Multi-level feature filtering algorithm based on MapReduce

Zhenling Zhang¹, Chao Wang² and Zhiyuan Li³

¹Zhengzhou Institute of mechanical and electrical engineering, China
²Zhengzhou Institute of mechanical and electrical engineering, China
*Corresponding author’s e-mail: 854909839@aaaa.aaa

Abstract: In engineering practice, we need to consult a large number of professional data to support theoretical innovation. If we rely on manual classification and screening one by one, it will take a lot of time and energy, and the classification accuracy cannot meet the requirements. In order to improve the work efficiency, a multi-level feature selection algorithm based on MapReduce is proposed. The improved Chi feature selection algorithm can be used for the initial screening, and then the noise words and pre quality features can be filtered by mutual information method. The experimental results show that the algorithm not only ensures the low time complexity of processing big data, but also improves the accuracy of text classification.

1. Introduction

Text classification is a research hotspot in data mining. It refers to dividing a series of texts into predefined categories. Feature selection is a key step in text categorization. It refers to the selection of words from text that best represent the text information as a reference for classification. Therefore, the quality of feature selection algorithm directly affects the accuracy of text categorization [1].

MapReduce is a programming model based on distributed parallel computing proposed by Google for large-scale data processing. Its core strategy is to divide and conquer it, and distribute a huge set of jobs equally to multiple nodes for processing [2]. By splitting and combining data, not only the ability of parallel processing data is improved, but also the system performance is greatly improved.

Traditional feature selection methods include Chi Square Statistic (CHI), Mutual Information (MI), Document Frequency (DF) and so on. Traditional feature selection algorithm considers fewer constraints and has low classification accuracy. Multi-level feature selection model can effectively solve the problem of low classification accuracy. A multi-level feature model based on feature Contribution Degree (FCD) and Latent Semantic Indexing (LSI) is proposed in document [3]. Literature [4] uses CHI-like extractors for primary selection, and then uses LSI-based genetic algorithm (GA) for secondary selection. Reference [5] proposes a multi-level feature selection model with information Gain Sequences (IGS) as primary selection and GA as secondary selection. Although the above methods can improve the classification accuracy, they also have obvious shortcomings. After completing the second-level feature selection, the set of feature words will lose part of the original semantics. The reason is that the multi-level feature selection has strong filtering ability and seldom considers the context. Literature [6] Firstly, IGS is used to make primary selection of text, and then Markov Blanket Filtering (MBF) algorithm is used to make secondary selection. To a certain extent, this method retains relatively complete original semantics, but it has high time-consuming. Literature [7] A modified TF-IDF method was used for preliminary screening, followed by a second screening using maximum correlation minimum redundancy (MRRMR), and an incremental search...
was used for secondary screening to reduce time consumption. This method overcomes the shortcomings of the above algorithm and optimizes the time consuming, but there is still much room for improvement.

To solve the above problems, while improving the feature selection algorithm, the original semantics are preserved as much as possible, and the time consumption of the model is low. This paper proposes a multi-level feature selection mechanism based on MapReduce. The improved CHI method is used for primary selection, and then the mutual information method is used to filter the CHI method. The noise words are generated and the appropriate feature words are prefixed. Considering that the mutual information method needs a lot of time, this model is loaded into the MapReduce framework to give full play to its advantages in dealing with massive data and improve the efficiency of model execution.

2. Multi-level feature selection mechanism based on MapReduce

2.1. Multilevel Feature Selection Model

To overcome the shortcomings of the CHI method described above, a two-level feature selection model based on improved CHI method and mutual information method is proposed. As a primary feature selection method, the improved CHI method (NCHI) can eliminate the influence of the traditional CHI method which only considers the frequency of documents but neglects the frequency of feature words. The feature word set obtained by the primary selection is then selected by mutual information method to filter out the noise words generated during the primary selection, as shown in Figure 1.

![Multilevel Feature Selection Model](image_url)

Figure 1 Multilevel Feature Selection Model

Unlike the traditional CHI method, NCHI introduces the intra-class frequency as a new parameter, which is recorded as CIF_{ij}. The intra-class frequency represents the maximum number of occurrences of the feature word Ti in all documents of category Cj. The calculation formula is shown in Formula (1).

\[
CIF_{ij} = \max_{u=1}^{M} tf_{iju}
\]

Among them, tf_{iju} is the number of times the feature Ti appears in the document Du under category Cj, and M is the number of documents in category Cj (M in the same table). Formula (1) shows that the larger the tf_{iju} is, the more frequent the feature word Ti appears in the document Du under category Cj, the more likely it is to be used as a candidate feature word for this category.

To sum up, the formula of NCHI method is shown in formula (2), and \( \chi^2_{max}(t_i, C_j) \) is the value of NCHI method.

\[
\chi^2_{max}(t_i, C_j) = CIF_{ij} \times \chi^2(t_i, C_j)
\]

Intra-class frequency introduced in NCHI can solve the problem that the traditional CHI method ignores the frequency of feature words, but it does not solve the problem that CHI method produces additional noise words. Therefore, in secondary selection, the use of mutual information method can effectively filter the noise words selected in NCHI primary selection process. It can be seen from data that noise words have the following characteristics: (1) they contain the characteristic word Ti but do not belong to the category Cj. (2) It belongs to category Cj but does not contain the characteristic word ti. The feature words Ti and Cj in this kind of noise words have very weak correlation. Mutual
information method is just a kind of classification method to examine the correlation of variables, so it can give full play to its advantages as a secondary selection, effectively filter the noise words generated in the primary selection. In addition, the excellent feature words are pre-positioned because of their high correlation and will not be omitted by truncation of the word range.

2.2. Multilevel Feature Selection Mechanism Based on MapReduce

The multi-level selection model proposed in this paper can effectively improve the classification accuracy, but the calculation of intra-class frequency and the use of mutual information method in NCHI method will cause a lot of time overhead and reduce the efficiency of model execution. Therefore, the model is loaded into MapReduce framework, and its advantages of efficient processing of massive data are utilized to reduce the execution time of multi-level feature selection model. The process of multi-level feature selection mechanism based on MapReduce is shown in Figure 2.

![Flow chart of multi-level feature selection mechanism based on MapReduce](image)

Figure 2 Flow chart of multi-level feature selection mechanism based on MapReduce

From Figure 2, we can see that the multi-level feature selection model based on MapReduce is divided into two parts: text training stage and text classification stage.

2.2.1 Text Training Phase

Figure 3 is a multi-level feature selection text training stage model based on MapReduce.
For a given set of training texts, the pre-processing is carried out first, including word segmentation, sentence breaking and word deactivation. The processed set of texts will be marked as a first-order training set and input into the training model. As shown in Figure (3), the text training phase consists of three MapReduce processes. The function of the MapReduce process is to calculate the intraclass frequency of the feature words. The first-order training set is partitioned into different nodes to execute Map function, and the calculation method of formula (5) is mainly executed. The key-value pairs based on < feature words, < category, tf|ju > are recorded as < Tid, < Cid, tf|ju > > after intermediate process processing, and finally through Reduce function, the maximum CIFij of tf|ju is found with Tid as the main key of classification, thus the key-value pairs < feature words, < category, tf|ju > > are obtained. The number of intra-class frequencies is recorded as < Tid, < Cid, CIFij >. The key-value pairs are input into (2) MapReduce processes as second-order training sets.

The core formula of Map stage is formula (6). The NCHI values of the obtained feature words Ti and category Cj are stored as key-value pairs and recorded as < Tid, < Cid, 2 new >. After the intermediate process, the reduction function is executed with Cid as the main key of classification, and the_2 new values obtained by descending order arrangement are pre-regulated. Fixed the first f, so as to get the initial eigenvectors of Cj, and finally put them into the primary training library.

After the data collection of the primary training library is completed, the secondary feature selection can be carried out. The primary feature vectors of different classes are divided into classes as a whole, and are input into different nodes to perform Map functions. The calculation method of the main execution formula (4) is used to obtain the key pairs After the intermediate process, Cid is used as the main key of the classification, and Reduce function is executed. The MI values obtained by the same descending order are taken from the pre-defined first s to obtain the secondary eigenvectors of the category Cj. Finally, the MI values are classified into the secondary training library. According to the above process, the multi-level feature selection training algorithm based on MapReduce is described as follows:

2.2.2 Text Categorization Phase

The data of the classification model comes from the pretreated test text set and is recorded as the first-order test set. In this process, the MapReduce process is used to generate document feature vectors by counting the number of times the feature word Ti appears in the test document. It should be emphasized that this process only cares about the frequency of the feature word Ti appearing in the document Du (referred to as tf|ju), and does not care about the document category. The first-order test
set takes text as a whole and is divided into different nodes according to blocks to execute Map function. Map function is used to calculate tfiu of feature words. After that, key value pairs < Tid, Did >, tfiu > are recorded as <Tid, Did >, tfiu > and processed by intermediate process. Finally, through Reduce function, the feature words with Did are arranged in descending order according to tfiu value, and the pre-defined first k values are taken to form document du. Eigenvector. After archiving, it is fed into the MapReduce next process.

The results of the MapReduce process are archived and recorded as second-order test sets, which are input to next MapReduce process. The role of the MapReduce process is to classify the second-order test sets using KNN classifiers in the second-order training library. After KNN classification, the key pair < Cid, Did > based on < Category, Document > is also processed through the intermediate process. Finally, Reduce function is executed to classify the documents with Cid on the basis of Category.

3. Experiments and analysis

3.1. Experimental Data
This paper uses the corpus provided by Fudan University (natural language processing group, international database center, Department of computer information and technology, Fudan University) to select six categories of documents as training set and test set. For unbalanced corpus, the selection is shown in Table 1.

| Corpus     | servo | optics | image | communicati on | mecha nical | artillery |
|------------|-------|--------|-------|-----------------|-------------|-----------|
| Training   | 2000  | 2000   | 1000  | 1000            | 50          | 50        |
| Test set   | 2000  | 200    | 1000  | 100             | 50          | 500       |

3.2. Experimental Settings
In this paper, the following aspects are considered in the experiment:

(1) Interference test. We will test whether MapReduce mechanism has interference or adverse effect on the dimension of feature selection vectors in multi-level feature selection model. Specific experiments are as follows: multi-level feature selection model is run on ordinary PC and Hadoop single node respectively, and text vector dimension is selected by comparing primary feature selection and secondary feature selection.

(2) Performance comparison experiment. We will test the performance of multi-level feature selection model. The specific experiments are as follows: comparing IGS-GA method proposed in document [5], TFIDF-MRMR method proposed in document [7], and NCHI-MI method proposed in this paper [in order to express concisely, NCHI-MI method is used to express the multi-level feature selection method proposed in this paper]. TFIDF-MRMR, IGS-GA and NCHI-MI are used to represent the three method names in the result analysis. ) Classification effect.

(3) Efficiency test. We will test the efficiency of the proposed multi-level feature selection model on Hadoop platform. The specific experiments are as follows: Comparing the execution time of IGS-GA method, TFIDF-MRMR method and NCHI-MI method proposed in this paper, and observing the execution acceleration ratio of NCHI-MI method on Hadoop platform with different number of nodes.

The machines used in the experiment are built according to the following configuration: CPU is Intel Core i5-6500 3.20 GHz, 8G memory, 2TB hard disk, the operating system is Ubuntu 14.04, Hadoop Version 1.2.1, Java Version 1.2.1. The K value of KNN classifier is 500 in the experiment.
3.3. Evaluation Index

(1) Accuracy is used to reflect the accuracy of classification results, which is recorded as $P$, and its calculation formula is shown in Formula (3).

$$P = \frac{\text{Number of correctly classified texts}}{\text{The total number of texts in this category}}$$

(2) The recall rate is used to reflect the ability to classify correctly. It is recorded as $R$ and its calculation formula is shown in Formula (4).

$$R = \frac{\text{Number of correctly classified texts}}{\text{The total number of texts classified into this category}}$$

(3) The F1 value is a comprehensive evaluation index based on the precision and recall rate, which reflects the comprehensive performance of the system. The calculation formula is shown in Formula (5).

$$F1 = \frac{2PR}{P + R}$$

Acceleration ratio is used to measure the performance of task parallel processing [15], which is denoted as $Ts$, and its calculation formula is shown in equation (6).

$$Ts = \frac{\text{Single processor execution time}}{\text{Multi node parallel execution time}}$$

3.4. Result Analysis

The results of interference test are shown in Table 2.

Table 2 Number of words picked up by PC and Hadoop single node at each stage

| Serial number | category   | PC single machine | Hadoop single node |
|---------------|------------|-------------------|-------------------|
|               |            | Primary characteristics Vector dimension | Second level eigenvector dimension | Primary characteristics Vector dimension | Second level eigenvector dimension |
| 1             | servo      | 863               | 698               | 851               | 676               |
| 2             | optics     | 719               | 565               | 739               | 591               |
| 3             | image      | 724               | 464               | 716               | 450               |
| 4             | communication | 688            | 571               | 680               | 563               |
| 5             | mechanical | 535               | 460               | 522               | 462               |
| 6             | artillery  | 618               | 463               | 626               | 475               |

Table 2 shows that the dimension of primary feature selection vectors and secondary feature selection vectors differ little on PC and Hadoop single nodes, and the difference between them is acceptable within the range of error and execution. In summary, the introduction of MapReduce technology has no obvious interference or adverse impact on the multi-level feature selection model.

Fig. 4, Fig. 5 and Fig. 6 respectively represent the experimental results of each accuracy rate, recall rate and F1 value under the three methods.
From Figure 4 to Figure 6, the following conclusions are drawn and analyzed:

(1) Generally speaking, for the three methods, the more text sets are trained, the higher the classification accuracy is. Because of the imbalance of corpus, the training text sets of "economy" and "medicine" are fewer, so the three methods cannot extract enough feature vectors for training database, resulting in poor classification accuracy; when corpus is relatively sufficient, the three methods can select enough feature words, and each index has been greatly improved; when corpus is enough, the three methods cannot extract enough feature vectors for training database. Each index of the method has also been improved to some extent, but the increase is not large and tends to be stable. This shows that increasing the size of corpus is no longer the main factor determining the improvement of the index, so it is necessary to improve the algorithm or improve the level of hardware.

(2) The average recall rate of the three methods is slightly higher than the average accuracy rate. The ability of correct classification is not only the premise of algorithm classification, but also the premise of improving classification accuracy. The three methods can distinguish positive and negative text sets well, but the classification accuracy is slightly lower, which shows that the ability of correct classification of the three methods is better than the accuracy of classification.

(3) NCHI-MI has a slight improvement in accuracy, recall and F1 value compared with TFIDF-MRMR. Generally speaking, NCHI-MI has little difference, but it is higher than IGS-GA. The classification effect of IGS-GA is quite excellent. Because of its strong filtering, it will lose part of the original semantics, so each index is lower than NCHI-MI and TFIDF-MRMR, which is also the bottleneck of IGS-GA. NCHI-MI and TFIDF-MRMR make improvements to IGS-GA losing original semantics, use relatively peaceful selection method, and then do redundant filtering and denoising optimization processing. The experimental data show that each index has been significantly improved, and NCHI-MI has better optimization effect.
Table 3  
Comparison of running time

| Serial number | Category       | Running time (ms) |
|---------------|----------------|-------------------|
| 1             | TFIDF-MRMR     | 179236            |
| 2             | NCHI-MI        | 197054            |

Efficiency test experiment. Table 3 records the running time of NCHI-MI and TFIDF-MRMR on a PC. The experimental data show the NCHI-MI execution time ratio on ordinary PC TFIDF-MRMR slows down by about 18 seconds. Considering the time-consuming calculation of MRMR, TFIDF-MRMR uses incremental search to obtain appropriate feature words, which reduces time consumption to a certain extent, but the acceleration effect is not optimistic. In this paper, NCHI-MI is placed on Hadoop platform for distributed parallel processing, which greatly reduces the time consumption.

Figure 7  
Acceleration Ratio of Different Nodes

In order to show the trend of acceleration ratio more intuitively, Figure 7 is a broken line chart of acceleration ratio calculated from the running time of different nodes. Data show that with the increase of Hadoop nodes, the system acceleration ratio increases. When the number of nodes is 2, 3 and 4, the acceleration ratio increases faster; when the number of nodes reaches 15, the acceleration ratio increases slowly, and the number of nodes will become saturated. When the number of nodes is too large, the processing time of the system is very short, mainly in data allocation, load balancing and so on.

4. Conclusion

In this paper, a multi-level feature selection model based on Mapeduce is proposed. The improved CHI method is used to solve the problem of ignoring the frequency of feature words. It is used as the primary feature selection. Then the noise words generated by the primary selection are filtered by mutual information method. The appropriate feature words are pre-positioned as the secondary feature selection. Finally, the excellent performance of large data is processed by Mapeduce technology. This two-level model is loaded on the Hadoop platform. The experimental results show that this mechanism can significantly improve the accuracy of text classification and the efficiency of classification processing.

It is worth noting that the mechanism proposed in this paper also has some shortcomings, and there are still many details to be improved. The classification accuracy of the training database composed of a small amount of training corpus is too low, and the second-level selection of mutual information method can still be improved, which will be the key research directions in the future. In addition, the "Multi-level Feature Selection Model + MapReduce Framework" proposed in this paper is a kind of basic mechanism that can be applied. Whether other algorithms can perform better or not is another key research direction.

Reference

[1] Dash M, Liu H. Feature selection for classification[J]. Intelligent Data Analysis, 2007, 1(3):131-156.
[2] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[C]// Conference on Symposium on Operating Systems Design & Implementation. USENIX Association, 2014:10-10.
[3] Meng Jialna, Lin Hongfei, Yu Yuhai. A two-stage feature selection method for text categorization [C]//2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE, 2010:1492-1496
[4] Kursat U A, Serkan G. A novel probabilistic feature selection method for text classification [J]. Knowledge-Based Systems, 2012, 36(6):226-235.
[5] Harun U. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm [J]. Knowledge-Based Systems, 2011, 24(7):1024-1032.
[6] Kashif J, Sammen M, Babri Haroon A. A two-stage Markov blanket based feature selection algorithm for text classification [J]. Neurocomputing, 2015, 157:91-104.
[7] Li Ji, Fu J, Jiang W, et al. Feature selection method based on MRMR for text classification[J]. Computer Science, 2016, 43(10):225-228.
[8] Huang Y, Li M, LV JC. New feature selection method based on CHI[J]. Computer Science, 2015, 42(5):54-56.
[9] Zheng Zhao, Lei Wang, Huan Liu, Jieping Ye. On Similarity Preserving Feature Selection[J]. IEEE Transactions on Knowledge & Data Engineer, 2013, 25(99):1-1
[10] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.
[11] Mashayekhy L, Nejad M, Grosu D, Zhang Q, Shi W. Energy-Aware Scheduling of MapReduce Jobs[J]. IEEE International Congress on Big Data, 2014, 26(10):32-39
[12] Han Lei, Sun Xuzhan, Wu Zhichuan, Chen Lijun. Optimization Study on Sample Based Partition on MapReduce[J]. Journal of Computer Research and Development, 2013, 50(Suppl.):77-84.
[13] Neil Gunther; Paul Puglia; Kristofer Tomasette. Hadoop Superlinear Scalability[J]. Queue. 2015, 13(5):1542-7730
[14] Fei X, Li X F, Shen C. Parallelized text classification algorithm for processing large scale TCM clinical data with MapReduce[C]// IEEE International Conference on Information and Automation. IEEE, 2015:1983-1986.
[15] Liu J, Zhu A, Qin C. Estimation of theoretical maximum speedup ratio for parallel computing of grid-based distributed hydrological models[J]. Computers & Geosciences, 2013, 60(10):58-62.