Intelligence Information Retrieval Based on Text Mining

Yang Ran[2], Qiang Liu[1,2,3,a] , Zhang BO[2], Zhoulong Li[2], Deng Ke[2]

[1] Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, No. 8007, Hongli Road, Futian District, Shenzhen, 518034, Guangdong, P.R. China

[2]School of Resources and Environment, University of Electronic Science and Technology of China, NO. 2006, Xiyuan Avenue, Wust Hi-Tech Zone, Chengdu, 611731, Sichuan, P.R. China

[3] Chengdu Institute, University of Electronic Science and Technology of China, NO. 2006, Xiyuan Avenue, Wust Hi-Tech Zone, Chengdu, 611731, Sichuan, P.R. China

[a]liuqiang_em@sina.com

Keywords: Intelligence; Text mining; Word segmentation.

Abstract. With the rapid development of Internet technology in recent years, the amount of text information carried by online electronic documents, e-mails, databases and other forms was increasing explosively. Traditional intelligence information processing methods were increasingly difficult to meet task needs. In this paper, we proposed a word segmentation strategy based on intelligence dictionary, and built a dictionary for intelligence information, which effectively improved the accuracy of word segmentation in intelligence texts.

Introduction

Research on text mining concentrated on text representation, feature extraction of high-dimensional text data, and multiple text mining algorithms for text categorization, or clustering [2].

For large-scale text data, different mining algorithms could be utilized to mine different knowledge models [3]. Common text mining tasks based on mining algorithms could be roughly divided into the following four categories: text classification, cluster analysis, association rules, and trend prediction [4]. Feldman's team members conducted textual data mining of tens of thousands of financial news stories reported by Reuters in France and analyzed the relative distribution of some stock market transactions with company organizations and individuals [5]. Google's research team predicted that the flu virus would explode in many places in the near future by analyzing a large number of users in different regions of North America and the United States in a short period of time by searching keywords such as "flu" and "flu prevention and control methods" on Google search tools [6]. Beat Wuthrich's team utilized text analytics tools to study economic reports and financial policy stories that were published on Web pages and often successfully predicted the general volatility of stock indices in major stock exchanges [7].

Common word segmentation tools

Text participle was the forerunner of Chinese natural language processing. It was difficult for Chinese natural language to do data mining without word segmentation. There had been a large number of open-source tools for Chinese word segmentation. Different text mining projects needed to adopt different preprocessing methods, and different word segmentation systems had great influence on the accuracy and efficiency of word segmentation. At present, there were many open source tools for Chinese word segmentation, as follows:

Jieba was a Chinese word segmentation component based on the Python programming language. Its main functions were word segmentation of Chinese text, part-of-speech tagging and word segmentation after text segmentation. Stitching segmentation function had the characteristics of simple operation, fast speed and high precision. Three kinds of modes were provided for the text segmentation: the first was the exact mode, suitable for text mining with high precision; the second was the full mode, in which all sentences could be formed and scanned into words with efficient segmentation; the third was the search engine model, which made precise segmentation at first, then repeated it again. There was also an important function in Jieba, which supported custom dictionaries. This open interface was in favor of users to build specialized lexicon based on specific areas.

Chinese word segmentation system (ICTCLAS) was developed by the Institute of Computer Technology, Chinese Academy of Sciences. Its working principle was based on a multi-layer hidden Markov model with excellent performance of Chinese word segmentation. The functions include word segmentation of Chinese text, identification of semantics of already segmented words, discovery of terminology vocabularies, discovery of new words and the addition of the word segmentation system which also allowed users to add dictionaries of specialized fields according to actual work needs [7]. This word segmentation system could be used in a variety of popular programming environments and had become one of the most popular Chinese word segmentation systems in current text mining technologies.

Common Chinese word segmentation tools were Standard Analyzer, Chinese Analyzer, IK Analyzer, MMSEG4J, Paoding, and more. In terms of specific tasks, some require the pursuit of word segmentation efficiency, and some needed to pursue word segmentation accuracy.

The Chinese version of the Chinese word segmentation tool (jieba) and the Chinese word segmentation system of the Chinese Academy of Sciences (ICTCLAS) were widely used in the real world. However, whether such open-source tools are applicable in the field of intelligence needed to be verified through experiments. The verification method adopted in this paper was to randomly select multiple materials from the intelligence material library as the test sample, and then use the method of segmentation of the dictation and ICTCLAS system to obtain the segmentation results. Through the comparison of the precision of the word segmentation results, it was concluded whether the open source segmentation tool was suitable for intelligence text classification.

Word segmentation based on specialized intelligence dictionary

The segmentation method studied in this paper was oriented to the practice of intelligence production. The language of the intelligence text had the characteristics of conciseness, high concentration and high repetition rate for keywords. This paper applied a special intelligence dictionary method to improve the performance of intelligence text segmentation.

Intelligence dictionary method could be generalized as the following processes: initialize intelligence dictionary, input word segmentation, deal with intelligence text structure, segment intelligence, disambiguate word sen**se,** update dictionary and preserve results. The flow of word segmentation in intelligence dictionary was shown in Figure 1.

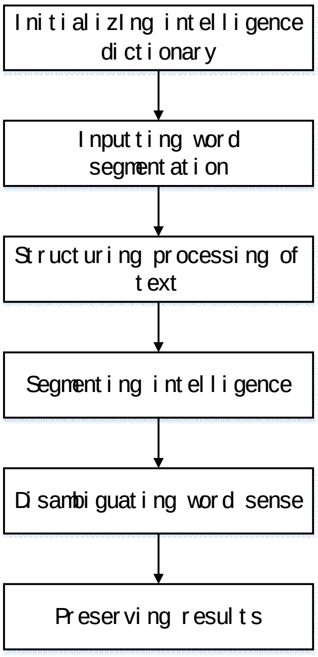
InitializIng intelligence dictionary
Inputting word segmentation
Structuring processing of text
Segmenting intelligence
Disambiguating word sense
Preserving results

Figure 1. Flow of word segmentation in intelligence dictionary

When removing stop words was completed, feature items not related to information topics were removed. Finally, the remaining features were more or less related to information topics, but their contributions were not the same. The weight calculation of the feature items was to give weight to each feature item. If a feature item had great contribution to the information theme, then its weight was large. Instead, if its contribution to the information theme was small, then its weight was small. In reality, feature items were usually arranged according to weights, ranging from large to small. The feature items with larger weights were preferred to be chosen.

The weight of feature items could be calculated with such methods as boolean weight, text frequency, entropy weight or TF-IDF weight.

In the TF-IDF weighting method, TF referred to the word frequency, which indicated the number of times that a characteristic item appeaed in a single specific text. IDF referred to the inverse text frequency, which expressed the reciprocal of a characteristic item appearing in the total text set. The main idea of the weight of TF-IDF was, if a feature appeared in a text in very high frequency (i.e. high TF value), and in other text appeared very low (i.e. high IDF value), then we would represent text category with the feature item. TF-IDF is the product of TF and IDF. The value of IDF was equal to the total number of texts divided by the logarithm of the number of the files containing the feature items. It was mathematically expressed as IDF=log (N/nj), where N was the total number of texts, and N represented the number of texts containing the entry t.

Acknowledgements

The Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources (No. KF-2016-02-007).

References

[1] Tan A H, Yu P S. Guest Editorial: Text and Web Mining[J]. Applied Intelligence,2003, 18(3):239-241

[2] Wuthrich B., Permunetilleke D. and Leung S.et al. Daily prediction of major stock indices from textual WWW data. In proceedings of the 4th International Conference on Knowledge Discovery. New York.1998

[3] Huang J.Improvement of Apriori Algorithm for Mining Association Rules[J]. Journal of University of Electronic Science & Technology of China,2003

[4] Bao Y, Ishii N, Du X. Combining Multiple k-Nearest Neighbor Classifiers Using Different Distance Functions[J]. Lecture Notes in Computer Science, 2004, 3177(1):634-641

[5] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases[C]//ACM SIGMOD International Conference on Management of Data. ACM, 1999:207-216

[6] Joachims T. Text categorization with support vector machines [J]. Fakultäten, 1999

[7] Grimaud A, Rouge L.Non-Renewable Resources and Growth With Vertical Innovations:Optimum,Equilibrium And Economic Policies.Journal of Environmental Economics and Management, 2003(45):433-4531