

Polarity Dictionary-based network public opinion polarity research

Chun Xu

School of Management
Wuhan University of Technology
WuHan, China
chunerbao@163.com

Abstract—In this paper, based on the polarity of the text on the basis of dictionary preference analysis techniques to explore the new, try to understand the technology, natural language text analysis technology into the tendency of public opinion analysis.

Keywords—public opinion polarity; polarity dictionary; polarity calculatet

I. INTRODUCTION

July 16, 2009, China Internet Network Information Center (CNNIC) announced the “24th China Internet Development Statistics Report”, in Beijing. The report shows that as at the end of June 2009, the scale of Chinese netizens reached 338 million, compared with the end of 2008 increased 13.4%, six months increased by 40 million; while the size of broadband Internet users reached 320 million take up the total number of Internet users 94.3%, higher than the end of 2008 increased by 3.7%.

February 28, 2009, Wen Jiabao exchange with netizens online, the incident is not only a reflection of this signal that the national leaders pay attention to the Internet, the most important is that as the Internet continues to develop in-depth, the trend that the network public opinion will gradually replace the important position the original paper-based media, and its influence is limitless. With 3G era, the channels of people to express their views of the current phenomena of social and practical issues and the diversification of communication ways have become more smooth and so fast, such as: news and reviews, search engine, BBS, and online community, current affairs forum, blog, wiki, instant messaging software, Email and SMS, etc, truly achieved 3W (no matter when and where you want to express which views can be realized immediately)[1].

Simultaneously, the network public opinion the most timely and most truly and accurately reflect the public opinion trends at all levels of society, it will become a social barometer of public opinion trends, and it's influence on society will growing. In the 16th National Congress of Communist Party of China, Party General Secretary Hu Jintao's report pointed out how to strengthen the capacity of the network public opinion and put forward three specific requirements, namely “how to channel the network of public opinion, how to response to a online hot topic timely, how to improve the ability of dispose the network emergencies”; June 20, 2008, Hu Jintao inspected the People's Daily and the People's Network, also noted: “The Internet has become the hub of ideological, cultural and information, an amplifier

of public opinion, we must fully understand the impact of Internet as the representative of the social new media”; Early in 2009, the restricted reference materials “Network of public opinion”, supply for the party and government leaders at all levels to read was published. The above signals show that the network public opinion has been attached great importance to the national manager. Analysis, research and process the network public opinion has become a problem demanding prompt solution. This article tries to build a polarity dictionary to solve the question that how to determine the polarity of the network public opinion text.

II. BACKGROUND

Many countries in the world emphasized on hotspots information discovery, network information dissemination mechanism. just like that key technologies of research. The current network public opinion on the hotspot information study are the following two directions: first, based on natural language processing techniques—word frequency statistical methods of research concerning the technology to the identification of unknown words, in English and Chinese word segmentation. Towards the hot topics for the rapid discovery of the problem, the existing measure index are Shared Goals based on BBS, Mutual information, Shared Context and other metrics proposed to use the multi-dimensional vector measure the activity of topic, which names topic activity measurement method. However, this approach can not represent a good situation in the circumstances that at a large number of topics which is not based on historical information, rapidly discover the sensitive discuss points, with a longer time lag or a higher computational complexity [2]. The second is from the perspective of data mining information on the discovery of hot spots, using the characteristics of a complex network to do classification and aggregation towards information, the process is much more complex.

In recent years, in the field of information processing there are a lot of software products in domestic, but the most remarkable are the product of Peking University Founder Technology Research Institute which names Founder Zhi Si-assisted decision support system for early warning of public opinion.[3]. This product has successfully achieved a massive public opinion against the Internet, automated monitoring real-time and effectively analysis to help the government departments to solve the traditional manual way to monitor the implementation of the difficulties of public opinion for the promotion of the strengthening of internet information monitoring, organize all forms forces to

finish the information and in-depth analysis, response to emergency network public events, have complete control of social conditions and have played a certain role in promoting the public opinion.

In addition to Founder's that product, the domestic in information processing technology, and natural language understanding technology and research in the field or at the international leading position. With regard to the many of the results of studies about Chinese natural language processing can be found at the CAS Institute of Computing Technology of the "CNLP Chinese natural language processing open platform" and found these resources reporting system has a great help for monitoring public opinion, WEB design and development [4].

However, there are still inadequate in hot information discovery algorithm of public opinion, only use the Chinese text clustering algorithm, but no specific hot spot discovery algorithm. For the analysis of public opinion tendency yet exist no mature product, need further development.

III. POLARITY DICTIONARY

For information on public opinion hot tendentious judgments, after making Chinese word segmentation and Chinese syntactic analysis, was still not able to fully realize the tendency to judge public opinion, the followed work , the most important step is to bias analysis [5]. This design method of this paper is using polar Dictionary to achieve the ultimate judge of the bias about the hot spot information of public opinion. The expected processing flow chart is as follows:

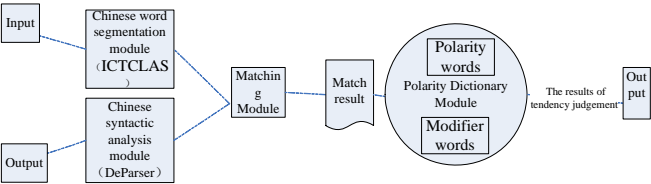


Figure 1. flowage structure

The polarity dictionary uses all of the corpus resources to built the Chinese polarity lexicon, the work for construct polarity dictionary includes of polarity words module and modifier words module [6]. Polarity words module primarily singled out one of the polarity of the word from corpus, and assure the characteristic of polarity word such as pos, definitions and their polarity values (according to appraise meaning and intensity level set as (-2,2), commendatory term polarity is positive, strong polarity value of 2, a neutral word polarity is 0, derogatory term polarity is negative, strong negative value of -2) for classification (including the field of polar and dynamic polarity). But the function of modifier is modify the polarity word to enhance or diminish their polarity, forward or reverse modification (setting the intensity level of the qualifier value is a strong positive,

positive pole intensity is 2, reverse strong as -1, reverse strong as -2). Matching the Chinese word segmentation and word analysis the results of Chinese syntactic, and input the results into polarity dictionary module polarity , matching them and then gradually to judge them, then come to the polarity value of each polarity word, each sentence, and finally use statistical algorithm to determine the bias of hot spot information public opinion [7].

In accordance with ontology classes, attributes, instances such a clear framework for building polarity dictionary ontology model[8]. For the position of the compliment and derogatory of polarity word, polarity value, add a relatively clear instance all have a clear description. Polarity dictionary ontology construction model is expected as follows:

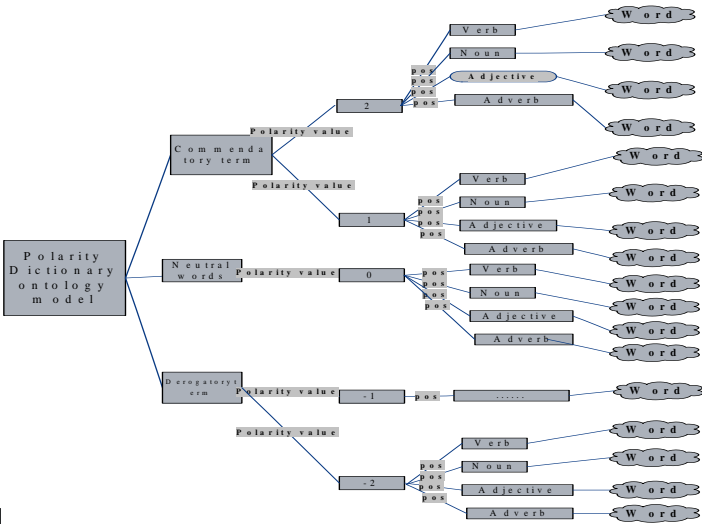


Figure 2. Polarity dictionary ontology model

The modifier is a further determine for the polarity word module, such as the polarity word "good", the polarity is compliment, according to the approach in this paper the polarity of the value of the word is set as 1, but if "good "to add after modifier" very "or" no ", the results of the word's polarity would be changed. This article was designed a simple algorithm to determine the polarity of text, the main idea is: use of Chinese word segmentation results will be a very important statements for initial position of the polarity word and the polarity value of the word, syntactic analysis do some corresponding description for the polarity of modifier words, and finally calculate the strength level of the polarity words which with a modifier word [9].

IV.POLARITY ALGORITHM

This paper designed a simple algorithm to judge the polarity tendency of a public opinion text , just as the following(the assumption is that X is a polarity word in the text):

- (1) If the polarity value of X is 1or -1.

- A. If the modifier strength value is 1, the polarity value of X in this text will *1.
- B. If the modifier strength value is 2, the polarity value of X in this text will *2.
- C. If the modifier strength value is -1, the polarity value of X in this text will *-1.
- D. If the modifier strength value is -2, the polarity value of X in this text will *-2.

(2) If the polarity value of X is 2or -2.

- A. If the modifier strength value is 1, the polarity value of X in this text will *1.
- B. If the modifier strength value is 2, the polarity value of X in this text will *1.
- C. If the modifier strength value is -1, the polarity value of X in this text will *-1.

In this paper, use DeParser to do syntactic analysis for sentences. Use of HIT "language technology platform (LTP) 2.0 beta version of" does the experiment [10].

(1) The test version also has a word segmentation and part of speech tagging feature, the results are as follows:

我们/r 不/d 相信/v 自己/r 经历/v 了/u 网络/n 民主/a ,/wvp 就/d 像/v 我
们/r 不/d 会/v 愚蠢/a 地/u 相信/v 经历/v 了/u 娱乐/v 选秀/n 就/d 经历/v
了/u 民主/a 一样/a ,/wvp 民主/a 不/d 是/v 有人/r 为/p 我们/r 做主/v ,/wvp
而是/c 人们/n 能/v 为/p 自己/r 做主/v ,/wvp 因此/c ,/wvp 人们/n 保留/v 对
/p 现实/n 不/d 妥协/v 的/u 真相/n 要求/n ,/wvp 人们/n 一再/d 要求/v 得到
/v 足以/d 相信/v 的/u 真相/n 为止/v ,/wvp 并/c 保留/v 质疑/v 一切/r 不/d
正当/a 的/u 权利/n ,/wvp 如果/c "wvp 躲猫猫/n 猫/n "wvp 不/d 想/v 成为/v
"wvp 莫须有/a "wvp ,/wvp 如果/c "wvp 躲猫猫/n 猫/n "wvp 不/d 想/v 成为/v
/v 千秋/nt 笑柄/n ,/wvp 那么/c 说/v 出/v 李荏明/nh 死/v 于/p "wvp 躲猫猫/n
猫/n "wvp 的/u 执法/v 机构/n ,/wvp 必须/d 证明/v "wvp 躲猫猫/n 猫/n "wvp
的/u 清白/a ,/wvp 公布/v 一/m 段/q 监控/v 录像/n 并非/v 难事/n ,/wvp 给/p
造成/v 李荏明/nh 死亡/v 的/u 曹华勇/nh 一/m 次/q 公开/a 公正/a 的/u 审理
/v 也/d 并非/v 难事/n ,/wvp "wvp 躲猫猫/n 猫/n "wvp 事件/n 网民/nh 调查/v
是/v 不/d 是/v 做秀/v 的/u 现代/nt 愚民/n ,/wvp 亦/d 将/d 在/p 彻底/a
的/u 公开/a 透明/a 中/nd 得到/v 证实/v ,/wvp

Figure 3. Segmentation Results

From the above results we can see, the system has done a good segmentation for the name, but for "躲猫猫" has not identified, the defect need to be solved depend on the unknown word processing of the hot spot information discovery of public opinion.

(2) Syntactic analysis

The system after segmenting is divided into three sentences, the follows will respectively analysis the polarity of its syntactic structure.

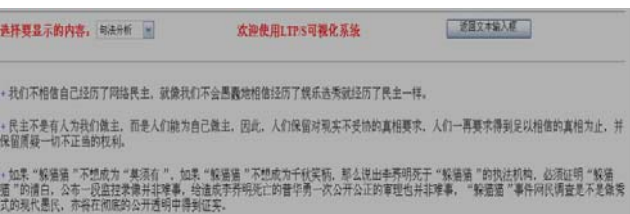


Figure 4. LTP Parsing

Select the first sentence to do the syntactic analysis of the polarity analysis.

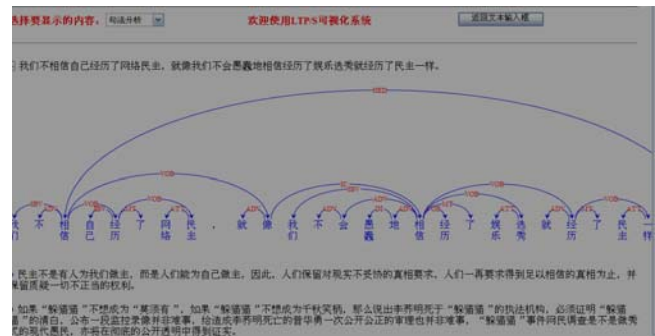


Figure 5. LTP Syntactic analysis

Although the screenshot itself is not complete, we can get the relationship between words in the sentence clearly from the figure. And SBV (subject-verb) means that subject-predicate relationship, ADV (adverbial) means adverbial-center relationship, VOB (verb-object) as a verb-object relations, DI means "地" the relationship between words, IC (independent-clause) as an independent clause , HED (head) as the core, MT (mood-tense) as the voice structure, ATT (attribute) means the attribute-center relationship [11]. According to this syntactic analysis results can be good to determine the relationship between modifier and the polar and laid a good foundation for analyzing the hot spot information given preference of public opinion.

(3) Polarity words and their values determination

This paper only just established the polarity dictionary model, so in this experiment for the polarity word and its nature determination are all adopted artificial selection manner. The results are as follows:

believe 1 democracy 1 silly -1 entertainment 1
compromise -1 truth 1 questioning -1 valid 1
right 1 groundless -2 laughingstock -2
die -1 cleanly 1 twister -1 death -1
public 1 fair 1 Showbiz-style -1
ignorant people -1 transparent 1

Dependencies:

bu xiang xin(ADV) bu hui xiang xin(ADV) bu
shi(ADV) shi zuo zhu(VOB) wei
zuo zhu(ADV)
neng zuo zhu(ADV) bu tuo xie(ADV) bu
zheng dang(ADV) bu xiang (ADV) xiang mo
xu you(vv)
bu xiang(ADV) xiang cheng wei(CMP)
cheng wei xiao bing(VOB) bing fei
nan shi(ATT)

(4) Bias analysis of the text

According to a simple algorithm this paper designed computing (polarity value * modified intensity value):

bu xiang xin1*-1=-1 democracy 1 bu hui yu chun-1*-1=1
 1=1 bu hui xiang xin 1*-1=-1 entertainment 1
 democracy 1 democracy 1 bu tuo xie-1*-1=1 truth 1 believe 1 truth 1
 questioning -1
 bu zheng dang-1*-1=-1 right 1 bu xiang mo
 xu you-1*-2=2 bu xiang xiao bing-1*-2=2 die
 -1

cleanly 1 bing fei nan shi-1*-1=1
 death -1 public 1 fair 1
 bing fei nan shi-1*-1=1 bu shi zuo xiu shi-1*-2=2
 ignorant people -1 public 1 transparent 1

Calculation results:

Commendatory number: 20 Derogatory Number: 7

Compliment value: 23 Derogatory Value:-7

$$\frac{23}{23+|-7|} \approx 77\%$$

Compliment ratio:

$$\frac{|-7|}{23+|-7|} \approx 23\%$$

Derogatory ratio:

From the above results can be seen, although the analysis is very rough, but the hot spots tendentious of public opinion still can be drawn [12]. The tendency of this text is commendatory. In fact from the people's understanding of point of view, this paragraph was also a positive means more dense, not means do a simple and crude assessment for the incident "躲猫猫", but an introduction of positive thinking towards the text.

CONCLUSIONS

Given the current Internet environment becomes increasingly complex and changeable, the number of Internet users has increased every year, people's lives are increasingly affected by the network, so quickly to discover the network hot spot of public opinion information to respond to network incidents has important practical significance [13].

This paper did preliminary studies about discovery techniques of hot spot on the network information of public opinion, and introduced the text tendentious analysis technique to analyze and research the hot tendentious of the information which has been found, a more comprehensive

exposition of the design and implementation issues, in particular, the text focuses on the application of analytical techniques of the text tendency and it's to be achieved. Of course, for the public opinion polarity of the algorithm is also the need a further debugging.

ACKNOWLEDGMENT

This paper is supported by National Natural Science Fund 70972094, National Natural Science Fund 70572079 and National Science Support Plan 2006BAH02A08.

REFERENCES

- [1] China Internet Information Center. The twenty-fourth of China Internet Development Statistics Report [J]. China's Internet Survey Report 2009 (7)
- [2] Markus Maier, Matthias Hein, Ulrike von Luxburg. Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters [J]. Theoretical Computer Science, 2009 (410):1749-1764.
- [3] Huifeng Tang, Songbo Tan, Xueqi Cheng. A survey on sentiment detection of reviews [J]. Expert Systems with Applications, 2009(36):10760-10773.
- [4] Li Xu, Zhenrong Zhuge, Jie Yang. Artificial Emotion and Its Recognition, Modeling and Applications: An Overview [J]. Proceedings of the 5th World Congress on Intelligent Control and Automation, June 15-19, 2004, Hang Zhou. P. R. China.
- [5] Qiang Ye, Ziqiong Zhang, Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches [J]. Expert Systems with Applications, 2009(36): 6527-6535.
- [6] Kristof Coussement, Dirk Van den Poel. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers [J]. Expert Systems with Applications, 2009(36): 6127-6134.
- [7] Chao Zhang. Research on the application of text orientation analysis in the public opinion monitoring system[D]. Beijing University of Posts and Telecommunications, 2008.
- [8] Chai Yumei, Xiong Delan, Zan Hongying. Research on Web Text Appraisable Classification[J]. Computer Engineering, 2006, 32 (17): 89-91
- [9] China Internet Information Center. The twenty-fourth of China Internet Development Statistics Report [J]. China's Internet Survey Report 2009 (7)
- [10] Li Juan, Zhang Quan, JIA Ning. Semantic orientation identification for Chinese opinion terms. Computer Engineering and Applications, 2009,45(2):131-133.
- [11] Xu Lin-hong, Lin Hong-fei, Yang Zhi-hao. Text Orientation Identification Based on Semantic Comprehension[J]. Journal of Chinese Information Processing, 2007(1).
- [12] Yu H., Hatzivassiloglou V. Toward Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences[J]. In Proceedings of EMNLP, 2003: 129-136.
- [13] Thruney, Peter, Littman Michael. Measuring Praise and Celeriticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003,21(4):315-346.