Research on dimension reduction methods facing massive high-dimensional web text data based on cloud computing

Deng Hui*

Library of North Sichuan Medical College, Nanchong, Sichuan, China, 637000

**Corresponding author's e-mail: denghuiswsw@126.com*

Received 6 October 2013, www.cmnt.lv

Abstract

The cloud model is introduced in the clustering dimension reduction process of the text data. In order to make the feature words selected meet this requirement, the cloud model theory is used for text feature selection, and association cloud filter together with distinction cloud filter is separately done for each feature in the training set; finally, the cloud feature space is obtained. Adopting the cloud computing model can not only allow the text information to be reflected more rationally but also ensure that the vector dimension will not be oversized to influence the machine learning ability. The cloud computing model can be introduced in the massive high-dimensional web text data; on one hand, speed of choosing the feature space can be increased, on the other hand, the data dimension reduction effect can also be enhanced.

Keywords: Cloud Computing; Text Data; Dimension Reduction; Feature

1 Introduction

Text data have no structure and they belong to nonlinearity, which may have a certain influence on computers while searching for information in the massive texts. The vector space model is needed to effectively solve this problem [1-4]. This model is able to convert some unstructured text contents into the linear text vectors; space between the text vectors is used to enhance the similarity of the text contents. Now this way has been widely used in many fields related to natural language processing. However, if the text is mapped to the text space, space dimensionality will become very huge; when such high-dimensional text vectors are processed with computer, the difficulties brought by the complexity of time and space will also become huge [5-8]. Therefore, when the text is reflected by the form of vector, not only such expressing way can rationally reflect the text information but also the oversized vector dimension cannot appear to influence the machine learning. Feature selection of the vector requires that the excessive amount of information cannot lose while the vector dimension is reduced. In this paper, dimensionality reduction methods for the massive high-dimensional web text data are researched from the perspective of cloud computing.

2 Feature selection in text categorization

2.1 COMMON FEATURE SELECTION METHODS IN TEXT CATEGORIZATION

In text categorization, one of the first problems to be solved is how to adopt an appropriate set of features to reflect the text content. All the training corpus texts in categorization calculation have category identifications, belonging to the process with guide; therefore, the originally existing category identifications can be used to dig out the relationship between the feature words and the categories themselves in these categories. The methods commonly used are: mutual information, information gain, X^2 statistics and so on. Below, these methods will be introduced and illustrated in details.

2.2 METHOD DESCRIPTION

2.2.1 Mutual information

Mutual information (M1) is from information theory, which mainly reflects the amount of information of association degree between two objects. It is embodied in the association degree between the feature word and another category in text feature selection. If the mutual information between a word and another category is larger than a certain threshold, then the word is considered to have certain association with the category; otherwise, the word is not associated with the category. Definition of mutual information is as follows:

$$MI(t_i) = \sum_j p(C_j) \log \frac{p(t_i/C_j)}{p(t_i)}, \quad (1)$$

where, $p(t_i/C_j)$ indicates appearance of the feature item t, and $p(C_j)$ indicates the proportion of the document in the category Cj, while $p(t_i)$ indicates the proportion of the document where the feature item t_i appears in the whole training set.

2.2.2 x^2 Statistics

The relationship between the feature word r and a certain category is assumed to match χ^2 distribution. This method is similar to the mutual information method; both need to quantify through the association degree between the target feature and the category, as well as the amount of information carried by the category.

$$\chi^2(t_i, C_j) = \frac{N(AD - BC)^2}{(A+C)(A+B)(B+D)(C+D)}, \quad (2)$$

where, N is the total number of documents contained in the training corpus; A represents the number of documents contained and belonged to; B represents the number of documents contained but not belonged to; C represents the number of documents not contained but belonged to; D represents the number of documents neither contained nor belonged to.

2.2.3 Information gain

Information Gain (IG) is a very effective feature selection method and is widely used in text data mining function. This method is different from the two methods above: information gain does not separately consider the relationship between a feature word and a category but regards all the categories in the training corpus as a whole, using the amount of information, which can be brought to the whole system, to measure the importance degree of the word. Information gain of the feature word is the difference value between the amount of information carried by the whole training set without considering feature word and the amount of information carried by the whole training set considering the feature word. Its definition is as follows:

$$Gain(t_i) = Entropy(S) - Expected\_entropy(S_{t_i}). \quad (3)$$

3 Application of cloud model theory in the field of natural language processing

Cloud model theory is the product of reasonable compatibility of fuzzy mathematics and probability theory, which accurately portrays the concept of indeterminateness in natural language through the numerical features. In recent years, it has been widely used in data mining, intelligent computing, knowledge discovery, network security and other fields [9-10]. Among which, the key Laboratory of Chinese Information Processing in Central China Normal University has done the relevant research according to the cloud model theory. The specific contents are as follows:

Chen Jinguang applies the cloud model theory in multi-document automatic abstracting technology, which refers to the methods of abstract unit selection related to the cloud model. It can be found in the experiments that this method is characterized in the strong adaptability and stable performance, which can be better reflected in English corpus and Chinese corpus. In addition, it has also obtained the better results in many international public evaluation corpuses.

Wan Jianyun applies the cloud model theory in the text categorization technology, adopting the cloud model theory to improve the traditional Naive Bayes Classification Algorithm and using the expectation in it to correct the association degree between the words and categories. This approach enables accuracy of Naive Bayes Classification Algorithm in Reuters corpus - 20 news group to promote 5.7%.

Lou Zhenxia applies the cloud model theory in information retrieval technique. From the perspective of uncertainty in the cloud model theory, she focuses on researching document rearrangement methods, thoroughly explores the association degree between the inquiring statements and the candidate documents, and rearranges the documents while grading them so that accuracy of information retrieval has been essentially improved.

4 Text feature selection based on the cloud theory

Text categorization is a machine learning process with guide. While doing automatic text categorization, a feature set of low dimension needs to be established to effectively express the text content. Feature selection is an indispensable step for whatever disaggregated model. This paper argues that a reasonable feature word needs to have enough association degree and distinction degree. Association degree describes the degree of the feature word contacting a category and distinction degree describes the ability that the word distinguishes the category it represents from other categories. In order to make the feature words chosen meet this requirement, the cloud model theory is used for text feature selection, and association cloud filtration together with distinction cloud filtration is separately done for each feature in the training set; finally, the cloud feature space is obtained.

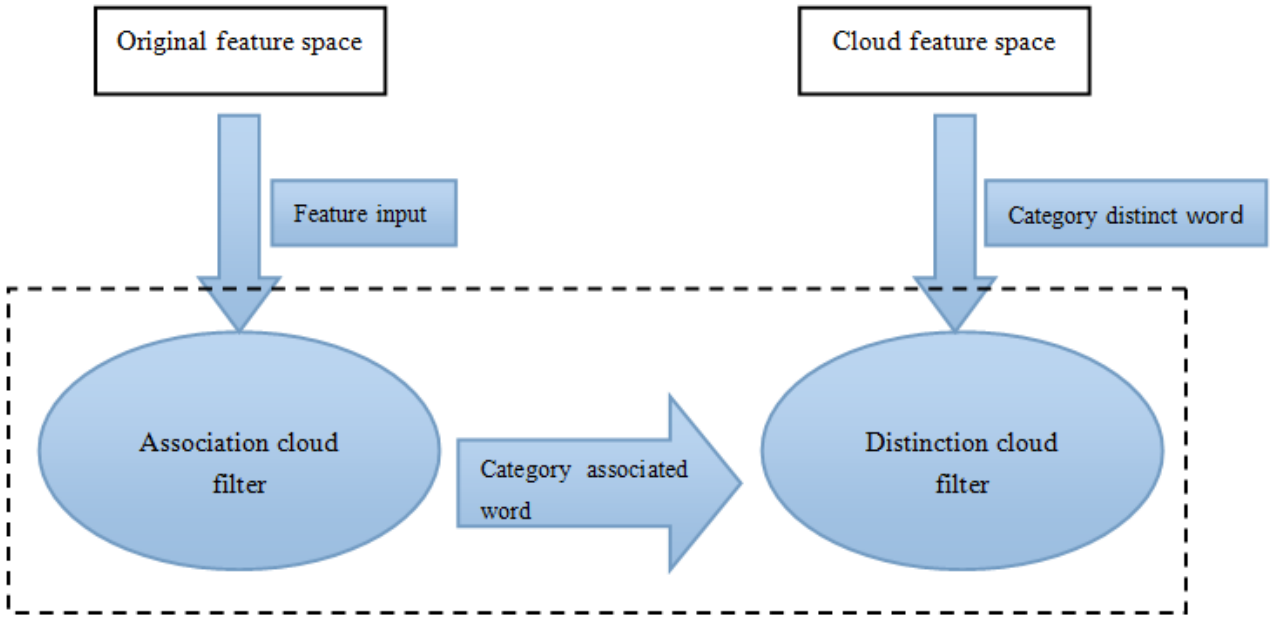

FIGURE 1 Setting up procedures of cloud feature space

5 Feature selection of text data dimension reduction process based on cloud computing

5.1 TEXT FEATURE SELECTION BASED ON TF-IDF

Cluster computing is without a guide; all the feature words that the machine faces to have no category identification with out the training corpus. Feature selection in the cluster has no category information for reference so the feature selection methods mentioned in the foregoing can not be directly used here. TF-IDF method is used to evaluate importance degree of a feature word to a document in a corpus set. Its main idea is that if a word appears in the document set many times and rarely appears in other documents, then the word would have a strong category distinction. In the TF-IDF evaluation criteria, the frequency of the feature word appearing in the document set needs to be determined and the distribution of the feature word in the document set needs to be macroscopically described. The calculation method is as follows:

$$idf(t) = \log(N / n_t + 0.01). \tag{4}$$

Corresponding to the clustering problem, N is the total number of clustering test document; n_t is the number of documents where the feature word t appears in the whole test set. 0.001 is the smooth factor and $idf(t)$ is avoided to be 0. Thus, the computational formula of TF-IDF is as follows:

$$tf\_idf(t) = tf(t) \times \log(N / n_i + 0.01). \tag{5}$$

The principle of TF-IDF method is easy to understand; if a word appears for less times, than the word is proved to be unrepresentative; while the word appears in many documents, it indicates that the word show no distinction to the documents. The principle of feature selection based on this method is to find the feature words that appear more times but are contained by the few documents. Using this method for feature selection can indeed reduce the dimension of feature space. However, after all, it is just a borrowed algorithm without very adequate theoretical basis. Therefore, the effectiveness of this method is not absolute. Because TF-IDF can directly calculate the distinction degree of the feature words between the documents without category as the guide, it is very appropriate to be used in feature selection of clustering documents.

5.2 TEXT FEATURE SELECTION BASED ON CLOUD THEORY

In the text clustering, document is usually directly represented by the vector space model (VSM). If there are two documents d_1 and d_2, they are respectively represented by the vectors v_1 and v_2. Similar degree between d_1 and d_2 is measured by the cosine values of v_1 and v_2.

$$sim(d_1, d_2) = \cos(v_1, v_2) = \frac{v_1 \bullet v_2}{|v_1 \bullet v_2|}. \tag{6}$$

If feature dimension reduction processing is not done, each word in the document will form each dimension of the document vector. Thus, the cost is very high during similarity calculation. Meanwhile, errors will be brought to similarity calculation because of existence of some useless features. For example, there are the two short documents A, B belonging to the same category.

A: Wang Ming thinks the performance of this computer is very good. B: Li Lei also thinks the performance of this computer is very good. Assuming that an effective feature selection is made, and then there are only these words in the feature space V: "computer", "performance", "good". Thus, Document A is represented as (1, 1, 1) and Document B is also expressed as (1, 1, 1), where 1 represents the frequency of the word appearing in the document. Thus similarity degree between Document A and Document B is 100%. If feature selection is not made, the similarity degree will reduce a lot. Therefore, importance of feature selection in the text clustering cannot be ignored.

5.3 K-MEANS TEXT CLUSTERING BASED ON CLOUD FEATURES

Determination of initial point number as well as selection of the initial point itself is a bottleneck problem that K-means algorithm is currently experiencing. Therefore, when using this text clustering method based on partition to do clustering calculation of similarity degree, larger floatability and instability usually appear in its clustering results. Research emphasis of this paper lies in how to select the effect feature words in the document set instead of improving the K-means algorithm so the artificial specified method is used to select the initial point.

5.3.1 Experiment description

To confirm the validity of the feature set reflected by Cluster-Filter in the text clustering, TF-IDF is used here to do compare experiments. In order to compare the effects of the two feature selection methods fairly, these two feature selection methods are adopted in the experiment to the feature subsets of the same scale, and the method of specifying the same initial point artificially is used in K-means.

5.3.2 Corpus sources

Six categories of computer, Experiments selected computer classes, art classes in six categories, literature and art, economy, sports, history class and aviation in Fudan Chinese Corpus are selected for the experiment. 50 documents are randomly extracted from each category, in total of 300 documents.

6 Conclusion

Preprocessing stage:

Clustering has no training corpus so the same techniques of the preprocessing are used and classified to carry out word segmentation, to stop words and to filter impurities for the text set in the clustering. Finally, all the feature words appearing in the text documents are counted. Then the set of feature words forms the original feature space V_SET.

Feature selection stage:

This stage is mainly to do dimension reduction processing for V_SET. Here, TF-IDF and Cluster-Filter are respectively used to carry out feature selection; the numbers of the feature words selected respectively are 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, and 1500.

Text representation stage:

The core algorithm of the text clustering based on K-means is the similarity calculation between the text vectors so it is necessary to represent all the documents in the test set as the vector space forms. Each dimension of the vector is each feature word in the feature subset and each value on the coordinate is the frequency of the feature word appearing in the text.

Acknowledgement

This work was supported by the Fund of North Sichuan Medical College (14ZB0200).

References

[1] Turcu Gabriela, Foster Ian, Nestorov Svetlozar 2011 Reshaping text data for efficient processing on Amazon EC2 *Scientific Programming* **19**(2-3) 133-45
[2] Zubi Zakaria Suliman 2010 Using text mining techniques in Electronic Data Interchange environment *WSEAS Transactions on Computers* **9**(8) 832-46
[3] MacKenzie I S, Soukoreff R W 2002 Text entry for mobile computing: Models and methods, theory and practice *Human-Computer Interaction* **17**(2-3) 147-98
[4] Háva Ondrej, Skrbek Miroslav, Kordík Pavel 2013 Supervised two-step feature extraction for structured representation of text data *Simulation Modelling Practice and Theory* **33** 132-43
[5] Goto Masayuki, Ishida Takashi, Suzuki Makoto, Hirasawa Shigeichi 2010 A theoretical analysis of document classification based on a high-dimensional vector space model - Asymptotic analysis of classification performance and distance measures *Journal of Japan Industrial Management Association* **61**(3) 97-106
[6] Wang Tai-Yue, Chiang Huei-Min 2011 Solving multi-label text categorization problem using support vector machine approach with membership function *Neurocomputing* **74**(17) 3682-9
[7] Lee Lam Hong, Wan Chin Heng, Rajkumar Rajprasad, Isa Dino 2012 An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization *Applied Intelligence* **37**(1) 80-99
[8] Kim Kwang In, Jung Keechul, Park Se Hyun, Kim Hang Joon 2001 Support vector machine-based text detection in digital video *Pattern Recognition* **34**(2) 527-9
[9] Wang Pingshui 2010 Survey on privacy preserving data mining *International Journal of Digital Content Technology and its Applications* **4**(9) 1-7
[10] Zhan Justin 2008 Privacy-preserving collaborative data mining *IEEE Computational Intelligence Magazine* **3**(2) 31-41

Author

Deng Hui, 1980, Nanchong, Sichuan, P.R. China

Current position, grades: Master degree the of lecturer North Sichuan Medical College, China.
Scientific interest: His research interest fields include Computer networks, data mining.
Publications: more than 15 papers published in various journals.
Experience: He has teaching experience of 12 years, has completed three scientific research projects.