International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.183-192
http://dx.doi.org/10.14257/ijdta.2016.9.8.18

Weighted Content Feature Text Recognition Algorithm Research

Zhou Chengyi

University of Science and Technology Liaoning,
Anshan, Liaoning, China, 114053
E-mail: askdccy@126.com

Abstract

Text data classification and retrieval is an important field of artificial intelligence, but because of the different characteristics and different language syntax, classification and retrieval also will be different; classification and identification retrieval traditional text data, using training data and segmentation the algorithm does not consider the specific locale, it is often an error. To solve this problem, we propose a weighted feature-based classification method, according to this method, the text data can be quickly and accurately classify; experiments show that the proposed algorithm can effectively improve the accuracy and speed of classification and retrieval.

Keywords*: Feature weighting; Classification algorithms; Text processing; Text Recognition*

1. Introduction

With the continuous development of information technology, in particular the popularity of Internet applications, electronic document dramatically increased, how to effectively organize and manage these vast amounts of information, and can quickly and accurately obtain the information users need in today's information technology sciences a major challenge. One of the effective management of electronic text is text classification method. Text classification is an important intelligent information processing technology, information filtering, of information retrieval, text databases and digital libraries and other great value. A text (not following basic distinction between "text" and "Document" two words meaning) is to issue a document classified into several predefined categories of one or a few, while the automatic classification of text is to use a computer program to implement such a classification [1-4].

First, the category system used to classify what is needed is predetermined. Such as Sina news classification system, Yahoo! Site Map classification level. Once this classification level, for a long period of time it is immutable, or even if you want to change, have to pay a high price (as much as down and rebuild a basic classification system).

Second, a document does not restrict only be assigned to a category. This classification of the problem concerning the subjectivity, for example, find a personal judgment relating to article 10 stated what part of the financial, banking or fiscal policy in the field, 10 people may give 10 different answers, so the article is likely to which they are assigned to multiple categories, but points to certain categories of convincing, and some people feel ambiguous Bale (confidence is not the same).

Of course, the real heavy use of text classification, is still based on the article relating to classification and, accordingly, to build up the system, comes as the search engines. Neri because of course, self-evident, I just wanted to give everyone mind you, text classification is not exactly the same web page classification. Information page contains far contained therein text (text) much more information on a web page classification, in addition to considering the contents of the text classification outside the chain into a chain

ISSN: 2005-4270 IJDTA

link information, the page file metadata itself, even this page contains a website structure and themes, can provide great help to the sort (such as Sina Sports Gallery pages are no doubt about sports), and therefore said text classification is actually a subset of the pages are also classified not too much [5-7]. Of course, a purely text categorization and page classification system is not a little difference at all. There is an important prerequisite for text classification: that can only be classified according to the text of the article, and not by means of encoding formats such as files, the article author, publication date and other information. And this information is often available on the web page, and sometimes the role is still great! So pure text classification systems in order to achieve comparable classification results, we must work hard on the theoretical basis and technical content itself.

In addition to search engines, systems and mass text messages to deal with such as digital libraries, archives management, *etc*., are the lingua franca of text classification. Text classification and other classification problem is not fundamentally different, the method can be attributed to certain characteristics to be classified according to the data to be matched, of course, an exact match is not possible, it must be (according to some evaluation criteria) selection optimal matching result, thus completing the classification.

Therefore, the core of the problem will be transformed into what features indicated with a text in order to ensure effective and rapid classification (note that both areas of need are often contradictory). So that its own text classification system since, has been a dominant feature different options for different methods factions.

The first matching word appears only if the class name and the same word (at most adding synonyms processing) to determine whether the document belongs to a category according to the document. Obviously, this simplistic classification methods cannot bring good results.

Later, the rise of the method over a period of time with the aid of knowledge engineering professional help, define a large number of inference rules for each category, if a document to meet these inference rules, it can be determined that category. Here the degree of matching a particular rule has become a feature of the text. Because the system is added to human judgment factor, much higher than the accuracy of matching words. But the disadvantage of this approach is still evident, such as quality classification depends heavily on the quality of these rules, which is dependent on the rules of the "people" is good or bad; another example rule-making is an expert level, labor costs significantly rise often unbearable; and knowledge engineering is the most fatal weakness replicability do not have any, for the financial sector to build a classification system, if you want to expand to the medical or social insurance, and other related fields, in addition to completely reinvent the wheel outside the no other way, often resulting in a huge waste of money and knowledge [9-10].

people realized what basis to determine what features should be attached to the text of the problem category, and even human beings themselves are not answered clearly, there are too many so-called "can be felt not explain in words," stuff in there. Most human judgment based on experience and intuition, so naturally someone will think of how to make a machine like humans themselves to their own lessons by a large number of similar documents to observe, as a future basis for classification.

Text classification application range is very wide. Below in text classification in information retrieval and analysis of emotional huge role [11-12], for example.

(1) Information retrieval

Text classification effectively raised the speed of the information retrieval, in text classification, information retrieval need artificial to complete, need to seek knowledge in artificial in the middle of a vast knowledge base, time-consuming, and the emergence of text classification for information retrieval plays a supplementary role, can make use of text classification determine the type of the query may beforehand, by narrowing the scope of the text of the query, the applicability of the information retrieval for text

classification also provides a good validation, in general, text classification can improve the accuracy of information retrieval, the reliability of text categorization provides a good verification.

(2) Sentiment analysis

With the advent of the era of network, the content on the web in particular the social networking site all kinds of text information organization and management is becoming more and more difficult, at the same time, the people to the emotional orientation of various comments on the website of the information classification gave more and more attention and interest. Sentiment analysis technology mainly expressed by commentators in the article the subjective emotion as a classification of objects, can automatically for reviewers to express the subjective emotion tendency to make judgments are for or against, for example, so that you can through the emotional tendencies and timely grasp the social attitude towards the matter of public opinion, public opinion. Specifically the method can be used to calculate the reviewers emotional attitude to his commentary point of view, in this article as a bridge, reviewers can be calculated and published the emotional intensity and emotion between the polarity, further reviewers can be calculated and then the similarity of the values between the submitter, emotional similarity and so on. It can realize through the analysis of the emotional to the division of social groups and tracking public opinion [13-15].

2. Related Works

2.1. Text Categorization

Text classification flow diagram as shown below:

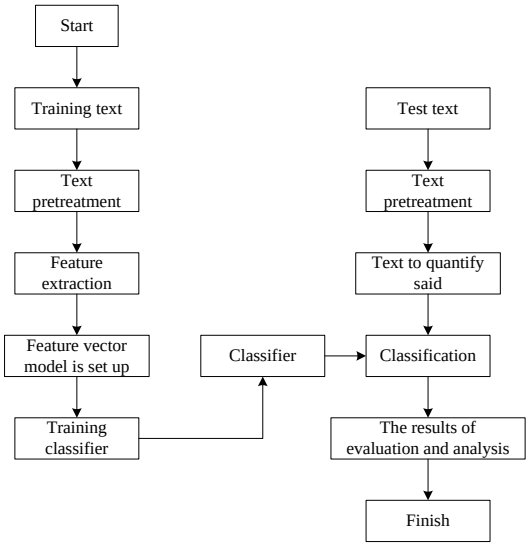


Figure 1. Text Classification Process in General

Text classification typically includes the expression of the text, the classifier selection and training, evaluation and classification results and other feedback process, in which the expression of the text can be divided into text preprocessing, indexes and statistics, feature extraction and other steps. Overall functional modules text classification system is as follows:

(1) Pre-processing: the original corpus formatted in the same format, to facilitate the subsequent unification process;

(2) Index: The document is divided into the basic processing unit while reducing the

cost of the follow-up process;

(3) Statistics: Frequency statistics, items (words, concepts) associated with the probability of classification;

(4) Feature extraction: extract the document reflects the theme of the features from the document;

(5) Classifier: classifier training;

(6) Evaluation: Test results classifier analysis.

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. Here are several common weight calculation method.

(1)The Boolean weighted

Boolean weighted, corresponding to the Boolean logical model, it with a characteristic in this text with or without as the judgment standard, the text has the characteristics of the item, don't consider it is appear multiple times in the text, it has a weight of 1, don't consider whether it appears in the core of a text paragraphs. Otherwise as "0", the weight of the Boolean calculation formula is as follows:

$$w_i = \begin{cases} 1, TF_i > 0 \\ 0, TF_i = 0 \end{cases} \quad (1)$$

Among them, the w_i said feature weights of t_i in the text, TF_i said text feature t_i is contained in the number of how many, this will tell you, if the characteristics in the text, whether once, twice, or many times, weight value is 1.

As you can see, the weight of Boolean expressions is concise, achieving in text categorization is not difficult, but the characteristic value of weight value is only 1 s and 0 s, did not see the text categories of contribution rate of characteristic value, so the weight of this calculation method is only suitable for simple text, is not widely used [8-9].

(2) Word frequency weighting

Word frequency of word frequency weighting indicates that text contains the characteristics of the number of times, it eliminates the Boolean weighted calculation does not consider feature occurrences of faults, using word frequency to measure the characteristic to the distinction between categories of contribution. So, in an article appears the characteristics of the item number is more, it the word of weight is bigger, computation formula is as follows:

$$w_i = TF \quad (2)$$

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. If the higher dimensional vector directly applied to text categorization algorithm, the whole process requires a lot of time, and the effect is often poor, in order to improve the classification accuracy and speed of classification, dimension must be carried out. Different characteristics study of the role of the different text classification, feature dimension reduction to d remove makes little

sense to text classification, keep work on text categorization of d. Here are two kinds of commonly used algorithm for dimension.

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. If training the feature item in the text of the "attendance" is low, show that the feature of categories of contribution rate is low, if the feature item in the text of the training "attendance" is higher, shows that the feature of categories of contribution rate is high, set up a digital measure, when DF values less than this value, just delete those characteristics, because of these characteristics items without the "representative" to set a threshold, when the DF value is greater than the threshold, also should make delete, because they are no "distinction" for the text.

DF algorithm has the advantage of calculation are not so tedious, easy to operate; 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

T_i, classification system in the case of contain it and didn't include it how much is the amount of information, and then two Numbers do bad, the resulting value is the gain, said characteristics of the effect of the classification system, through the calculation of the value, indicating characteristics of t_i on the size of the contribution of text category.

IG algorithm has the advantage of presence or absence of a feature item all reflected in the calculation formula, the results of the calculation more accurate; 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 based on bayes' theorem, is to use statistical methods of text categorization, the algorithm used is to assume that the characteristics of the text between the items were independent of each other, using probability text category. Text of the final category is the category of the designated by the maximum probability is. Naive bayes algorithm process is as follows:

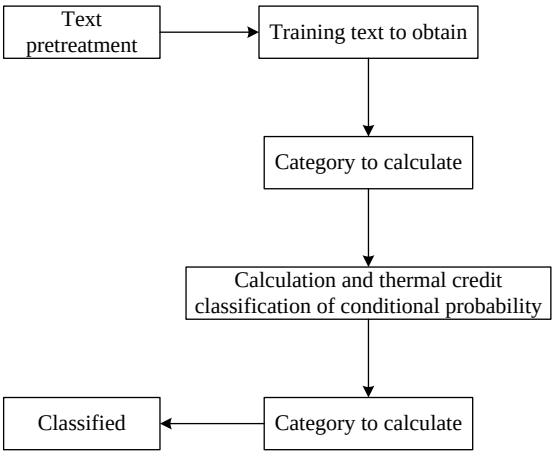


Figure 2. Naive Bayes Algorithm Process

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. But the limitation is that in many cases can't ignore the relationship between a text characteristic, all the algorithm has great limitations in practical application.

(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. Training the SVM algorithm the basic idea is: the original text through a nonlinear mapping relationship, vector projection to the high dimension space, mapping relation can be defined by inner product function, and then in the high dimension vector space to look for a "hyperplane", one kind of sample is separated from other types of samples, converted into a binary classification problem.

The advantage of the SVM algorithm is suitable for solving the problem of high-dimensional, and can improve the generalization performance, under the condition of limited samples can get global optimal solution, and does not exist the problem of dimension reduction; But the limitation is sensitive to missing data, no general solution to the nonlinear problem.

(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. The advantage of KNN algorithm is run not cumbersome, no parameters to judge; But the limitation is that the k value of the selected long time consuming, but compared with other algorithms, KNN algorithm is simple, has been widely used.

(4)Neural network algorithm

Neural network algorithm is a kind of with people's perception of the characteristics of the neural network based on text categorization algorithm, by McCulfoch put forward in 1943.NN algorithm the basic idea is: the stage of training to complete the connection weights between input and output of optimization test, to obtain the optimal classification effect; Classification stage main text is to be classified by the input layer to the appropriate output of text classification. BP neural network algorithm is one of the representative, as shown in the figure below, a total of three layers, including the input layer to receive information from the outside world, the middle layer receives input layer

passed signals, and processing; Output layer classification results [14-15].

The number of neurons by constantly trying to find the optimal value. The connection between the three layers of weight, is the goal of the BP neural network training, through continuous training, get the optimal value. Training process continuously testing, if the output is correct, then the weights are unchanged, or continue to test, until they get the right results, such as the ownership of the heavy test, you can apply the neural network for text classification. Neural network algorithm has the advantage of distributed parallel processing ability, to noise of high robustness and adaptability, the limitation lies in the training process, however, requires a lot of parameters, the learning process is relatively slow.

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 hownet 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.

Experimental data using computer information and technology department of fudan university international center database natural language processing group of corpus 2, but the following problems: (1) the corpus contains two parts of the training set and test set, contains more than 9000 documents, respectively, but there were nearly 1500 documents to repeat;(2) the training set and test set C35 - parts of the Law document already through word segmentation processing, but the segmentation result is poor;(3) some articles only head, not the actual content, some categories exist empty document;(4) some short document because keywords after word too little, the classification of basic doesn't work.

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, experimental results on the corpus as Table 2:

Table 2. Each Classification Algorithm Classification Result Table

Classification algorithm		Precision	Recall	F-score
SVM	Before processing	0.91	0.83	0.85
	After processing	0.76	0.85	0.87
NB	Before processing	0.74	0.76	0.78
	After processing	0.78	0.77	0.79
NN	Before processing	0.75	0.72	0.76
	After processing	0.77	0.70	0.78
KNN	Before processing	0.88	0.81	0.86
	After processing	0.90	0.84	0.89

Can be seen from Table 2, compared with traditional data processing algorithm, to deal with the text data, the algorithm in the Precision, Recall, F-score three indicators on most improved. Only a few data due to sample selection problem decline phenomenon, but the overall algorithm adaptability is good, to achieve the desired effect.

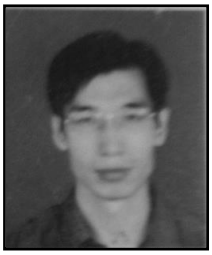
5. Conclusion

According to data the difference value synonyms phrases and a large text database mining, a weighted classification method based on semantic features, the method based on characteristics of the weighted value of the analysis of the text meaning of the phrase, and the characteristic weight values are based on the contents of the database reconfiguration, making computing more accurate data classification, and its low computational complexity algorithms for online text recognition and classification, test the validity and accuracy of the algorithm.

References

[1] C. Kaliszyk, "Urban J. Stronger Automation for Flyspeck by Feature Weighting and Strategy Evolution", Blanchette J.pxtp. third International Workshop on Proof Exchange for Theorem Proving, **(2013)**, pp. 87-95.
[2] Z. Xu, Y. Yang and I. Tsang, "Feature Weighting via Optimal Thresholding for Video Analysis", IEEE International Conference on Computer Vision. IEEE, **(2013)**, pp. 3440-3447.
[3] J. Chai, H. Chen and L. Huang, "Maximum margin multiple-instance feature weighting", Pattern Recognition, vol. 47, no. 6, **(2014)**, pp. 2091-2103.
[4] X. B. Zhi, J. L. Fan and F. Zhao, "Robust local feature weighting hard c-means clustering algorithm", Neurocomputing, vol. 134, no. 4, **(2014)**, pp. 20-29.
[5] M. M. B. Ismail and H. Frigui, "Unsupervised clustering and feature weighting based on Generalized Dirichlet mixture modeling", Information Sciences, vol. 274, no. 274, **(2014)**, pp. 35-54.
[6] A. Davide, D. F. Carlotta and C. Duccio, "Explaining diversity in metagenomic datasets by phylogenetic-based feature weighting", Plos Computational Biology, vol. 11, no. 3, **(2015)**.
[7] H. Dorksen and V. Lohweg, "Combinatorial Refinement of Feature Weighting for Linear Classification", Emerging Technology and Factory Automation (ETFA), 2014 IEEE. IEEE, **(2014)**, pp. 1-7.
[8] M. Nazari, J. Shanbehzadeh and A. Sarrafzadeh, "Fuzzy C-means based on Automated Variable Feature Weighting", Lecture Notes in Engineering & Computer Science, vol. 2202, no. 1, **(2013)**.
[9] A. Valencia and K. Verspoor, "BioC: a minimalist approach to interoperability for biomedical text processing", Database the Journal of Biological Databases & Curation, vol. 2013, no. 3, **(2013)**.
[10] D. R. Davidson and A. Ozer, "Automatic language identification for dynamic text processing: US", US 8464150 B2[P], **(2013)**.
[11] K. Yessenov, S. Tulsiani and A. Menon, "A colorful approach to text processing by example", Acm Symposium on User Interface Software & Technology, **(2013)**, pp. 495-504.
[12] S. Sabour, "Text processing in information retrieval system using vector space model", Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, **(2014)**, pp. 1-6.
[13] C. S. Hills, R. Pancaroglu and B. Duchaine, "Word and Text Processing in Acquired Prosopagnosia", Annals of Neurology, vol. 78, no. 2, **(2015)**, pp. 258-271.
[14] A. Rajdho and M. Biba, "Plugging Text Processing and Mining in a Cloud Computing Framework", Internet of Things and Inter-cooperative Computational Technologies for Collective Intelligence. Springer Berlin Heidelberg, **(2013)**, pp. 369-390.
[15] A. Y. Bredikhin and N. E. Sergeichev, "Method for Automated Text Processing and Computer Device for Implementing Said Method", US20150293902 [P], **(2015)**.

Author



Zhou Chengyi, received the B. Sc degree in math from Liaoning University and the M. Eng degree in University of Science and Technology Liaoning, CHINA in 1985 and 2007 respectively. He is currently engaged in teaching. He is currently researching on Data Mining, Formal Concept Analysis.