Review on the technology of network public opinion monitoring system

Wenyi He[1*], Zhuoling Bai[2], Shiyuan Xu[3]

[1]*School of Public Administration, Central South University, Changsha Hunan 410083, China*

[2]*School of Public Administration, Central South University, Hunan Urban Professional College, Changsha Hunan 410083, China*

[3]*School of Software, Central South University, Changsha Hunan 410075, China*

Received 1 March 2014, www.cmnt.lv

Abstract

This paper introduced the technology related to the network public opinion monitoring system. According to the work flow of public opinion monitoring system, the paper summarizes the research status from three aspects of network information collection, pre-processing and analysis.

Keywords: public opinion monitoring, crawler, text classification

1 Introduction

In the modern society, the network has become a significant information spreading media. According to a relevant survey organized by CNNIC, the number of Chinese network users' totals to 632 million, among which 527 million are mobile phone users. The prevail rate of the network reached 46.9% by June 2014. The data fully displays the influence of the Internet. Unfortunately, the expansion of the Internet as an information spreading media has caused several problems. Network public opinion is one of the most typical embodiment. Due to the openness of the social network platform, network users would comment on daily hot news on different forums and microblogs, causing great social impact with rendering of network-public-opinion on tiny issues. In a sense, the network public opinion influences the development of society. It is of great importance to monitor network public opinion, guide opinions properly and effectively, and publish information in time. The network public opinion system can be organized into three parts, including the network information collection, pre-processing and opinion analysis. This paper studied the implementation techniques of these three divisions.

2 Information collection

At present, the medium of network public opinions include microblogs, BBS, forums and so on, which are basically dynamic webpages made up of tuple information. A majority of the current network public opinion monitoring systems collect information by the Meta-searching technique and the Web Crawler technique.

2.1 THE META-SEARCHING TECHNIQUE

Most of the users get information through search engine. However, the search engine fails to collect all the internet information and can only provide hundreds of searching results. Therefore, people start to study the Meta-searching based on the current search engine. The Meta-searching engine consists of three key techniques, processing retrieval request, calling retrieval interface and displaying searching results.

The Meta-searching engine works as it integrates multiple search engines and submits users' retrieval request to different search engines through interfaces. Several search libraries are searched simultaneously, duplicate information will be removed, fragment information merged, and the searching results are sorted and filtered [1]. Even though the Meta-searching engine is able to make up for the demerits of traditional search engines effectively, the information integration and sorting requires further study. Li Qinqin put forward personal search engine based on the interest of users. The search engine can extract the characteristics of users' personalized behavior and build an interest database. Search results will be integrated and displayed in reasonable orders with users' most interested information showed in the first few rows. Wu Weibin proposed a method based on the weight of members' search engine indices. Initial matrix is constructed by experts questionnaire survey, consistency is amended by correcting algorithm based upon Maximum Deviation Method. When the judgment matrix is screened, we may confirm the weight of members' search engine.

* ***Corresponding author's*** e-mail: wenhua1711@qq.com

2.2 THE WEB CRAWLER

The Web Crawler, also called the Web Spider, is a script program that will automatically extract web information based on certain rules. Traditional Web Crawler can execute from pre-custom-made URL links, download web pages in URL queue, recognize every hyperlinks in these pages and add these hyperlinks in the queue.

The Web Crawler technique divided into different research directions with the development of the Web Crawler technique. The current network public opinion monitoring system mostly adapts Web Crawler based on topic. This kind of Web Crawler will collect topic-related pages instead of non-related pages, which distinctly saves system resources and improves information collecting speed. The Web Crawler based on topic satisfies people's requirement of obtaining specific areas of information, and is also a popular direction of the current study on Web Crawler. However the problem of how to specify the program theme and improve the extent and precision requires further analysis. Bai Yuzhao analyzed multiple topic Crawlers and applied probabilistic model to calculate priority values of each URL in order to filter and sort URL. He designed a topic Crawler based on probabilistic model and settled the problem of strategy singularity [5].

The visiting strategies of the Web Crawlers are divided into three types:

1) Deep-prior search. Deep-prior strategy interpret links as tree structure, which can reach to enough depth of a certain links. But the program easily gets stuck [2].

2) Breadth-prior search. Breadth-prior strategy means that the program can only search the next page after the search of the current page is done. The information obtained can guarantee the relationship between links due to its high quality [3]. This strategy is often applied when information quality is demanded. On the other hand, the obtained pages include large amounts of unwanted information [4].

3) Best-first search. The Best-first search obtains similar and relevant pages based on re-designed web analysis algorithm. Only the web pages that passed the prediction and analysis of the algorithm will be visited. Unfortunately, the disadvantage of the Best-first strategy is that as a local search algorithm, it needs further improvement by combining with practical application. Otherwise, large amount of pages will be dropped on the grab path [5].

3 Preprocessing of information

After the information collecting work is done, the information obtained should be preprocessed to build a foundation for text mining work later. The information preprocessing techniques include the extraction of web text information, Chinese participle technology, the formal representation of text and so on.

3.1 EXTRACTION OF WEB TEXT INFORMATION

A complete page consists of multiple content pieces, including navigation, headline, main body, link, copyright, etc. Among all these information, the user mainly concerns on headline and main body. The rest of the contents shows little relevance to web page. The extraction of web text information is to extract pre-determined set of entities, relations and events, and record these information with structural representation [6].

At present, there are two ways to extract web text information:

1) Method based on templates. This method set targeted web pages that are randomly extracted as samples, and translate them into DOM tree to judge similarity of these pages. In this way, the web pages get clustered and then we can draw and modify identical types of template. This method ensures high precision, briefness and is easy to operate, enabling people to deploy quickly. Corresponding template should be configured on certain pages in advance, and pre-set demand information are extracted to precisely collect website information. The method is designed for processing small amount of information sources. Chen Zhi'ang worked out an automatic web text extraction algorithm based on templates to solve the problem of the web page noise and complexity of template generation for unstructured information.

2) Method based on page structure. This method automatically extract web page information by analyzing page structure and intelligent node and generate real-time extracting rules instead of pre-setting page template. The method based on page structure provides better versatility compared with the previous method, but shows low precision and highly operation difficulty. Duan Xiaoli [7] put forward an information extracting method based on web page information and HTML feature tag. On the basis of interpreting web pages into DOM tree, text blocks can be obtained according to DOM tree structure. Then the feature of noise information contained in the text block are analyzed to remove the noise information. The author tested the accuracy and recall rate of the method by experiments.

3.2 WORD SEGMENTATION

Word segmentation is the foundation of information retrieval and the premise of text mining. Word segmentation can be divided into three types [8]:

1) Segmentation based on character string match. Electronic dictionary is required to identify segmentation. The algorithm is simple since it only involves string comparison. The matching algorithm can be divided into positive match and negative match according to matching direction, or maximum match and minimum match according to matching length. Xiong Zhibin [9] proposed an improved Trie tree structure. Tree nodes record the location of character strings and words. Child nodes

adopted Hash Lookup Mechanism and optimized the positive maximum matching algorithm on Chinese word segmentation. The effectiveness is tested through experiments. In a mature analyzing system, the segmentation based on character string match is preliminary processing. Other processing methods are required to cooperate with it for further application. Wang Xigang [10] put forward a bilateral matching algorithm based on positive and negative maximum matching, which improves the accuracy of segmentation to certain extent.

2) Understanding-based segmentation. This segmentation is constructed by three pars including word segmentation subsystem, object semantic analysis subsystem and general control. Machines are applied to simulate people's understanding towards language, making it highly capable of ambiguity and new word recognition. Complete and accurate rules are required to help the machine understand sentences. However, the algorithm is so complex that no mature methods are available on the market. Since it's difficult to translate voice messages into information that machines can read directly due to the complexity and conceptuality of Chinese language, segmentation based on understanding remains at research stage.

3) Statistics-based segmentation. This method requires training set construction, which provide foundation of word segment model for statistic machine. Then the model can be applied on text for segmentation. Unfortunately, the statistic method tends to extract words of highly frequency that fail to form phrases. Training set is required to count words and ensure the precision of segmentation. The algorithm is commonly applied despite of its complexity.

Besides these three algorithm, experts put forward new segmentation methods. Yang Xiaojia [11] proposed segmentation for certain field based on ontology and syntax analyzing. By establishing ontology to analyze and search words intelligently, the demerit of semantic drop-out of traditional methods that failed to consider context information can be avoided. The effectiveness is tested by experiments. Hu Juxin [12] took the limitation of the current segmentation into consideration, and proposed a self-learning segmentation algorithm that combines character string segmentation and statistic segmentation. This method was successfully applied into duplicate checking system of science and technology projects. In general, mature segmentation systems are usually composite methods of several segmentations, such as the Massive Chinese intelligent word segmentation system of Tianjin Hailiang [13]. The segmentation precision of ICTCLAS [14] developed by Chinese Academy of Science has reached 98.45%.

3.3 FORMAL REPRESENTATION OF WEB PAGE TEXT

The formal representation of texts means texts can be classified and represented by computable features. The choice of words are required to determine the size of the feature and realize the formal representation of texts. The current representation methods include Boolean Model, Vector Space Model, Cluster Model, Probabilistic Model and Knowledge-based Model.

The Vector Space Model is widely applied these years with its practicability. This model extract features of texts based on the size of the features. The number of the processing words should be reduced as possible on the premise of ensuring text categorization accuracy. Only in this way the number of vector dimensions can be reduced. The evaluation functions that are commonly used at present include DF, IG, MI, CHI and so on. Li Jianlin [15] studied multiple text features extracting methods and proposed a PCA-CFEA, which can quickly achieve text features dimension reduction. Multiple feature extracting algorithm is applied to obtain typical features and filter atypical features. Texts are categorized by SVM classifier, thus improving the efficiency of text categorization.

4 Public opinions analysis technology

4.1 TEXT CATEGORIZATION

Text categorization means dividing information contained in the text-set according to pre-set classification system. The prevailing text categorization methods are as follows:

1) Bayesian method, which calculate the probability of texts appearing in different categories based on Bayes theorem. The demerits of Bayesian method is that it sets a strict standard on preliminary training samples, and text feature attribute are mutual independent.

2) K-nearest neighbors. This algorithm is commonly used in text categorization as a prevailing pattern recognition algorithm with simple principles. However, all the training samples need to be restored, making it inefficient when dealing with massive data classification. The effectiveness shows a strong relation to K value. Further solutions are required for the determination of K value.

3) SVM algorithm [16]. SVM algorithm applies vector model to represent training sample and solve quadratic programming problem based on model. This algorithm substitutes vector representations into the optimal classification function and categorize texts according to the evaluated values. The SVM algorithm requires fewer samples compared with the first two algorithms. When the amount of samples is limited, SVM algorithm provides optimal categorization results.

4.2 TEXT CLUSTERING

Text clustering is the extraction of important documents' characteristics. Identical documents are clustered to form themes through cluster algorithm. The current text clustering methods are as follows:

1) Hierarchy process. This method regards categorizations as hierarchical and assembles clustered texts into tree structure. It can be divided into condensation

and division methods according to organization forms. Typical hierarchy process methods include CURE, Chameleon and so on.

2) Partitioning. This method split text-set into several subsets based on certain rules. Each subsets represent a model. Typical partitioning methods include K-means algorithm [17], which can cope with massive data. This algorithm can realize clustering of large document effectively. However, the initial clustering centre is randomly selected, leading to a sub-optimal cluster results. The selection of the K value also has an effect on clustering result. Yang Shanlin proposed distance cost function to test validation of the optimal cluster number. Yang built corresponding mathematical model and designed a new K-value optimal algorithm. The optimal solutions and upper limit for k-value are given. The rationality of $k\max \le \sqrt{n}$ was proved in theory, while the effectiveness of optimal algorithm was testified.

3) Density-based method. Representative algorithm is DBSCAN [18], which is capable of finding classifications in arbitrary shapes. However, it requires large system memory. Feng Shaorong minimize the impact of global variable Eps by dividing data according to the idea of "divided and rule". Parallel processing method and dimension reduction techniques are involved to increase efficiency of dimension reduction and decrease system memory requirements. Incremental processing method is adopted to eliminate the affection of adding and deleting of data. The effectiveness of the new algorithm is testified by experiments.

4) Grid-based method. Typical algorithms include STING, CLINQUE.

5) Model-based method [19]. This kind of algorithm is mainly based on statistics and neural network.

4.3 THE APPLICATION OF PUBLIC OPINION ANALYSIS

An important application field of the text categorization is topic tracking. Topic tracking focus on themes that people are interested in and dig out information that belongs to this subject from massive news. Text categorization techniques are requested to test whether the obtained texts match the search theme and filter noise data [20]. Zhu Huanmin studied web link relations and content relevance and put forward the concept of link network diagram based on public opinion evolution. He proposed a method to calculate and update the relevancy degree between nodes and topics in the network diagrams, and designed a topic tracking method based on link network diagram. Experiments showed that this method not only ensures the accuracy, but also increase the recalling rate of topic tracking and dig out related web pages in links. Xia Chunyan proposed a topic-update-based algorithm GTKNN according to current vector model algorithms. The effectiveness of the algorithm is testified by experiments.

The application field of text clustering is topic detection. Existing web page information can be categorized into relevant topics using text clustering techniques. The cluster results of texts lead to new public opinion issues. When selecting cluster algorithm, the type of data, the purpose of clustering and practical application should be considered. Zheng Feiran put forward a method to discover news topics from microblogs by construct composite weight to quantify the popularity of words. Content relevance model was adopted to support quantity cluster algorithm. Experiments showed that this method can detect news topic effectively from massive information.

5 Conclusion

It's difficult to ignore the influence of network public opinions with the development of the Internet. Therefore, reliable and effective techniques are demanded to monitor public opinions and guide the direction of public opinions in a right track. This paper analyzed three steps in the monitoring of network public opinions, and concluded that the study on Web Crawlers, information preprocessing, text clustering and classification requires further improvement. A more effective algorithm should be designed. Among all the application fields of public opinion analyzing, topic detecting and tracking are hotspots for future improvement and share broad prospects for development

References

[1] Kumar N, Nath R 2013 A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering *International Journal of Computer (IJC)* 1 10
[2] Coppin B 2004 Artificial intelligence illuminated *Jones & Bartlett Learning Publishers Sudbury Massachusetts*
[3] Deo N 2004 Graph theory with applications to engineering and computer science *PHI Learning Pvt. Ltd India*
[4] Skiena S S 2008 The Algorithm design Manual *Second Edition Springer Verlag London Limited*
[5] Qin JL,Zhou YL,Chau M 2004 Building domain specific web collections far scientific digital libraries: a meta search enhanced focused cn Aing method 6
[6] Zeehne K L 1997 A literature survey on information extraction and text summarization *Carnegie Mellon University Pennsylvania*
[7] Duan X-l, Wang Y, Gu J 2012 Content extraction of theme pages based on body feature and page structure *Computer Engineering and Applications* **30** 48
[8] Cao Y-G, Cao Y-Z, Jin M-Z 2006 Information Retrieval Oriented Adaptive Chinese Word Segmentation System *Journal of Software* 3 17
[9] Xiong Z-b, Zhu J-f 2014 Forward Maximum Matching Algorithm Based on Improved Trie Tree Structure *Computer Applications and Software* 5
[10] Wang X-g, Wang Z, Chen H 2013 Research on Chinese Word Segmentation and Page rank in Regard to Search Engine *Computer Applications and Software* 9
[11] Yang X-j, Jiang W, Hao W-n 2008 Implementation of Field Word Segmentation Based on Ontology and Syntax Analysis *Computer Engineering* 23 34

[12] Hu J, Ju X-g 2013 Self-learning Algorithm for Chinese Word Segmentation in Scientific Research and Project Application of Duplicate Checking System *Bulletin of Science and Technology* 6
[13] Munteanu D 2010 Vector space model for document representation in information retrieval Annals of Dun area de Jos 1
[14] Salton G, Lesk M E 1968 Computer evaluation of indexing and text processing *Journal of the ACM* 1 15
[15] Li J-l 2013 Combination of feature extraction in text classification algorithm based on PCA *Application Research of Computers* 8 30
[16] Alham N K, Li M, Liu Y 2011 A Map Reduce-based distributed SVM algorithm for automatic image annotation *Computers & Mathematics with Applications* 7 62
[17] Jain A K 2010 Data clustering: 50 years beyond K-means *Pattern Recognition Letters* 8 31
[18] Ester M, Kriegel H P, Sander J 1996 A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise *Proc. of the 2nd International Conference on Knowledge Discovering in Databases and Data Mining Massachusetts AAAI Press USA*
[19] Mc Nicholas P D, Murphy T B 2010 Model-based clustering of microarray expression data via latent Gaussian mixture models *Bioinformatics* 21 26
[20] Aksoy C, Can F, Kocberber S 2012 Novelty detection for topic tracking *Journal of the American Society for Information Science and Technology* 4 63

Authors

Wenyi-He, 1973, Hangzhou City, Zhejiang Province, China.

Current position, grades: School of Public Administration, doctor.
University studies: Central South University.
Scientific interest: cultural industry, sports industry, internet communication and social issues, ethics, religion.
Publications: 7 papers, 6 books.

Zhuoling-Bai, 1985, Huaihua City, Hunan Province, China.

Current position, grades: School of Public Administration, doctor; Hunan Urban Professional College, lecturer.
University studies: Central South University.
Scientific interest: musicology, music information and music education.
Publications: 4 papers.

Shiyuan-Xu, 1990, Xianning City, Hubei Province, China.

Current position, grades: School of Software, master.
University studies: Central South University.
Scientific interest: web front-end.