Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Opinion Groups Identification in Blogosphere Based on the Techniques of Web Mining and Social Networks Analysis

I-Hsien Ting[1+], Chia-Shung Yen[2]

[1]Department of Information Management, National University of Kaohsiung, TAIWAN

[2]Department of Communication, National Chung Cheng University, TAIWAN

Abstract. Recently, it is a very important issue about how to identify opinion groups in blogosphere, due to it is essential for many applications who want to utilize the information in blogs. Web mining is now claimed as the most powerful tool to extract use information and social networks analysis is a tool for us to identify the relationship and structure in a network. Therefore, in this paper, we will propose a system architecture to apply the techniques of web mining and SNA for opinion groups identification in blogs. A model about how to measure the similarity of different groups by using the mean of SNA is also discussed in this paper. Furthermore, an experiment has been performed as an empirical study as well as the analysis results will be discussed in this paper.

Keywords: Social Networks Analysis, Blogosphere, Opinion identification, web mining

1. Introduction

Blogosphere now are very popular platform for users to post and share articles with each other, no matter traditional blogs or micro-blogs (such as twitter and plurk). Recently, there are numerous data and information aggregrated in blogs and which has becoming a very valuable database[1]. Therefore, blogs now is a new target for different applications, such as marketing[2], crime detection, politic[3], etc. For these applications, the most important issue is about how to identify the opinions in blogs[4]. For the application of politic, it is essential to understand the opinions of citizens for a public policy or a social event. This is also very helpful for the prediction of election results, especially for such countries with very clear political stance[5].

Currently, most applications in this research area is using the techniques of web mining to extract useful information in blogs. Normally, the data from blogs are semi-structured and the process of natural language processing (NLP) is always necessary. However, web mining can only be used to identify groups by using the techniques of classification or clustering, but it can't be used to illustrate the structure and relationship in a group. Social Networks Analysis (SNA) is a methodology which can be used to identify the nodes, the roles and the social relationship in social networks[6]. We believe the results of opinion groups identification will be better, if the technique of web mining and SNA can be combined together. Therefore, the objective of this paper is to propose an approach to combine web mining and SNA together for opinion gropus identification.

The rest of the paper is structured as follow. In section 2, we will provide a brief literature review about social networks analysis and web mining as well as a discussion about current works in this research area. The research methodology and system architecture will be proposed in section 3. The experiment design and analysis results are included in section 4, and we will conclude this paper in section 5 with some fugure research suggestions.

+ Corresponding author. Tel.: +886-7-5919751; fax: +886-7-5919328.
E-mail address: iting@nuk.edu.tw

2. Literature Review and Related Works

The research methodology of social network analysis is developed to understand the relationship between "actors", and the term actor can be a person, an organization, an event or an object [7]. In a social network, each actor is presented as a node and each pair of nodes can be connected by lines to show the relationships. The social network structure graph is a graph that formed by those lines and nodes, and social network analysis is therefore a methodology that used to understand the graph and the relationships and actors in the social network [8].

The most important measurements of SNA include network size, diameter, density, centrality and structure holes [9]. Size is a measurement to measure the amount of nodes or links in a network, and the measurement of diameter is to measure the amount of nodes between two nodes in a network. Density is used to calculate the closeness of a network [10]. These measurements are common used in many social network related researches and will be used in this paper as well.

Traditionally, researches about SNA are mainly focus on small group of actors and are process manually in most cases. However, with the rapid growth of Internet and web techniques, more and more data have been collected and it has become a hard task to process these data by only the mean of manually.[11] Therefore, the scholars of information technology and computer science are starting to devote related researches to deal with these research issues and web mining is consider as the most suitable techniques to analyze the data from web [12].

Web content mining, text mining or natural language processing are very useful techniques that can be used for social network analysis. For example, web content mining can be used to categorize or classify the documents of social networking website, especially for blog or text forum analysis to categorize or classify the articles of blogs. The article categorization is usually the first task for many social networks analyses or applications.

Web usage mining plays an important role in social networks analysis as well. It is useful for the social network analysis of social networks extraction. The usage data and users' communication in social networking website can be transformed to relational data for social-networks construction [12].

In Adamic and Glance 2004, they have proposed a method based on the pre-defined blogs which have been divided into two different groups which support different parties in 2004 US election. This paper is the only and most valuabel paper who focus on how to detect opinion groups in blogosphere for political application. In this paper, they use a visualization tool to illustrate the connection and relationship of the users [5] as well as use TF-IDF for keywords extraction for analysis. However, they do not utilize web mining related techniques and the information of the network strcutre to detect the opinion groups, which has been considered as very important information for groups detection [4]. Therefore, in this paper, we will propose an approach, which is based on web mining techniques and social networks analysis. We expect the appraoch can of help to identify opinion groups in blogoshpere.

3. The Research Methodology and System Architecture

In order to achieve the idea that proposed in this paper, we will introduce the research methodology and system architecture in this section. The methodology can be divided into two phases. The first phase is the model training phase and the second phase is the opinion groups identification phase. The second phase is also the system architecture that proposed in this paper.

Figure 1 shows the first phase of the research methodology and this phase is also called the traning phase. In this phase the training data is the pre-defined data from blogs for an special event. The data will be identified by an expert to divide the opinions into positive and negative opinion. Then, the data will be pre-processed and using web content mining technique to extract keyword based on TF-IDF (Term Frequency-Inverse Document Frequency). The TF-IDF value will then be used to rank the keywords. Furthermore, the relationship of the users in the same group will be measured by counting the responses to an article in the blogs and then the users' matrix can be established.

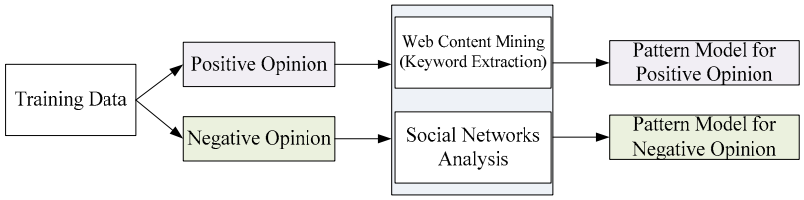

Fig.1 The training phase as the first phase of the research methodology

For social networks analysis, we will take degree centrality, closeness centrality, betweeness and cluster coefficient into account [6]. The four SNA measurements and the top five keywords (the number of keywords depends on the requirement of different applications) will be used to form the pattern model for positive opinion group and negative opinion group. A pattern model can be presented as formulation 1. In formulation 1, i=1 or 2, m=5 and n=4.

$$PM_i = \{K_{1...m}, SN_{1...n}\} \tag{1}$$

Figure 2 is the second phase of the research methodology, it is also the system architecture of the prposed system. In this phase, we will collect real data from blogs. The data will be pre-processed in order to filter-out useful information (the detail of the information will be discussed in section 5) and store in a database. Same as the first phase, the keyword of the data will be extracted and the four SNA values will be measured.

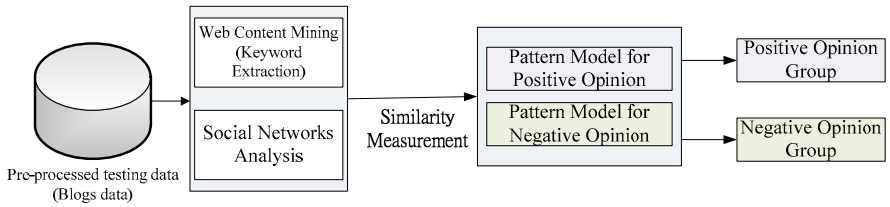

Fig.2 The second phase (system architecture) of the research methodology

For each user in the database will obtain an user model. The user model can then be used to measure the similariy between users and the two defined pattern models. The user model can be presented as formulation 2. In formulation 2, i denotes the number of users in the database, m=5 and n=4.

$$UM_i = \{K_{1...m}, SN_{1...n}\} \tag{2}$$

In the research methodology, the similarity of each user will be measured by using the cosine similarity to compare with the pattern model of positive opinion group or negative opinion group. When the similarity is higher then a threshold, it will be classified as either positive or negative opinion. Finally, we can identify the two different types of groups by using the techniques of web content mining and social networks analysis.

4. Empirical Study and Analysis Results

4.1 The Training Phase

In this section, we will use the real data from a very famous blogs platform in Taiwan (http://www.wretch.cc/) as the empirical study. In this empirical study, 100 users are selected for each opinion groups (positive and negative) which are defined by a domain expert. Therefore, the entire blogs of 200 users are collected with about 3000 web pages.

These blogs will be pre-processed and stored in a database with two tables, the fields in table 1 are Message_ID, Message_URL, Date, User_ID, Message_Content. Table 2 is mainly designed to store the information about users' response to a message in table1 and the fields are Response_ID, Message_ID, User_ID, Date, Response_Content.

After performing the step of keywords extraction and social networks analysis, the pattern model for positive and negative opinion groups. Table1 shows the pattern models.

Tab.1 the pattern models of positive and negative opinion group

	Positive pattern model	**Negative pattern model**
Keyword1 (TF-IDF)	Great (0.512)	Disagree (0.677)
Keyword2 (TF-IDF)	Good decision (0.341)	Unreasonable (0.336)
Keyword3 (TF-IDF)	Nice (0.107)	Stop (0.111)
Keyword4 (TF-IDF)	Understand (0.044)	Stupid (0.097)
Keyword5 (TF-IDF)	Like (0.035)	Ridiculous (0.055)
Degree centrality	2.414	4.112
Closeness centrality	1.741	2.121
Betweeness centrality	2.322	3.819
Cluster coefficient	2.865	4.7

In table 1, the top five keywords have been listed (the number of keywords depend on the requirement of different applications) as well as the four SNA measurements. The keywords are originally in Chinese language and have been translated to the same meaning in English. From table 1, it shows that the differences of the keywords in the two opinion groups are very significant. Furthermore, the SNA measurements are also significance enough for us to distinguish the two different opinion groups.

4.2 The Opinion Groups Identification Phase

In section 4.1, the two pattern models have been defined. In this section, the matrixes of the pattern model will then be used to identify opinion groups from the real data from the blogs platform. Before performing the process, we have selected 1000 users which have posted related blogs to a particular political event.

The blogs of 1000 users then will be processed by using keyword extraction and social networks analysis. After performing the process, each user will has one user matrix which is the same design to the pattern models. The similarity 1000 users' matrix and the two pre-defined pattern models. In this paper, we have set a threshold of 0.5 for classify each user into the two groups. For example, if a user's matrix is (Great 0.512, idea 0.412, Nice 0.107, Next 0.012, Like 0.145, 2.11, 1.89, 2.12, 2.88) then the cosine similarity value to positive pattern mode of this example is 0.895 which is greater than 0.5. This user will then be classify to positive opinion group. However, the threshold could be different to different application and 0.5 is only for explanation in this empirical study.

5. Conclusion

In this paper, we have proposed an approach to use the techniques of web mining and social networks analysis to identify opinion groups in blogosphere, particular for political event and issues. This methodology is very useful for the governments or parties to understand the opinion of citizens. This is also very helpful for election prediction which probably can be used as another reference in addition to traditional opinion poll. The empirical study that provided in this paper also shows very good result to classify users to appropriate groups. The visualization of users' structure also presents very significant and interesting results. We can identify the two opinion groups very easily from the figure.

In this paper, we only focus on two groups with very different opinion, which can be distinguished easier. If there are more opinions for an political event, then it would be more difficult for the approach to analyze. Not only in the training phase but also in the opinion identification phase. Therefore, in the future, we suggest related researches can focus on more complex groups. Furthermore, related applications can also be designed for the proposed approaches, which we think will be of benefit to this research area.

6. Acknowledgements

The corresponding author would like to thank the funding and support from National Science Council (NSC) in Taiwan, NSC 100-2410-H-390-004.

7. References

[1] Chang P.-S., Ting I.-H., and Wang S.-L., 2011, “Towards Social Recommendation System Based on the Data from Microblogs,” 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 672-677.

[2] Wang K.-Y., Thongpapanl N., Wu H.-J., and Ting I.-H., 2011, “Identifying Structural Heterogeneities between Online Social Networks for Effective Word-of-Mouth Marketing,” 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 418-422.

[3] Cepela N. T., and Danowski J. a, 2009, “Automatic Mapping of Social Networks of Political Actors from Large Collections of News Stories,” 2009 International Conference on Advances in Social Network Analysis and Mining, pp. 212-218.

[4] Song X., Chi Y., Hino K., and Tseng B., 2007, “Identifying opinion leaders in the blogosphere,” Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM ’07, p. 971.

[5] Adamic L. A., and Glance N., “The political blogosphere and the 2004 U.S. election: divided they blog,” In International Workshop on Link Discovery (LinkKDD, pp. 36-43.

[6] Scott J., 2002, Social Network Analysis: A Hand Book, SAGE Publication.

[7] Wellman B., and Berkowitz S. D., 1988, Social Structures: A Network Approach, Cambridge University Press, Cambridge, England, UK.

[8] Lento T., Welser H. T., Gu L., and Marc S., 2006, “The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System,” Proceedings of the 15th International World Wide Web Conference, Edinburgh, Scotland, UK.

[9] Burt S. P., 1992, Structural Holes, Harvard University Press, Cambridge, MA, USA.

[10] Furukawa T., Matsuo Y., Ohmukai I., Uchiyama K., and Ishizuka M., 2007, “Social Networks and Reading Behavior in the Blogosphere,” Proceedings of ICWSM 2007, Boulder, Colorado, USA, pp. 51-58.

[11] Godbole N., Srinivasaiah M., and Skiena S., 2007, “Large-Scale Sentiment Analysis for News and Blogs,” Proceedings of ICWSM 2007, Boulder, Colorado, USA.

[12] Ting I.-hsien, 2008, “Web mining techniques for on-line social networks analysis,” Proceedings of the 2008 International Conference on Service Systems and Service Management, IEEE, pp. 1-5.