Construction of Implicit Semantic Multi-label Text Fast Clustering Model based on Big Data

Dawei Zhao, Gang Chen*

College of Humanities & Sciences of Northeast Normal University, Changchun, Jilin, 130117, China

Keywords: Big Data, Semantic Multi-label, Construction Method

Abstract: Aiming at the conceptual ambiguity and underlying semantic structure of multi-label text classification, an integrated classification method is proposed to combine random forest (RF) algorithm and implicit semantic index (LSI). Through the random segmentation of vocabulary, the diversity of integration is increased, different orthogonal projections of low-dimensional implicit semantic space are obtained, and LSI is performed on the basis of orthogonal projection in low-dimensional space. Random forest can effectively solve the binary classification problem, and implicit semantics reveals the underlying semantic structure of the text. The combination of the two can represent the diversity of the group and the individual accuracy. The experimental results on the Yahoo dataset verify the effectiveness of the proposed method, which is superior to other methods in terms of Hamming loss, coverage, first error and average accuracy.

1. Introduction

With the rapid development of information technology, the scale, scope and depth of database applications continue to expand. The Internet has developed into the world's largest information base and the most important channel for disseminating information on a global scale. Whether it is a commercial enterprise, a research institution or a government department, in the past several years, a large number of data stored in different forms have been accumulated. Because these materials are very complicated, relying solely on the database's query retrieval mechanism and statistical methods are far from meeting the display needs. It is urgently required to automatically and intelligently transform data into useful information and knowledge to achieve the purpose of decision-making services. Data mining is a new data processing technology for developing information resources to meet this need and develop rapidly. Data Mining is the extraction of information that is hidden in the prior, but potentially useful, from a large number of incomplete, noisy, fuzzy, and random application data. And the process of knowledge, in a nutshell, data mining is the extraction or "mining" of knowledge from large amounts of data. The research scope of data mining involves association rules mining, classification rule mining, clustering rule mining, trend analysis, outlier analysis, and evolution analysis. The information and knowledge acquired by data mining has been widely used in a variety of applications, including business management, production control, market analysis, engineering design and scientific exploration.

2. Construction of invisible semantic index and text similarity quick calculation model

With the explosive growth of data scale and the increasing number of descriptive features, the dimensions of data sets are getting higher and higher. In such data sets, many traditional methods of information retrieval and pattern recognition are computationally efficient and accurate. Has been severely restricted. Therefore, the reduction in dimensionality brought about by semantic indexing (or feature projection) is crucial for the analysis and processing of text. Among them, LSI can effectively overcome the vocabulary matching problem. The method assumes that there is an underlying implicit semantic structure in the data, which statistically captures the implicit association structure in the text term, and retrieves the concept index and retrieves it in the text collection. LSI is completely unsupervised, and its essence is to detect the most representative features of text, rather than the most distinctive features. In order to apply implicit semantic

DOI: 10.25236/dpaic.2018.033

indexing, text uses the "text-by-term term" matrix xn, d in the vector space model (n is the number of texts, d is the number of words), and performs SVD decomposition from high-dimensional input space to low-dimensional Find a linear map in the implicit space so that most structures can be interpreted in the data.

Through the processing of microblog information by big data technologies such as web crawling, web page cleaning, text segmentation and stop words, a large amount of microblog text with only a small number of feature words can be obtained. Although each Weibo text contains only a small number of feature words, if all Weibo texts are applied to generate a feature vocabulary, the number of feature words it contains will be extremely large, possibly containing hundreds of thousands or even millions. Characteristic words. The large number of feature words will bring great difficulty to the similarity calculation and clustering of the text, so it is necessary to reduce the number of feature words contained in the feature lexicon. At present, the feature word screening methods of the commonly used feature lexicon include word frequency method, information gain method, document frequency method, $\chi2$ statistical method, expected cross entropy method and mutual information method, among which the document frequency method is simple and efficient, and It is more suitable for feature word screening of large-scale text information than several other methods. Literature and proved that the document frequency method has the simplest implementation and the lowest algorithm complexity through calculus and experiment, and its feature word screening effect is similar to other methods. Since the feature words in the microblog feature lexicon are derived from massive microblog information, the feature word screening method selected must have both high efficiency and accuracy. Therefore, the document frequency method is used to filter the feature words. After the feature lexicon is determined, the feature words in the text that are not part of the feature lexicon are deleted according to the feature vocabulary, thereby further reducing the number of feature words of the text, and speeding up the similarity calculation and clustering speed of the text. When calculating the similarity of text, it is necessary to count the number of identical feature words in the two texts by: first comparing each feature word of text with all feature words of text if the feature words are the same, then The comparison value is 1, otherwise it is 0; finally all the alignment values of all the feature words in text 1 are accumulated.

3. Text similarity threshold determination and analysis

The setting of text similarity threshold is one of the key links of text clustering, and its threshold setting directly affects the effect of text clustering. If the text similarity threshold is set properly, the text clustering effect is good; otherwise, the text clustering effect is poor. Due to the high similarity of the microblog texts of the same topic, the similarity of the microblog texts of different themes is lower, and the text similarity threshold is the similarity between the texts of different themes, and the similarity between the texts of the same theme. Inside. In order to quantify the optimal text similarity threshold γ, the simulation of the value of γ is carried out, and the optimal value of γ is determined by simulation results and analysis. Take 4 different sets of data, each group has 1000 microblogging public opinion information, including different lyric topics, and each group contains different lyric themes. In the simulation experiment, the four sets of data are processed separately by big data technology and document frequency method; then the value of γ is set, and according to the constructed text similarity quick calculation model and text autonomous clustering model clustering The microblogging public opinion of the group; Finally, the clustering accuracy rate $P\gamma$ of each group is calculated. If the number of microblog information correctly clustered is $N\gamma$, then $P\gamma = N\gamma /1000$. Through the above simulation operation, the relationship between the clustering accuracy rate $P\gamma$ and the similarity threshold γ of each group can be obtained.

4. Chinese text clustering model

The model is a simple model based on set theory and Boolean algebra. In the Boolean model, the weights i and jW of the index feature iT in the document jD are binary, ie, {0, 1}i jW ∈ , that is, a single text representation can be called a vector in the feature space, in the vector Each component weight is 0 or 1, and this Boolean model is called the Boolean Approach. Due to the ambiguity of weights, the Boolean model can only be used to calculate the relevance of user queries to documents in information retrieval, and it is not possible to use this model to calculate the deeper similarities between two documents. Based on the classical Boolean model, the researchers also proposed the Extended Boolean Approach, so that the correlation can be a number between [0, 1]. Boolean model is a kind of representation model based on set theory and Boolean algebra. Its representation and calculation can be transformed into vectors to achieve equivalent. It is a class vector model.

Probabilistic retrieval model is another mature model in the field of information retrieval, and has achieved good results in many systems. The probability model is an abbreviation of a series of models. It takes into account factors such as word frequency, document frequency and document length. It combines documents and user interests (queries) according to a certain probability relationship, and measures two probability by probability in the probability measure space. The semantic similarity of the text. In information retrieval, P (Relevance | Document, Query) is mainly calculated, and the Probabilistic Ranking Principle (PRP) is used to judge the degree of correlation between different documents and the same query. P(Relevance | Document, Query) represents the probability that the document Document is related to the query for the query Query. According to different assumptions, the calculation formula of P (Relevance | Document, Query) can be used to derive different probability retrieval models. Probabilistic search models include BIR (Binary Independence Retrieval), INQUERY, etc. Among them, the most widely used is the OKAPI model, which has achieved success in the field of information retrieval and has achieved good results in many TREC (Text Retrieval Conference) evaluations.

The language model is also essentially a model based on probability and statistics. In the language model, each document, the entire corpus, and related queries are treated as language models. The distance between the query and the document and the correlation between the document and the document are measured by calculating the distance between the language models. Language models generally fall into two categories in terms of their research direction. One is a rule grammar based on linguistic knowledge, and the other is a statistical language model. At present, the corpus-based statistical language modeling method has become a trend. This method acquires linguistic knowledge in large-scale real corpora by deep processing, statistics, and learning of the corpus. The statistical language model considers language as a kind of probability distribution in the alphabet, and calculates the probability that any sequence of letters becomes a language unit (sentence, paragraph, article, etc.) of the language through the probability distribution. A feature set forms a distribution in a document jD. This probability distribution is called a language model. In language model research, the Language Information Institute of Carnegie Mellon University (CMU) and the Center for Intelligent Information Retrieval of the University of Massachusetts (UMass) The LEMUR system developed in cooperation is an information retrieval system that implements a language model. It has achieved good results in TREC, and this system is also a useful tool for studying language models.

5. Conclusion

In the text similarity calculation and text clustering, considering the different focuses of microblogging public opinion analysis, the two scenarios of unknown lyric theme and lyric theme are known, and the text similarity quick calculation model and text autonomous clustering model are constructed respectively. The experimental results show that: this method can quickly obtain the microblogging public opinion, greatly shorten the time consumption of public opinion acquisition, and to a certain extent meet the urgent need for rapid acquisition of microblogging public opinion, which can provide real-time monitoring and analysis capabilities for network public opinion. Certain method support.

Acknowledgements

Fund Project:

1) This work was partially supported by The Education Department of Jilin province science and technology research project "13th Five-Year" Kyrgyzstan UNESCO Zi [2016] No. 159th (English translation)

2) Jilin Provincial Department of Education “Thirteenth Five-Year” Social Science Research Project, Ji Jiaoke Wenhe [2016] No. 506, Research on Software Professional Training Mode of Independent College

3) Jilin Province Education Science "Thirteenth Five-Year Plan" 2018 annual project, the project approval number is GH181034, the research and practice of the wisdom classroom teaching mode in the "Internet +" era.

4) Jilin Higher Education Society 2018 Higher Education Research Project, Question No.: JGJX2018D415, Research and Practice of Computer Teaching Reform under the Mode of "Internet +" Applied Talents Training

References

[1] Wang Chao, Pan Zhenggao K-mean clustering based on Laplacian matrix [J].Journal of Suihua University, 2017, 37(9):153-155.

[2] Qi Chao, Gu Yu. The grammatical metaphor perspective of ambiguity [J]. Caizhi, 2009, 0(24): 169-169.

[3] Wang Jun, Liu Sanmin, Liu Tao.A Dynamic Classification of Dynamic Data Streams with Noise [J]. Journal of Neijiang Teachers College, 2017, 32(8):51-55.

[4] Ning Li. Innovative Studio Training Mode Broaden the Way of Learning and Producing Talents [J]. Journal of Nanning Vocational and Technical College, 2017, 22(4).

[5] Li Wei, Wang Rujuan. A review of the research on implicit semantic index in big data environment [J]. Electronic test, 2018, 0 (14): 115-116.