it performs much better than social network analysis models, i.e. DEG, CLO and BET. One possible reason is the sparse network structure. Many sub topics are in the bulletin board system so these topics form several isolated sub networks.

| | Recall | Precision | *F*-measure | Accuracy |
|---|---|---|---|---|
| DEG | 0.45 | 0.50 | 0.48 | 0.56 |
| CLO | 0.36 | 0.40 | 0.38 | 0.48 |
| BET | 0.64 | 0.70 | 0.67 | 0.72 |
| INV | 0.73 | 0.80 | 0.76 | 0.80 |
| ITM | 0.82 | 0.90 | 0.86 | 0.88 |

Table 1: Results of models evaluated by recall, precision, F-measure and accuracy

## 5 Conclusions and Further Work

Word of mouth (WOM) has a powerful effect on consumer behavior. Opinion leaders have stronger influence on other members in an opinion society. How to find opinion leaders has been of interest to both practitioners and researchers. Existing models mainly focus on quantitative features of opinion leaders, such as the number of posts and the central position in the social network. This research considers this issue from the viewpoints of text mining. We propose an integrated text mining model by extracting three important features of opinion leaders regarding novelty, expertise and richness of information, from documents. Finally, we compare this proposed text mining model with four quantitative approaches, i.e., involvement, degree centrality, closeness centrality and betweenness centrality, evaluated by human judgment. In our experiments, we found that the involvement approach is the best one among the quantitative approaches. The text mining approach outperforms its quantitative counterparts as the richness of document information provides a similar function to the qualitative features of opinion leaders. The proposed text mining approach further measures opinion leaders based on features of novelty and expertise.

In terms of possible future work, some integrated strategies of both qualitative and quantitative approaches should take advantages of both approaches. For example, the 2-step integrated strategy, which uses the text mining-based approach in the first step, and uses the quantitative approach based on involvement in the second step, may achieve the better performance. Larger scale experiments including topics, the number of documents and testing, should be done further in order to produce more general results.

## References


Agarwal, N., Liu, H., Tang, L. and Yu, P. S. 2008. Identifying the Influential Bloggers in a Community. Proceedings of WSDM, 207-217.

Arndt, J. 1967. Role of Product-Related Conversations in the Diffusion of a New Product. Journal of Marketing Research, 4(3):291-295.

Flynn, L. R., Goldsmith, R. E. and Eastman, J. K. 1996. Opinion Leaders and Opinion Seekers: Two New Measurement Scales. Academy of Marketing

He, J. and Chen, L. 2008. Chinese Word Segmentation Based on the Improved Particle Swarm Optimization Neural Networks. Proceedings of IEEE Cybernetics and Intelligent Systems, 695-699.

Jiang, L., Ge, B., Xiao, W. and Gao, M. 2013. BBS Opinion Leader Mining Based on an Improved PageRank Algorithm Using MapReduce. Proceedings of Chinese Automation Congress, 392-396.

Katz, E. and Lazarsfeld, P. F. 1957. Personal Influence, New York: The Free Press.

Keller, E. and Berry, J. 2003. One American in Ten Tells the Other Nine How to Vote, Where to Eat and, What to Buy. They Are The Influentials. The Free Press.

Khammash, M. and Griffiths, G. H. 2011. Arrivederci CIAO.com Buongiorno Bing.com- Electronic Word-of-Mouth (eWOM), Antecedences and Consequences. International Journal of Information Management, 31:82-87.

Kim, D. K. 2007. Identifying Opinion Leaders by Using Social Network Analysis: A Synthesis of Opinion Leadership Data Collection Methods and Instruments. PhD Thesis, the Scripps College of Communication, Ohio University.

Kim, S. and Han, S. 2009. An Analytical Way to Find Influencers on Social Networks and Validate their Effects in Disseminating Social Games. Proceedings of Advances in Social Network Analysis and Mining, 41-46.

Kritikopoulos, A., Sideri, M. and Varlamis, I. 2006. BlogRank: Ranking Weblogs Based on Connectivity and Similarity Features. Proceedings of the 2nd International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications, Article 8.

Li, F. and Du, T. C. 2011. Who Is Talking? An Ontology-Based Opinion Leader Identification Framework for Word-of-Mouth Marketing in Online Social Blogs. Decision Support Systems, 51, 2011:190-197.

Myers, J. H. and Robertson, T. S. 1972. Dimensions of Opinion Leadership. Journal of Marketing Research, 4:41-46.

Page, L., Brin, S., Motwani, R. and Winograd, T. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University.

Rosen, E. 2002. The Anatomy of Buzz: How to Create Word of Mouth Marketing, 1st ed., Doubleday.

Song, X., Chi, Y., Hino, K. and Tseng, B. L. 2007. Identifying Opinion Leaders in the Blogosphere. Proceedings of CIKM'07, 971-974.

Zhou, H., Zeng, D. and Zhang, C. 2009. Finding Leaders from Opinion Networks. Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics, 266-268.