Application of Internet segmentation research based on Natural Language Processing technology in enterprise public opinion risk monitoring

IOP Conf. Series: Journal of Physics: Conf. Series **1187** (2019) 042007 doi:10.1088/1742-6596/1187/4/042007

Application of Internet segmentation research based on Natural Language Processing technology in enterprise public opinion risk monitoring

Di Liu[1, a], Jiangwen Su[2], Lihua Song[2] and Zhen Qiu[1, b]

[1]State Grid Information and Telecommunication Group, Beijing 100000, China;

[2] Fujian Yirong Information Technology CO., Ltd, Fuzhou 350000, China.

[a]liudi@sgitg.sgcc.com.cn, [b]qiuzhen@sgitg.sgcc.com.cn

Abstract: With the advent of the mobile Internet era, the network has become a distribution center of various information such as media, entertainment, sports, economy, politics and so on. A large amount of information is generated and disappeared on the network every day. How to effectively extract and identify the relevant data, and judge and analyze them is an important part of the corporate public opinion control. This paper uses natural language processing technology to study the word segmentation of text information on the network, and applies it to the risk detection of corporate public opinion.

1. current situation of public opinion research

Public opinion analysis is a process of deep thinking, analysis and Research on the public opinion aiming at this problem and getting relevant conclusions according to the needs of specific problems. Public opinion analysis is a subject based on many fields such as language and social communication, news communication and so on. It is different from the early paper media era. After the arrival of the Internet era, the new information modes of Web pages, news portals, short videos, microblogs and other fast-food formats have posed great challenges to public opinion analysis in the past. At the same time, because of mobile interaction. With the advent of the Internet era, the massive information and communication modes have been difficult to cover the previous public opinion analysis models. The data on the network end grow exponentially and erupt. It is very likely that an event will be completed, originated, fermented and erupted in a day or two. All these put forward new choices for public opinion analysis.

1) Definition, origin and development of Internet public opinion in China

Network public opinion refers to the network public opinion which is popular on the Internet and has different views on social issues. It is a form of expression of social public opinion. It is a strong influence and tendentious opinion of the public on some hot and focus issues in real life through the Internet. Network public opinion is a collection of netizens' emotions, attitudes, opinions, opinions, expressions, dissemination and interaction, as well as follow-up influence, with the network as the carrier and events as the core.

Network public opinion refers to the social and political attitudes, beliefs and values that people have towards public issues and social managers through the network around the occurrence, development and changes of intermediary social events in a certain social space. It is the sum of the beliefs, attitudes, opinions and emotions expressed by more people about various phenomena and

problems in society. The formation of Internet public opinion is rapid and has a great impact on society. With the rapid development of the Internet in the world, the network media has been recognized as the "fourth media" after newspapers, radio and television, and the network has become one of the main carriers reflecting public opinion.

Internet public opinion is a reflection of social public opinion in the Internet space and a direct reflection of social public opinion. Traditional social public opinion exists in the folk, in the public's ideas and daily comments on the streets and lanes. The former is difficult to capture, while the latter is fleeting. Public opinion can only be obtained through open and secret visits, public opinion surveys and other means. The acquisition efficiency is low, the sample is small and easy to flow biased, and the cost is enormous. Big. With the development of the Internet, the public often express their opinions in the way of informationization. The network public opinion can be easily accessed by means of automatic grasping technology of Turing public opinion network, which is efficient, information fidelity and full coverage.

2) Research methods of public opinion

In recent years, China has made great efforts to use technical means to dig and analyze a large amount of network public opinion information in depth, so as to quickly compile public opinion information, thus replacing the complicated work of manual reading and analysis of network public opinion information. The key technologies related to Internet public opinion are summed up as two types: monomer technology and systematization technology.

(1) Network public opinion collection and extraction technology: network public opinion is mainly formed and disseminated through news, forum/BBS, blog, instant messaging software and other channels. The carriers of these channels are mainly dynamic web pages, which carry loose structured information, making effective extraction of public opinion information very difficult. Mei Xue et al. (2007) realized the extraction and integration of dynamic web page data to a certain extent through the method of automatic generation of web page information extraction Wrapper, which has a certain processing accuracy and extraction efficiency.

(2) Network public opinion topic discovery and tracking technology: Internet users discuss a wide range of topics, covering all aspects of society, how to find hot and sensitive topics from the mass of information, and track its trend change has become a hot research topic. Early research ideas of Alan James, J. Allan, G. Hulten, Qiaozhu Mei and others are based on text clustering, that is, the keywords of text are the characteristics of text. Although this method can aggregate text under a large category of topics, it does not guarantee the readability and accuracy of topics. Duan Jianguo et al. (2007) improved this idea and realized topic discovery and tracking: transforming text clustering into topic feature clustering, and reorganizing and utilizing language text information flow according to events.

(3) Network public opinion tendentiousness analysis technology: through tendentiousness analysis, we can make clear the subjective reflection of the feelings, attitudes, views, positions and intentions of network communicators. For example, Sina's "news mood ranking" divides the mood of users when they read news comments into eight levels as shown in Figure 2-1. The analysis of the tendency of public opinion text is actually an attempt to use computer to achieve the goal of extracting the emotional direction of the author of the text according to the content of the text. Tang Huifeng, Xu Linhong, Li Yanling and others (2007) devoted themselves to the tendentiousness analysis technology of online public opinion texts: by judging the characteristics and types of tendentious feature words in the network environment, and making the identification and annotation of the tone polarity, they constructed an Internet-oriented tendentious mood dictionary and constructed a certain scale of standard data.

(4) Multi-document automatic summarization technology: news, posts, blog posts and other pages contain spam information. Multi-document automatic summarization technology can filter the content of pages and extract summary information to facilitate query and retrieval. To a certain extent, researchers have realized the automatic generation of messages from network public opinion information, and can browse and retrieve information through browsers.r further research on Chinese orientation analysis.

3) the application of computer and machine learning in public opinion.

It is a long-standing pursuit of people to communicate with computers in natural language. Because it not only has obvious practical significance, but also has important theoretical significance: people can use their most accustomed language to use computers, without spending a lot of time and energy to learn various computer languages which are not very natural and habitual; people can also further understand human language ability and intelligence through it. The mechanism.

At the same time, computer is used to analyze and calculate large-scale complex systems, deal with massive text technology, and use statistical and artificial intelligence methods to replace manual text classification and discrimination. At the same time, the analysis and processing based on historical data can quickly complete the judgment and analysis of regional and global hot words in a short time in the future, so as to achieve the role of prevention and control.

2. application of machine learning algorithm in natural language understanding

1) machine learning algorithm text processing technology

Similar to traditional machine learning algorithm, the main modes of text processing technology based on text machine learning algorithm are word segmentation, extraction, discrimination and vectorization.

A Chinese text is a string consisting of Chinese characters (including punctuation marks). Words can form words, phrases can form phrases, phrases can form sentences, and then some sentences can form paragraphs, sections, chapters and chapters. No matter at all levels mentioned above: characters (words), words, phrases, sentences, paragraph,... Or there are ambiguities and polysemy in the transition from the next level to the next level, that is, a string of the same form can be understood as different word strings, phrase strings and so on in different scenarios or contexts, and has different meanings. Generally speaking, most of them can be solved according to the corresponding context and scene. That is to say, in general, there is no ambiguity. This is why we do not feel ambiguity in natural language and can communicate correctly with natural language.

Above all, a Chinese text or a string of Chinese characters (including punctuation marks, etc.) may have multiple meanings. It is the main difficulty and obstacle in natural language understanding. Conversely, an identical or similar meaning can also be represented by multiple Chinese texts or multiple Chinese character strings. And to quantify the data sets.

The process of selecting a subset of related features from a given set of features is called feature selection. Feature selection is from the feature set T={t_1,... In t_s}, we choose a real subset T '={t_1,... T_ (s')} satisfies (s s s). Among them, s is the size of the original feature set, and s^ 'is the size of the selected feature set. The criterion of selection is that feature selection can effectively improve the accuracy of text. Selection does not change the nature of the original feature space, but only chooses some important features from the original feature space to form a new low-dimensional space. Text feature selection can effectively reduce the dimension of text representation.

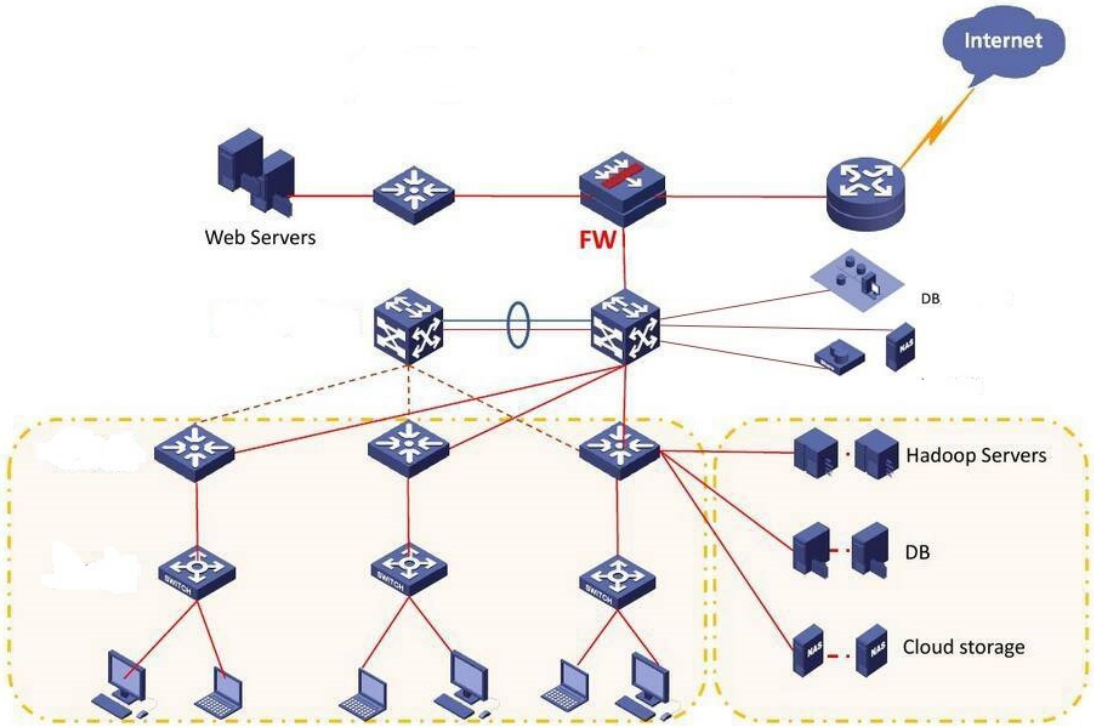

Fig. 1. collection and processing of data information

2) implementation of machine learning algorithm for natural language understanding.

(1) using supervised learning to generate a historical data training model for predicting the type of text.

(2) Mainly delete duplicate data, correct or delete erroneous and invalid data, and check the consistency of data. For example, if the length of text is less than 13, it is meaningless to delete it.

(3) set a set of categories to classify words according to the sample set manually.

(4) using Classified Thesaurus to deal with the processed text information two times.

3) the application and implementation of thesaurus technology.

Word segmentation technology is a technology that search engines use various matching methods to segment keywords according to user's keyword string after query processing for keyword string submitted by users.

The main methods are: string matching method of word segmentation; word segmentation; statistical word segmentation.

Based on these different word segmentation methods, the extracted text constitutes a database and is updated according to the text at any time. The comparative sample database is the word segmentation database. The specific implementation algorithm is not described in detail here, and can refer to the relevant literature at home and abroad.

3. data acquisition method of massive information processing foundation and web crawler method

1) massive information processing technology

The so-called massive information processing means that the amount of data is too large to be solved quickly in a relatively short time and can not be loaded into memory at one time. Based on these problems, many algorithms have been put forward to solve this problem.

In terms of time complexity, we can use ingenious algorithms with appropriate data structures, such as Bloom filter/Hash/bit-map/heap/database or inverted index/trie tree. In terms of spatial complexity, divide and conquer /hash mapping.

The basic methods of massive data processing are summarized as follows:

Divide and conquer /hash mappings + hash statistics + heap / fast / merge sort;

(1) double barrel division;

(2) Bloom filter/Bitmap;

(3) Trie tree / database / inverted index;

(4) external sorting;

(5) Hadoop/Mapreduce for distributed processing.

2) web crawler technology

The object of general search engine is Internet pages. At present, the number of Internet pages has reached 10 billion. So the first problem faced by search engine is how to design an efficient download system to transmit such a large amount of web pages data to the local area and form a mirror backup of Internet pages locally. Web crawler can play such a role to complete this arduous task. It is a key and basic component of search engine system.

Firstly, the crawler system carefully selects part of the web pages from the Internet pages, and takes the link addresses of these pages as seed URLs, and puts these seeds into the waiting URL queue. The crawler reads the waiting URL queue in turn, and parses the URLs through DNS, and converts the link addresses into the corresponding IP addresses of the Web server. Then give the relative path name to the web page downloader, which is responsible for the download of the page. For downloaded local pages, on the one hand, they are stored in the page library, waiting for subsequent processing such as indexing; on the other hand, the URLs of downloaded pages are placed in the crawled queue, which records the URLs of webpages that have been downloaded by the crawler system to avoid duplicate crawling of the system. For the newly downloaded Web pages, extract all the link information and check it in the downloaded URL queue. If it is found that the link has not been crawled, it will be placed at the end of the queue to be crawled, and the corresponding web pages of the URL will be downloaded in the subsequent crawl scheduling. In this way, the formation of a cycle until the URL queue to be grabbed is empty, which means that the crawler system will be able to grab all the pages have been grabbed, at this time completed a complete round of grabbing process.

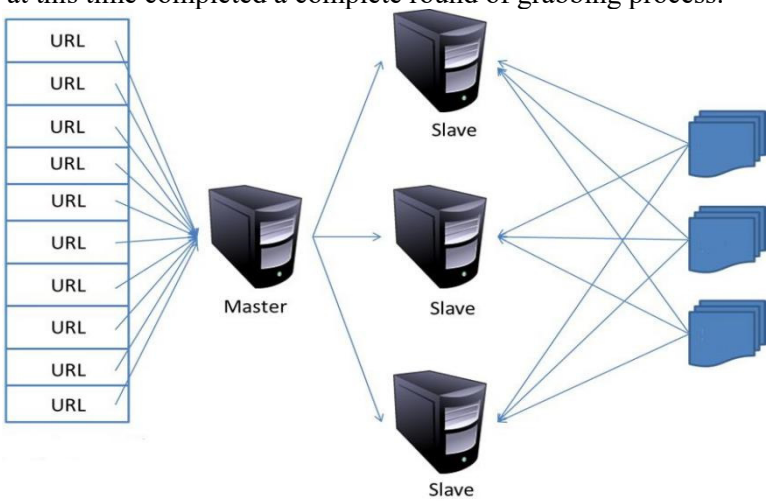

Fig. 2. schematic diagram of web crawler principle

4. the establishment of public opinion monitoring model.

1) public opinion text data analysis

Firstly, according to the latest five-year news reports on the group, two categories of positive news and negative news are selected. At the same time, the text of the group is pre-classified and classified according to the subsequent impact and duration. Finish the text sorting of the thesaurus.

The characteristic text file is segmented and the corresponding sub library is formed.

2) model establishment

By using the established word segmentation library and using the Jieba package of Python language to learn, a model of public opinion discrimination and correction is constructed.

The specific steps are as follows:

[1]. The jieba. cut method accepts three input parameters: a string requiring participle; a cut_all parameter to control whether full mode is used; and an HMM parameter to control whether HMM model is used.

[2]. The jieba.cut_for_search method accepts two parameters: a string requiring participle; and whether to use the HMM model. This method is suitable for search engine to build inverted index segmentation.

[3]. the strings for pending participles can be Unicode or UTF-8 strings and GBK strings. Note: it is not recommended to enter the GBK string directly, it may be unreasonably decoded into UTF-8.

[4]. The structure returned by jieba. cut and jieba. cut_for_search is an iterative generator that can use the for loop to obtain each word (unicode) after segmentation, or use the

[5].jieba.lcut and jieba.lcut_for_search return directly to list

[6]. Jieba. Tokenizer (dictionary = DEFAULT_DICT) creates a new custom word segmenter, which can be used to use different dictionaries at the same time. Jieba.dt is the default participle, and all global participle related functions are the mapping of the word segmentation device.

3) model screening

The initial data and the latest network text are used for preliminary learning, and the selected words in the lexicon are adjusted appropriately according to the results, and the segmentation strategies and corresponding algorithms are adjusted.

5. data learning and adjustment of models

1) public opinion text data collection

Python crawler technology is used to collect data on the main portal websites in our country, and it is textualized and saved as. TXT file format. And classify them according to different data sources.

2) data cleaning and cleaning

The collected data were first screened based on events, space and correlation degree. The obvious irrelevant data and text are eliminated in the first round. The remaining saves are processed in the next step and segmented and vectored.

3) adjustment analysis of learning strategies

According to python's learning curve and effect, the strategy and parameters are adjusted. At the same time, different word segmentation libraries are adopted based on the time limit. The results show that the algorithm convergence achieves the effect of text learning.

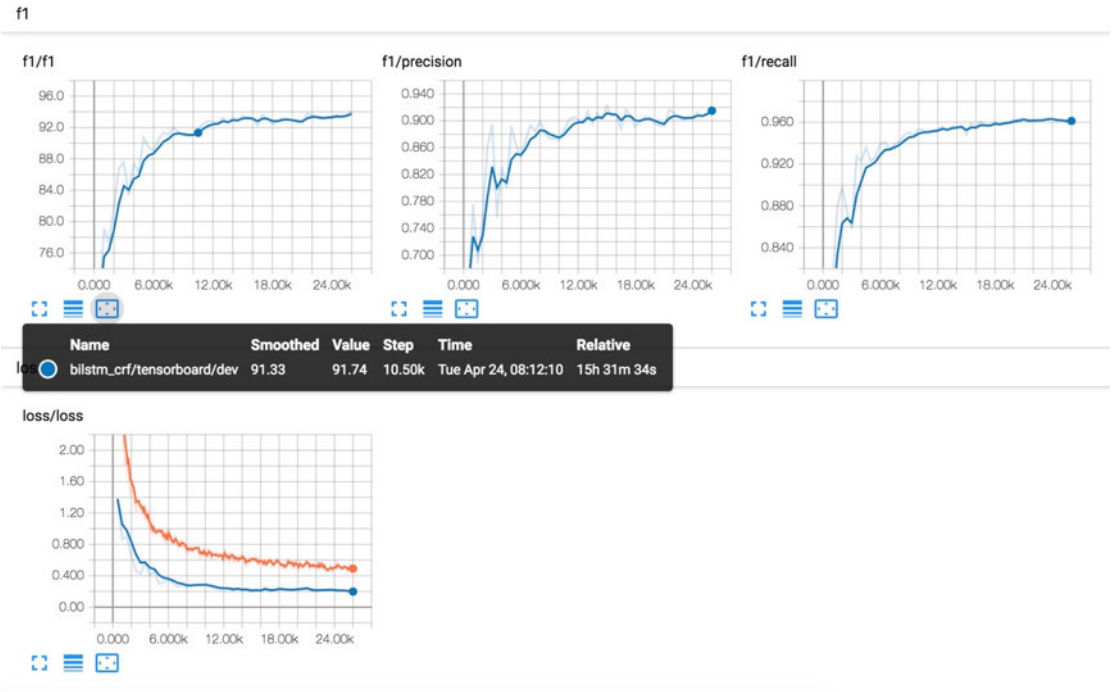

Fig. 3. convergence curve of machine learning algorithm for model

6. summarize the advantages and disadvantages and the main points to be improved.

Using the natural language processing technology of machine learning algorithm to detect and discriminate public opinion system is convenient and fast, and can achieve the purpose of public opinion monitoring. However, the timeliness of the model is relatively short, and the problem of long training time is to be solved. At the same time, sometimes the screening of invalid text is one of the bottlenecks of this technology. So effective segmentation algorithm is the focus of future research.

References

Acknowledgement: This work was supported by State Grid Technical Project (No. 52110418002W).

References:

[1] Zeng run hi. Network public opinion control mechanism research [J]. library and information work, 2009, 53 (18): 79-82.
[2] Zhu Yihua, Zhang Chaoqun, Zheng Dejun, et al. Research on Internet Public Opinion Management from the Perspective of Information Ecology [J]. Information Theory and Practice, 2013, 36 (11): 90-95.
[3] Guo Jianqiang, Zeng Wangfeng. Discussion on the Change of Network Public Opinion Management in the Age of Big Data [J]. Guangxi Social Science, 2015 (8): 145-149.
[4] Zhao Jinping, Zhang Xinyu. Public opinion management strategy for social security incidents [J]. news research guide, 2018 (9).
[5] Lin Yiou, Lei Hang, Li Xiaoyu, et al. Deep Learning in Natural Language Processing: Methods and Applications [J]. Journal of University of Electronic Science and Technology, 2017, 46 (6).
[6] Tian Dong, Zhang Xining. Realization of Weak Supervisory Knowledge Acquisition System Based on Natural Language Processing [J]. Foreign Electronic Measurement Technology, 2017, 36 (3): 60-63.
[7] Bei Chao, Hooper. The influence of language priori knowledge on natural language processing tasks of neural network models [J]. Chinese Journal of Information, 2017, 31 (6).
[8] Ma Yuchun, Song Hantao. Web Chinese text segmentation technology research [J]. computer applications, 2004, 24 (4): 134-135.

[9] Zhang Zhongyao, Ge Wancheng, Wang Liangyou, etc. Research and design of Chinese word segmentation technology based on MMSEG algorithm [J]. Information technology, 2016 (6): 17-20.
[10] Liu Xinliang, Yan Shanshan. Implementation and application of Chinese word segmentation based on Python [J]. Computer and Information Technology, 2008 (11): 85-88.
[11] Xu Xiao, Zhang Weizhe, Zhang Hongli, et al. WAN Distributed Web Crawler [J]. Journal of Software, 2010, 21 (5): 1067-1082.