Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (16): 3315-3319, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3315.3319

A Novel Public Opinion Mining Method on Microblog Platform

[1]Zhao Zhe, [1]Xiang Yang, [2]Zhang Bo, [1]Zhang Qi and [3]Pan Tao
[1]College of Electronics and Information Engineering, Tongji University,
201804, Shanghai, People Republic of China
[2]College of Information, Mechanical and Electrical Engineering,
Shanghai Normal University, Shanghai, 200234, People Republic of China
[3]Shen Hua Hollysys Information Technology Co. Ltd., Beijing, 100000, People Republic of China

Abstract: Microblog has been an important platform for expression of public opinion towards policy decisions. One key challenge for policymakers is to mine public opinions from microblog platforms as soon as possible. In order to deal with the challenge, this paper proposes a Topic Detection and Tracking (TDT) algorithm based on self-adjusting vector space model (VSM) and an opinion mining method based on comments. Furthermore, an innovative opinion mining system is developed, using mciroblog as the opinion mining platform and combining natural language processing techniques with similarity calculation and polarity calculation. A series of related experiments are employed to verify the efficiency and maneuverability of the algorithm.

Key words: Opinion mining, TDT, microblog, polarity calculation

INTRODUCTION

In the 21st century, with the continuous development of information technology, the Internet has become the "fourth media" after press, radio and television. The emergence and rapid development of microblog made it became a very important medium for people to receive information, express their opinions and communicate. In these popular “new medias”, microblog is definitely a leader of them. Microblog can gather public's opinions and disseminate them in a very short time, affect people's behavior even have a great influence on the whole society. Therefore, the public opinion expressed on microblog has become a vital force for governments and should not be ignored in social development (Liu, 2007).

There are many existing works which focus on public opinion mining (Pang *et al*., 2002; Zhao and Li, 2009; Kim *et al*., 2009). These methods apply machine learning, natural language processing, association rules and ontology to opinion mining. However, as we know, the number of microblog users is huge and the texts which contain public opinions are short, unstructured, updated every second and disseminated through relationships. These new features imply that traditional topic tracking algorithm and public opinion mining method may not be suitable for microblogs. This paper addresses a novel method of mining public opinions on microblog platform which is composed by two core issues: TDT algorithm based on self-adjusting VSM and an opinion mining method based on comments. Our main motivation of this work is to track policy decisions published on microblog, make track evaluation and sentiment analysis of users' comments and eventually give feedback to the relevant decision-making departments. Our research utilizes association rules for theme extraction and natural language processing techniques for polarity analysis.

The organization of this study is as follows: section 2 reviews related work of topic tracking and public opinion mining. Section 3 details the TDT algorithm based on self-adjusting VSM. Section 4 expatiate the comments based opinion mining method. Section 5 shows our experimental results. Finally, section 6 concludes the paper and gives future work.

RELATED WORKS

Traditional topic tracking methods mainly include two research trends: Based on knowledge and based on the statistics. The former focuses on the analysis of association and inheritance relation between the contents of different reports based on specific domain knowledge. The latter adopts statistical methods to decide the relevance of reports to the policy decision, according to probability distribution of the characteristics. Here are some representative researches of topic tracking. Paper proposes a topic tracking algorithm based on language

Corresponding Author: Zhang Bo, College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, 200234, People Republic of China

model and the KL-divergence clustering algorithm (Yamron *et al*., 2000). Li *et al*. (2005) introduce historical news event detection method based on probability retrieval model. Public opinion mining is a wide topic and sentiment analysis is a very important part of opinion mining. Foreign scholars began to widely focus on sentiment analysis research since the mid-1990s. The main types of research methods: Uses Machine learning based traditional text categorization technique for sentiment analysis (Jia *et al*., 2011); Zhu *et al*. (2006) explore using Polarity dictionary for sentiment analysis. Topic tracking and public opinion mining have a rapid development in recent years with wide-ranging application in natural language processing, information filtering and management.

TDT ALGORITHM BASED ON SELF-ADJUSTING VSM

In this study, the topic tracking task is to identify and track microblog messages and comments related to the topic and provide web links of these microblog messages and comments on the condition that we have already know the content or the topic of the new policy decision.

Build vector space: Vector Space Model is a text representation model (Cortes and Vapnik, 1995; Vapnik, 1995). In this model, each text can be represented by a multi-dimension vector in the vector space:

$$\phi(d) = (tf(t_1,d), tf(t_2,d), \ldots, tf(t_N,d)) \in R^N \quad (1)$$

where, t_i indicates a word, N is the total number of words in the word list, $tf(t_i, d)$ is the appearance frequency of word t_i in text d, this appearance can be absolute word frequency or relative word frequency. Absolute word frequency is the word's actual appearance frequency in the text. Relative word frequency is the normalized frequency. The most commonly used formula to calculate RWF is TF-IDF and we use TF-IDF equation in this study:

$$tf(t,d) = \frac{f(t,d) \times \log(\frac{N}{n_1} + 0.01)}{\sqrt{\sum_{t \in d}[f(t,d) + \log(\frac{N}{n_1} + 0.01)]^2}} \quad (2)$$

where f(t,d) is the absolute word frequency of word t in text d. N is the sum of texts used for training model. n_1 is the number of texts that contain word t. Denominator is the normalization factor.

The process of building vector space model using raw data from microblog as Fig. 1 shows.

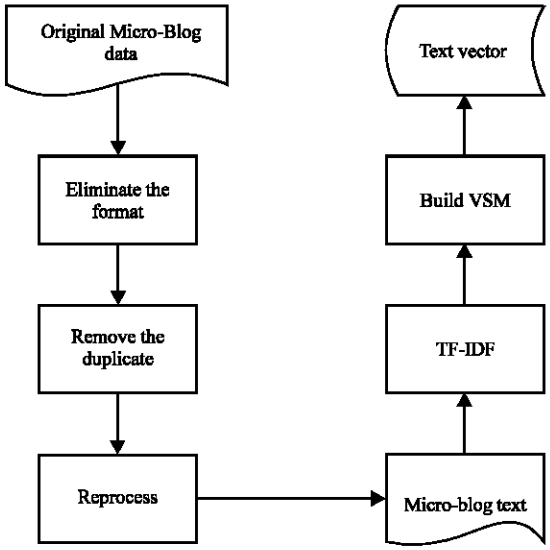

Fig. 1: The process of building VSM

Semantic combined similarity calculation: After building Vector Space Model, we can judge topics of two texts are close or not though computing the similarity of vectors of the texts. Semantic meaning combined cosine similarity calculation method is: First, build a list of synonyms and analogous words, before calculating the cosine distance of text vector x and y, go through all the dimensions of x and y, if a text vector contains the word in the list of synonyms and analogous words, select a unified word to replace the original one.

Then, two microblog vectors similarity calculation formula is as follows:

$$similar(D_A, D_B) = \cos(\vec{V}_A, \vec{V}_B) = \frac{\sum_{k=1}^{n}(W_{Ak} * W_{Bk} * similar(T_{Ak}, T_{Bk}))}{\sqrt{\sum_{k=1}^{n} A_{tk}^2} * \sqrt{\sum_{k=1}^{n} B_{tk}^2}} \quad (3)$$

In above equation, D_A and D_B respectively represent text A and text B, $\vec{V}_A$ and $\vec{V}_B$ represent the corresponding text vectors of D_A and D_B, T_{AK} is the Kth vector component of $\vec{V}_A$, W_{AK} is the weight of T_{AK}. T_{BK} is the Kth vector component of $\vec{V}_B$. W_{BK} is weight of T_{BK}. $similar(T_{AK}, T_{BK})$ is the similarity of T_{AK} and T_{BK}, the formula used to calculate $similar(T_{AK}, T_{BK})$ is as follows:

COMMENTS BASED OPINION MINING METHOD

Comments based opinion mining model: In this study, the object of study is the policy comments and event comments, besides general characteristics of product comments, policy and event comments are sudden, easy to transfer to another topic and full of emotion.

In order to facilitate the semantic polarity analysis described below, this paper divided subjective comments sentences into the following two categories:

- **Strong polar subjective sentence:** These statements express single and explicit semantic meaning
- **Weak polar subjective sentence depending context:** Determining the polarity of these sentences relies on the context

For the above two different situations, we propose a method for calculating the polarity as shown below:

- **Polarity analysis of strong polar sentence:** To use statistical methods, that is, to build polar dictionary
- **Polarity analysis of weak polar sentence:** In this case, simply using polar dictionary can't solve the problem. In order to deal with this kind of polar analysis, this paper proposes a polarity analytical method based on combination of natural language processing techniques and statistical methods

Information selection: The randomness of microblog comments leads to the quality of comments varies greatly. So it is very important to design a useful information selection system. This paper uses AC automation for information selection.

The Fig. 2 is the pattern tree of the set of patterns p = {the, she, his, hers}.

AC automation extended from the pattern tree is a 6-tuple:

$$M = \{Q, A, goto, fail, S, F\} \quad (4)$$

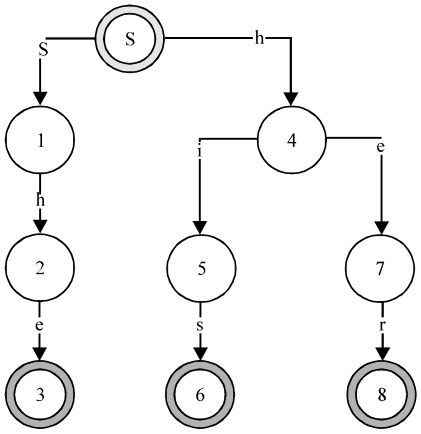

Fig. 2: Pattern Tree

where, Q is a finite set of states. A is a finite input character table. S is the initial state. F is the set of final states. Both S and F are the subset of Q. Function goto and function fail are transition functions defined as follow.

- goto (s, a): Begin from s, the current state, turn to the next state along the path labeled a
- fail (a): If the next state along the path labeled a in the pattern tree can=t be found, then fail(a) = s

Thus, the process of finding patterns in the string has been transformed into a search process of pattern tree, starting from the root node along the path labeled with the characters:

- If the automatic machine can reach the final state, then the corresponding pattern exists in the main string
- Else it doesn't exist in the main string

THEME EXTRACTION BASED ON ASSOCIATION RULES

In this study, we use Apriori algorithm on mining association relationship between words and phrases to extract comment theme. Process of theme extraction based on Apriori algorithm (Table 1) is showed as below.

- Scan the entire comment text, calculate each word's frequency of occurrence in the text. If there is just only one noun in the word list, then regard the noun and its qualifier as the theme
- According to the statistical result got in step 1, remove the words whose occurrence frequency below the minimum threshold from the word list. The rest forms the 1-item candidate set
- According to the set obtained in step 2, get a 2-item candidate set though Cartesian operations. Re-scan the entire comment text. Elements in the set will also be removed if their occurrence frequency below the minimum threshold
- By this way, we can get K-item frequent set. If the K-item frequent set contains only one candidate, the algorithm exits

Polarity analysis: The procedure of polarity analysis method based on natural language processing techniques is:

- Polarity dictionary construction: use HowNet

- Word segmentation: use Hightman, an open source project
- Contextual polarity calculation of polarity word: the contextual polarity of polarity word may not equal to its original polarity. The contextual polarity is calculated according to its context and the dependence relationship between polarity words in the sentence

In order to facilitate the calculation of polarity of polarity words, we quantize the polarity and strength of the words (Table 2).

Strength quantization meets the following equation:

$$0 \leq \text{Strenght}(x) \leq 1$$

To calculate the specific contextual polarity of the word, it needs to construct a negative word dictionary and an emphasis word dictionary. The construction of negative word dictionary just needs a database for negative words, but the emphasis word dictionary needs to record emphasis words and strength of the words.

The algorithm description of polarity analysis method based on natural language processing techniques is as follows:

Algorithm Polarity analysis:

Step 1: Word segmentation, represented by "cut(comment)", the result represented by "cut_result"

Step 2: Traverse "cut_result", if the traversing word (represented by "word") is adjective, turn to step (3), else continue step (2)

Step 3: If "polarity(word)!=0", then turn to the next step, else turn to step(2). "polarity(word)" represents the polarity of "word"

Step 4: "search(word)", search the polarity word dictionary for the original polarity of "word"

Step 5: Traverse the words which has dependences relations with "word", represented by "relation_w", if "relation_w" is a negative word, then "polarity(word)= -polarity(word)", if "relation_w" is an emphasis word, then "polarity(word)*=strength(relation_w)

Step 6: If the traverse of "cut_result" ends, then output "polarity (word)", else turn to step (2)

Table 1: Polarity quantization

Positive	Polarity = 1
Neutral	Polarity = 0
Negative	Polarity = -1

Table 2: Strength quantization

General	Strength = 0.5
Very	Strength = 0.8
Extremely	Strength = 1.0

EXPERIMENTS AND RESULTS

Experiment of TDT algorithm based on self-adjusting VSM: This experiment tested the example of the State Council issued the policy of house price control based on Sina microblog platform. The input of policy decision information is as follows:

The State Council executive meeting held in January 26th, 2013 required further regulation of the real estate market and raised down-payment rates on second mortgages to 60%. The meeting also emphasized the responsibility of local government in the price control. Local governments should know the targets of price control and announce to the society.

After the algorithm test, the outputs are as in Table 3.

Seen from the experiment results, the TDT algorithm based on self-adjusting VSM can accurately track decision-making information and feed the microblogs related to the policy decision back. It can be concluded that TDT algorithm has strong practicality.

Experiment of polarity calculation method based on natural language processing techniques: This experiment was based on TDT algorithm experiment. The purpose was to test the feasibility of the polarity calculation method based on natural language processing techniques. The data were comments of a microblog captured by TDT algorithm, as shown in the Table 4.

Table 3: Microblogs got from the Internet

Num	Content
1	Many policies have been made for regulation of real estate. Will the new policy work? It's hard to say
2	At present, all means haven't hit the point. Keeping prices at a reasonable level the key is to make occupants make full use of their house. If everything revolves around this goal, I do not believe prices will not decline
3	Radical ideas, not stand scrutiny
4	No land, currencies, interest rates and other means of regulation, what local governments can do to bear the responsibility to curb housing prices?
5	Property tax is levied sooner or later. I care about the fairness and transparency of execution
..	....
100	New policy for regulation of the real estate market, all local governments should clear price control targets and announced to the society, raise down-payment rates on second mortgages to 60%. If the price of real estate declines, who will benefit?

Table 4: Comments of a microblog got from the Internet

Microblog	Many policies have been made for regulation of real estate. Will the new policy work? It's hard say. The problems of interest rates and the supply of land haven't been solved. The future supply shortage is expected to remain unchanged. The market is difficult to have a fundamental change.	
Comments	1	There is no moral bottom line.
	2	Radical ideas, not stand scrutiny
	3	Perpetual regulation
	4	The key point is who can benefit from the policy? People or politicians?
	5	I agree.
	Y	YYYY
	781	The problem can't solved by market economy is also very difficult to solve by planned economy

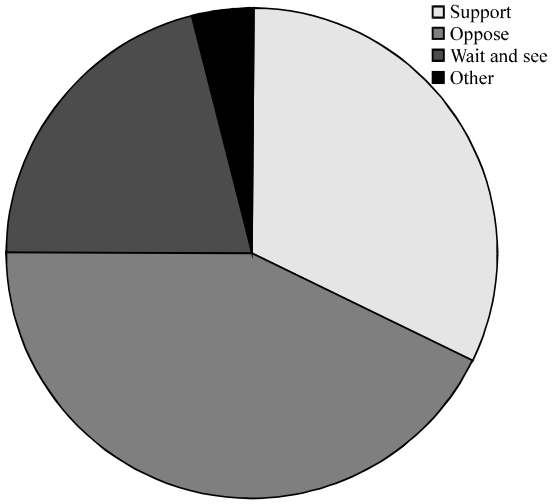

Fig. 3: Result of polarity analysis

Use the polarity calculation method based on natural language processing techniques to analyze the comments, the output result of polarity analysis is as in Fig. 3.

CONCLUSION

Based on the analysis of domestic and foreign development and current status of topic tracking and opinion mining, this paper proposes a TDT algorithm based on self-adjusting VSM and an opinion mining method based on comments and presents the overall design of an opinion mining system ,develops this system based on microblog platform. The results of the experiment demonstrate the feasibility and effectiveness of TDT algorithm based on self-adjusting VSM and polarity calculation method based on natural language processing techniques. In the future, we will consider using natural language processing technique to adjust text vectors for topic tracking. Because of the complexity of the Chinese language, deeper approach research of polarity analysis is needed to improve the performance of our public opinion mining method.

ACKNOWLEDGMENTS

We thank National Natural Science Foundation of China (71171148 and 61103069), Shanghai Committee of Science and Technology (11DZ1501703 and 11dz1210601), Innovation Program of Shanghai Municipal Education Commission (13YZ052) and National High-Tech Research and Development Plan of China (863)(2012AA062203) for supports.

REFERENCES

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Machine Learn., 20: 273-297.

Jia, Y.S., J. Min and Y.Q. Zhou et al., 2011. The research of Chinese text classification techniques based on machine learning. Comput. Knowl. Technol., 21: 5194-5196.

Kim, W., J. Ryu, K.I. Kim and U.M. Kim, 2009. A method for opinion mining of product reviews using association rules. Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, November 24-26, 2009, Seoul, Korea, pp: 270-274.

Li, Z., B. Wang, M. Li and W. Ma, 2005. A probabilistic model for retrospective news event detection. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 2005, Salvador, Brazil, pp: 106-113.

Liu, Y., 2007. Overview of Network Public Opinion Research. Tianjin People's Publishing House, Beijing, China.

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical methods in natural language processing, Volume 10, July 6-7, 2002, Philadelphia, PA., USA., pp: 79-86.

Vapnik, V., 1995. The Nature of the Statistical Learning Theory. Springer-Verlag, New York, USA.

Yamron, J.P., S. Knecht and P. van Mulbregt, 2000. Dragon's tracking and detection systems for the TDT2000 evaluation. Proceedings of the Broadcast News Transcription and Understanding Workshop, February 28-March 3, 2000, Herndon, VA , USA., pp: 75-79.

Zhao, L. and C. Li, 2009. Ontology based opinion mining for movie reviews. Proceedings of the 3rd International Knowledge Science, Engineering and Management Conference, November 25-27, 2009, Vienna, Austria, pp: 204-214.

Zhu, Y.L., J. Min and Y.Q. Zhou, 2006. How Net based lexical semantic orientation calculation. J. Chinese Inform., 20: 14-20.