Advanced Science and Technology Letters
Vol.134 (MAS 2016), pp.41-47
http://dx.doi.org/10.14257/astl.2016.134.07

Feature weighting classification algorithm in the application of text data processing research

Zhou Chengyi

University of Science and Technology Liaoning,
Anshan, Liaoning, China, 114053
askdccy@126.com

Abstract. With the development of network technology, the text information resources also exploded growth. How to store the massive text data, classification and retrieval become a growing concern. So will feature weighting algorithm is introduced into the text classification data processing, data noise reduction processing, using the word to word frequency combination analysis, and through the word relevance computation method to calculate feature weight. Through experimental analysis, the method can make the calculation more accurate, more effective classification effect.

Keywords: Feature weighting; Classification algorithms; Text processing

1 Introduction

The emergence of the Internet have created a new era, search information enables people to retrieve the information all over the globe, extended to various fields. Humans have to roam in the ocean of knowledge, but it is difficult to obtain their knowledge. In the process of text categorization, no rules text after word segmentation processing into a regular word set, and then after processing of dimension reduction algorithm can to debug training classifier using the algorithm and the results of evaluation. Text classification application range is very wide. Below in text classification in information retrieval and analysis of emotional huge role.
Therefore, as a hot research topic in data mining, text categorization makes maximum utilization of information, to promote the interconnection. The development of the network has a profound realistic effect [1-3].

2 Related works

2.1 Text representation

Information retrieval is the theory basis of text categorization, the trained classifier to no category marked text classification. Text classification flow diagram as shown below:

ISSN: 2287-1233 ASTL

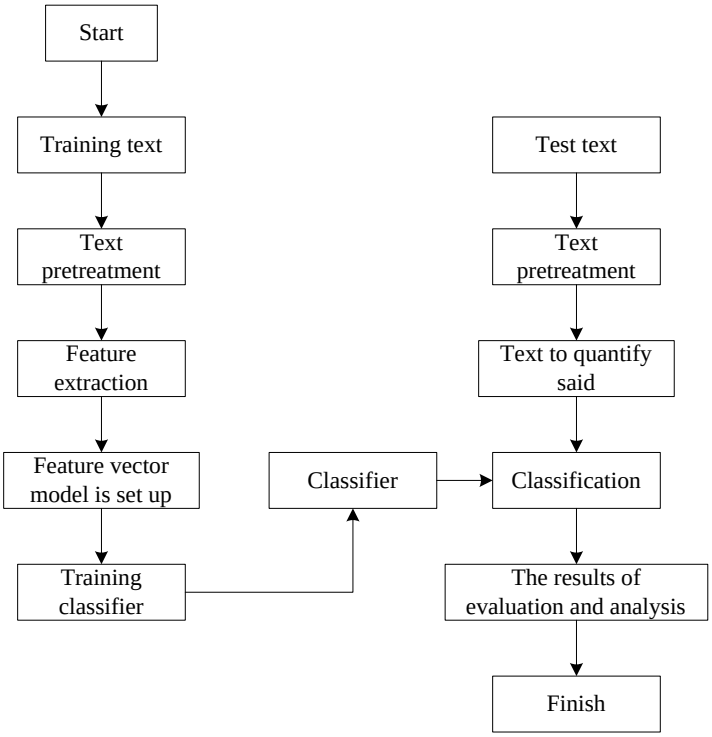

Fig. 1. Text classification process in general

From the Internet to get the text style each different, the computer could not identify the original text, and machine learning algorithms are used for the text, a can apply these texts to text classification, so the first step should be carried out on the text pretreatment, the original text into a particular style, the form of the computer can recognize. The current text representation model is diversiform, but in the aspect of data mining and machine learning, is widely used model there are three main kinds, the following will, in turn, is introduced, through the establishment of the text model, realize to the processing of the original text [4-5].

(1) The Boolean logical model

Boolean logic model is the most easy to understand the text classification model, whether it to text contains the feature as the judgment standard, has the characteristics of the term in the text, its weight value is "1", or "0",The advantages of Boolean logic model:1) The text said relative to other model is simple, no complicated calculation process;2) simple calculation and fast algorithm;

(2)The vector space model

Text representation model of vector space model is put forward by Gerald Salton put forward in 1975.The model USES mathematical vector to realize the text said, the vector is mainly consists of text feature item and feature of the weight of two parts. Among them, the text category contribution ability evaluation with the weight of the size of the measure.

(3) Probability reasoning model

Text representation model is still have a kind of common probability reasoning model, can be divided into the dual independent probability model, the double poison distribution probability model, regression model and so on. The main idea of this model is: through calculation according to the probability to reduce the order to sort on the text, in order to distinguish belongs in the category of the text. It will text of probability calculation and text to the user's attraction, the text semantic added to the formula to calculate the model [6-8].

2.2 Weighting

Build a complete text representation model, to appear in the text of the feature to complete the calculation of weight, to a certain text vectorization. Text representation model, corresponding to different ways of calculation, but the general principle is: if the text of the characteristics of "participation" is higher, it to the distinction between the text category contribution rate is bigger, the weight value is set, if all text in a feature of "participation" is high, the text of the category to distinguish the contribution rate is small, weight value is set. And the calculation method is affected by the length of the text, the longer the length of the text, the more occurrences may feature, some irrelevant feature weights may give larger value, but some less occurrences, but for classification plays a very important role in the characteristics of a given less weight, and thus influence classification effect, so the word frequency weights were not used widely [9-10].

3 Feature weighting classification algorithms

3.1 Feature dimension reduction algorithm

Text representation model is completed, the generated is a high dimensional vector space, dimension may be hundreds of thousands, millions, in the high dimensional vector of each dimension represents a feature weights. Part features selected from the high dimensional space vector points vector into a new space vector, namely the original space vector filter, so as to realize the transformation of high-dimensional to low dimension, retained during the transformation characteristics of classification

work, remove the classification characteristics of the role of is not very big.Here are several kinds of commonly used feature selection methods [11-13].

(1) Document frequency

Document frequency algorithm idea is: to measure the size of the contribution of a feature item category is in a category based on the characteristics of the text of the number of how many. But a limitation lies in ignoring the contribution rates of text categorization, some rare word comes once in a document, there may be some word or two, but a very important part in the work of text classification, if simply rely on DF, delete the words, may affect the classification effect.

(2) The information gain

IG algorithm for text in the light of characteristics of each item for computing, which looks one feature in text. Limitation lies in the calculation process is complicated, however, need to each feature calculation, global investigation characteristics affect classification system, they cannot be calculated separately for a category.

3.2 Text classification algorithm

(1) Simple Bayesian algorithm

Naive Bayes algorithm has the advantage of due to the assumption of text feature item and there is no relationship between a feature, so the calculation is not trival, is not sensitive to missing data.

(2)Support vector machine (SVM) algorithm

Support vector machine (SVM) algorithm is put forward in 1995 by Vapnik machine learning algorithms, the algorithm with the theory of structural risk minimization principle and VC knowledge support, is mainly used to solve binary classification problems.

(3) K-nearest neighbor node algorithm

K - nearest neighbor node algorithm is put forward in 1968 by Cover and Hart of machine learning algorithms. KNN algorithm principle is: to be classified by calculation the text of the text and training focus close degree between the text, find out the nearest to text classification k text, observe the k most belongs to which category in the text of the document, is to be attributed to the category classification of text [14-15].

4 Experiment and result analysis

4.1 Word frequency combined

Traditional dimension reduction algorithms such as mutual information, information gain, such as principal component analysis without considering the relationship between the words, simply by counting out some vector, this is likely to get rid of some low occurrences but useful for classification of words. Application Lin words synonym word similarity calculation method to realize feature of merger, to a certain extent, to strengthen the feature weights, weakening the characteristics of weight, so that you can put the feature representation of ascension to the "concept" level. Use of word processing, Lin text further reduce the dimension of the space vector, it makes calculation more accurate.

4.2 Word relevance computation and data noise reduction

Of traditional text classification only simple text preprocessing of the test, then the trained classifier to classify, this treatment can produce error, the result of the classification in word frequency based on the analysis of the combined use of how net words correlation algorithm for calculation, the characteristics of correlation between the characteristics of the text of the test vector weights assignment again, so that the calculation of similarity between the two text more accurate. In order to improve the classification accuracy, the corpus is the noise reduction processing: remove duplicate files under the training set and testing set file; Delete the training set and testing set C35 - Law under the file folder; Delete all the length is less than 400 document; The text renumbered, facilitate the realization of the classification.

4.3 The experiment design and analysis

Experimental data using computer information and technology department of fudan university center for international database natural language processing group of corpus, among them, the training sets the text, Answer to test the text, corpus contains 10 classes, before and after noise reduction processing document type distribution shown in the table below:

Table 1. Before and after noise reduction processing document type distribution

Category	Train		Answer	
	Before processing	After processing	Before processing	After processing
Art	735	413	722	410
Literature	36	0	38	0
Education	51	0	63	0
Philosophy	33	0	41	0
History	426	410	429	405

Space	540	470	622	484
Energy	36	0	35	0
Electronics	25	0	26	0
Communication	29	0	28	0
Computer	1259	962	1258	934

Experiment with more than 9 classes as experimental object, text categorization effect evaluation standard of the Precision, Recall, F-score three values. With not learning support vector machine algorithm for data processing, naive Bayesian classification algorithm, neural network classification algorithm, K neighbor algorithm with the processed data to study the four algorithms. Only a few data due to sample selection problem decline phenomenon, but the overall algorithm adaptability is good, to achieve the desired effect.

5 Conclusion

Based on synonym word Lin and hownet knowledge under the condition of sufficient research, puts forward a new algorithm of text data processing: the text based on semantic feature weighted classification algorithm, this algorithm first word for word Lin merger analysis, on this basis, using hownet words correlation feature weight assignment again, it makes calculation more accurate, and through the experiment.

References

[1] Kaliszyk, C., Urban, J.: Stronger Automation for Flyspeck by Feature Weighting and Strategy Evolution [J]. Blanchette J.pxtp .third International Workshop on Proof Exchange for Theorem Proving, 2013:87-95.
[2] Xu, Z., Yang, Y., Tsang, I.: Feature Weighting via Optimal Thresholding for Video Analysis[C]// IEEE International Conference on Computer Vision. IEEE, 2013:3440-3447.
[3] Chai, J., Chen, H., Huang, L.: Maximum margin multiple-instance feature weighting [J]. Pattern Recognition, 2014, 47(6):2091-2103.
[4] Zhi, XB., Fan, JL., Zhao, F.: Robust local feature weighting hard c-means clustering algorithm[J]. Neurocomputing, 2014, 134(4):20-29.
[5] Ismail, M.M.B, Frigui, H.: Unsupervised clustering and feature weighting based on Generalized Dirichlet mixture modeling [J]. Information Sciences, 2014, 274(274):35-54.
[6] Davide, A., Carlotta, D.F., Duccio, C.: Explaining diversity in metagenomic datasets by phylogenetic-based feature weighting.[J]. Plos Computational Biology, 2015, 11(3).
[7] Dorksen, H., Lohweg, V.: Combinatorial Refinement of Feature Weighting for Linear Classification.[C]// Emerging Technology and Factory Automation (ETFA), 2014 IEEE. IEEE, 2014:1-7.
[8] Nazari, M., Shanbehzadeh, J., Sarrafzadeh, A.: Fuzzy C-means based on Automated Variable Feature Weighting [J]. Lecture Notes in Engineering & Computer Science, 2013, 2202(1).

[9] Valencia, A:, Verspoor, K.: BioC: a minimalist approach to interoperability for biomedical text processing[J]. Database the Journal of Biological Databases & Curation, 2013, 2013 (3)
[10] Davidson, DR., Ozer, A.: Automatic language identification for dynamic text processing: US, US 8464150 B2 [P]. 2013.
[11] Yessenov, K., Tulsiani, S., Menon, A.: A colorful approach to text processing by example[C]// Acm Symposium on User Interface Software & Technology. 2013:495-504.
[12] Sabour, S.: Text processing in information retrieval system using vector space model[C]// Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, 2014:1 - 6.
[13] Hills, CS., Pancaroglu, R., Duchaine, B., Word and Text Processing in Acquired Prosopagnosia [J]. Annals of Neurology, 2015, 78(2):258-271.
[14] Rajdho, A., Biba, M.: Plugging Text Processing and Mining in a Cloud Computing Framework [M]// Internet of Things and Inter-cooperative Computational Technologies for Collective Intelligence. Springer Berlin Heidelberg, 2013:369-390.
[15] Bredikhin, AY., Sergeichev, NE.: Method for automated text processing and computer device for implementing said method: , US20150293902[P]. 2015.

