Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 139 (2018) 293–298

Procedia Computer Science

www.elsevier.com/locate/procedia

The International Academy of Information Technology and Quantitative Management, the Peter Kiewit Institute, University of Nebraska

A Chinese Message Sensitive Words Filtering System based on DFA and Word2vec

Fei Wu[a]*, Yuxiang Cai[b]

[a]*Department of Scientific and Information, State Grid Fujian Electric Power Company, Fuzhou, 350003, Fujian, China*
[b] *State Grid Fujian Information & Telecommunication Company, Fuzhou, 350001, Fujian, China*

Abstract

In this paper, a Chinese message sensitive words filtering system applied in an instant messaging environment is proposed. Firstly, the message sentence is segmented, and the segmentation result is corrected by using the association algorithm based on information entropy and point mutual information. The traditional DFA algorithm is used to construct the dictionary tree for sensitive word recognition, which effectively improves the recognition speed. Secondly, on the basis of the completion of the recognition, the pre-trained word vector model is used to match the words in the sensitive words list and the word segmentation results, and the words with higher similarity with the sensitive words are added to the sensitive words list to achieve the expansion and improvement of the sensitive thesaurus.

Keywords: sensitive word filtering, DFA, word vector, information entropy, point mutual information

1. Introduction

In order to prevent adverse information related to political and administrative violations in the process of corporate office communication, research work is carried out on sensitive words filtering, and the harmful information is removed through informational means for information that violates national laws and regulations. This harmful information mainly includes three levels, namely national security, social security and enterprise security. National security is a high-voltage line, and each special project has a special person to follow up

* Corresponding author. Tel.: +86-591-87076759; fax: +86-591-87076759.
E-mail address: 28692589@qq.com.

10.1016/j.procs.2018.10.271

because the performance of the different logical information is different. Social security includes content information such as pornography, bans, and fraud. Enterprise security is mainly to protect corporate confidential information.

In the instant messaging system of power companies, the text information containing sensitive vocabulary is identified and monitored, reminded, filtered and analyzed. Establish a comprehensive corporate vocabulary of information violations, support the setting of custom sensitive words, cover sensitive information of various types of enterprises, and conduct effective information and information supervision. For the case of sensitive words, research tracking and positioning to the sending time, sender, number of transmissions, etc., to maintain a good image and interests of the company

2. Related Work

At present, Chinese sensitive word filtering, common methods are BM algorithm based on string sliding comparison [1], multi-pattern matching WM algorithm [2] and classical KMP algorithm [3]; decision-based tree-based deterministic automaton (Deterministic finite automaton, DFA) algorithm [4-5]; and conventional methods based on regular expressions [6]. The BM algorithm compares the pattern string with the text string and compares it word by word. Then, according to the position of the text string, the next character of the pattern string position is found to determine the corresponding offset size, which can effectively reduce the number of matches; WM adopts pattern diversity. Thoughts, in the domain name filtering application, the method of reverse ordering and optimizing the hash function for the domain name improves the processing efficiency. Since the recognition of sensitive words in IM system emphasizes real-time [7], the above method based on string comparison is difficult to ensure real-time recognition and filtering of sensitive words when the sensitive vocabulary is relatively large and the message is updated frequently [8-9]. The DFA algorithm is one of the most commonly used algorithms in IM systems. The complexity of the algorithm is independent of the size of the sensitive word database and only related to the length of the sensitive words and the length of the message. Through word segmentation, correction, and removal of stop words, etc., it is possible to effectively reduce the number of times of traversing messages and improve the recognition speed [10].

The update of the sensitive thesaurus is the process of obtaining the unregistered sensitive words from the message. The basis is to extract words from the messages that are similar to the existing sensitive words. The word vector technique can solve this problem very well [11]. In simple terms, word vector technology is to convert words into dense vectors, and for similar words, the corresponding word vectors are similar [12]. For the uncovered sensitive words encountered, after the sensitive lexicon is updated, it can be compatible by regularly training the new word vector model [13]. The word vector model can be trained twice, and only need to train some unregistered words [14].

3. Methodology

3.1. Overall framework

In this paper, on the basis of the traditional DFA algorithm to construct the dictionary tree for sensitive word filtering, the efficiency of the filtering system is obviously improved by correcting the result by segmenting the input sentence and introducing the correlation algorithm. In addition, based on the filtering, this paper also adds a collection module of sensitive words synonym based on word vector, which is used to dynamically update the sensitive word list. The overall framework of the sensitive word filtering system in this paper is shown in Fig. 1.

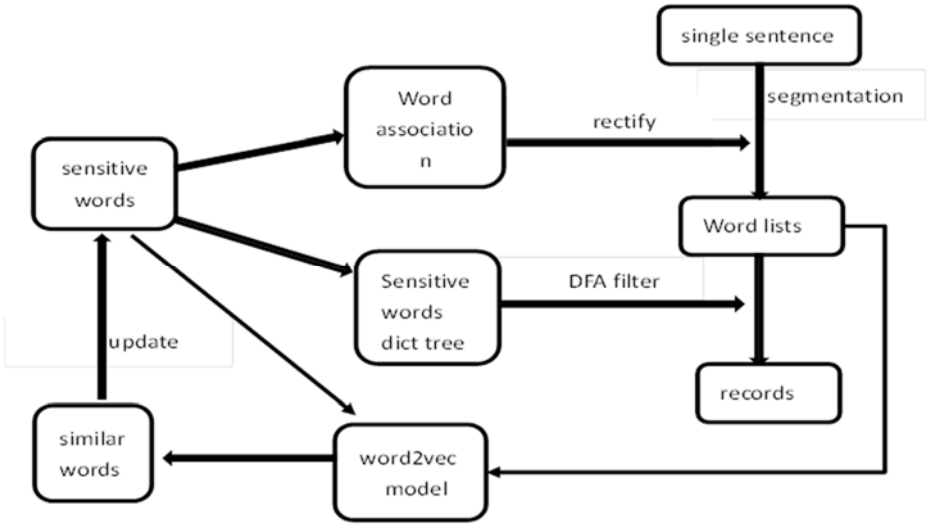

Fig. 1 DFA and word2vec based message sensitive words filtering system Framework

This framework is mainly divided into two parts, sensitive word filtering and sensitive word dictionary update. In the actual production environment, because the sensitive words filtering is for the instant messaging system, the response speed should be very fast, and the sensitive word update needs to calculate the similarity between all the words after the word segmentation and the existing sensitive words, and calculate The amount is slightly larger. Therefore, the two modules are concurrently performed after the word segmentation correction.

3.2. Sensitive words filtering module

Firstly, construct an association dictionary and a sensitive word dictionary tree according to the artificially pre-set list of sensitive words. Each input of the instant messaging system is composed of (name, timestamp, sentence) triples. The common word segmentation tool is used to segment the sentence, and then the correlation algorithm based on information entropy is used to correct the correlation algorithm. As follows: Assuming two adjacent Chinese characters *a* and *b*, let a be on the left side. From the inside, use the mutual information of points to examine the degree of mutual dependence of *a* and *b* from the perspective of mutuality [15]. From the outside, *a* is the center. Word, using information entropy to examine the ability of *a* to construct words. Give the following formula (1).

$$\mathrm{C(a,b)} = \begin{cases} \frac{PMI(a,b)}{H(a)}, & PMI(a,b) > 0,\ H_r(a) > 0 \\ 0, \text{ others} \end{cases} \tag{1}$$

Among them, the definition of PMI is formula (2), *p(a,b)* is the probability that the word appears in the sensitive lexicon , *p(a), p(b)* are the probabilities that a and b appear in sensitive lexicon, respectively.

$$\mathrm{PMI(a;b)} = \log_2 \frac{p(a,b)}{p(a)p(b)} \tag{2}$$

The definition of $H(a)$ is formula(3) and formula(4),which correspond to the corresponding left information entropy and right information entropy, respectively. A is a set of words in the sensitive lexicon that form words on the left side with the center word a, and B is a set of words in the sensitive vocabulary that form words on the right side with the center word a. Then use the stop word dictionary to remove the stop word [16], and finally use the dictionary tree to match the corrected word segmentation results in turn, and finally output the (name, timestamp, sensitive words list) triplet for subsequent statistical analysis work.

$$H_l(\mathrm{a}) = -\sum_{\varpi \in A} P(\varpi a|\varpi) \log_2 P(\varpi a|\varpi) \quad (3)$$

$$H_r(\mathrm{a}) = -\sum_{\varpi \in B} P(a\varpi|\varpi) \log_2 P(a\varpi|\varpi) \quad (4)$$

3.3. Sensitive words updating module.

On the other hand, the pre-trained word vector model is used to match the similarity of the word segmentation, and the words close to the sensitive words are identified and added to the sensitive word list, thus realizing the self-learning process of the sensitive word list [17]. In order to increase reliability, manual judgment can be added to the identified approximate sensitive words. For words that are not covered by the word vector model, the word vector model can be periodically trained by recording the sentences in which the words are not recognized.

4. Experimental results

4.1. Sensitive words filtering system platform

To test sensitive words filtering method in this paper, we coded with a common laptop with Ubuntu 14.04 LTS operation system and 4G RAM and Core i3 CPU. The interpreter we used is python 2.7 .

4.2. Sensitive words filtering and Sensitive words list updating

In order to simulate the real production environment, this paper obtains the real dialogue model training corpus, enterprise sensitive vocabulary and Chinese word vector model based on 120G network encyclopedia and news corpus training. From the sensitive words of the enterprise, 100 sensitive words are selected to form the initial list of sensitive words. 20000 sentences are selected from the dialogue corpus to identify them. From the perspective of recognition speed, the comparison with the simple DFA algorithm is carried out. The experimental results are as Fig 2. As shown in Fig.2, after segmentation and correction, the use of DFA recognition is much more efficient than the simple use of DFA algorithm recognition, especially when the sample size is larger, the total time-consuming advantage is more obvious. The recognition accuracy is almost the same, so the method of this paper is more effective.

At the same time, the number of sensitive words in the sensitive lexicon has also increased, as shown in Fig. 3. The sensitive lexicon grows very quickly, but it does not form a closed loop. In order to prevent the divergence of sensitive words, it should be properly added to the manual review measures to correct it.

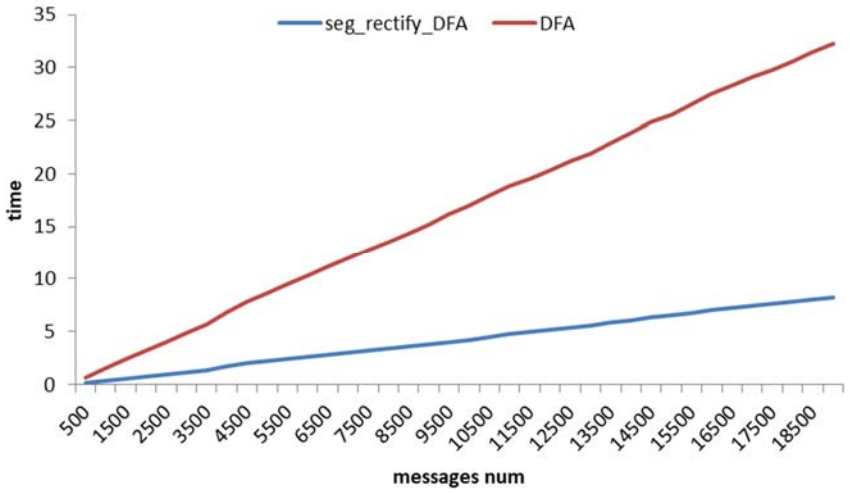

Fig.2 Comparison of seg_retcify_DFA and DFA

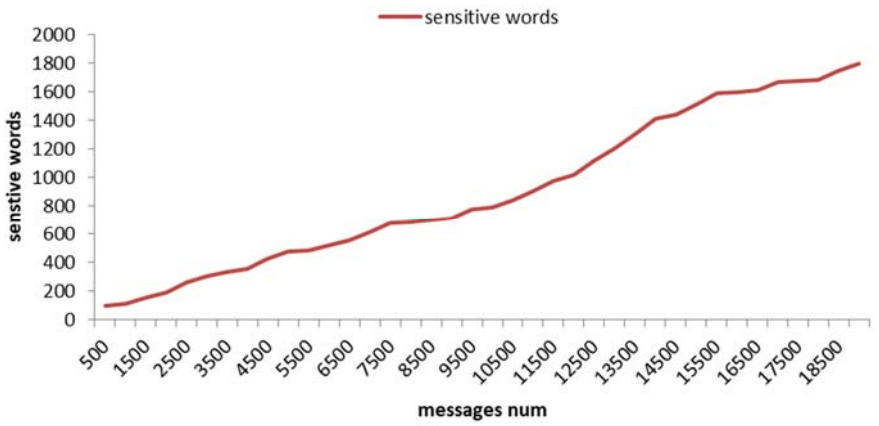

Fig.3 sensitive words updating

5. Conclusions

In this paper, the relevance algorithm based on information entropy and point mutual information is used to correct the word segmentation result, and combined with DFA algorithm for message sensitive word recognition, which effectively improves the speed and accuracy of DFA algorithm recognition. Secondly, this paper uses the pre-trained word vector model to achieve good results for the expansion and improvement of sensitive lexicon.

References

[1] SUN Wenjing, QIAN Hua.Impmved BM algorithm and its application in network intrusion detection[J].Computer Science,2014, 40(12):174-176(in Chinese)
[2] CHU Yanjie, LI Yunzhao, WEI,Qiang.Improved multi-pattern matching algorithm[J].Journal of Xidian University(Natural Science Edition)2014,41(6):174-180(in Chinese).
[3] HAN Guanghui, ZENG Cheng.Theoretical research of KMP algorithm[J].Microelectronics&Computer,2013,30(4):30-33
[4] DENG Yigui, WU Yuying.Information filtering algorithm of text content-based sensitive words decision tree [J] .Computer Engineering,2014,40(9):300-304(in Chinese).
[5] Bordihn H, Holzer M, Kutrib M. Determination of finite automata accepting subregular languages[J]. Theoretical Computer Science, 2009, 410(35): 3209-3222.
[6] Wyschogrod D, Leibman L. System and method for determining the start of a match of a regular expression: U.S. Patent 7,305,391[P]. 2007-12-4.
[7] Lang A K, Kosak D M. Information filter system and method for integrated content-based and collaborative/adaptive feedback queries: U.S. Patent 6,775,664[P]. 2004-8-10.
[8] Roger E S, Cundiff D M, Olson A E. Sensitive information handling on a collaboration system: U.S. Patent 8,151,200[P]. 2012-4-3.
[9] Bosch H, Thom D, Heimerl F, et al. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2022-2031.
[10] Wang M, Li X, Wei Z, et al. Chinese Word Segmentation Based on Deep Learning[C]//Proceedings of the 2018 10th International Conference on Machine Learning and Computing. ACM, 2018: 16-20.
[11] Arabacı M A, Esen E, Atar M S, et al. Detecting similar sentences using word embedding[C]//2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018.
[12] Qian Y, Du Y, Deng X, et al. Detecting new Chinese words from massive domain texts with word embedding[J]. Journal of Information Science, 2018: 0165551518786676.
[13] YanSong S S, JingLi Tencent A I. Joint Learning Embeddings for Chinese Words and their Components via Ladder Structured Networks[J].
[14] Heimerl F, Gleicher M. Interactive Analysis of Word Vector Embeddings[C]//Computer Graphics Forum. 2018, 37(3): 253-265.
[15] Liu K, Xu L, Zhao J. Co-extracting opinion targets and opinion words from online reviews based on the word alignment model[J]. IEEE Transactions on knowledge and data engineering, 2015, 27(3): 636-650.
[16] Na D, Xu C. Automatically generation and evaluation of Stop words list for Chinese Patents[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2015, 13(4): 1414-1421.
[17] Xue B, Fu C, Shaobin Z. A study on sentiment computing and classification of sina weibo with word2vec[C]//Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, 2014: 358-363.