Research on Bayes-based Text Automatic Classification

Xuan Zhang [1, a*]

[1]Office of Academic Affairs, Tianjin Electronic Information College, Tianjin China

[a] zhangxuanlive@126.com

Keywords: text automatic classification; Bayes; classification algorithms; feature extraction

Abstract. Enormous amount of information on the Internet, there are several of information and it is so complicated. Information retrieval is of blind and too much redundant information is in search results. In order for a user to much more effective at getting the information they needed, This paper researches the method of page text automatic classification based on the classification algorithm of Naive Bayes. Responding to the structure of pages, the paper analyses the structure components which are useful to the classification in the page tags in detail. And we apply Naive Bayes algorithm to classify with these effective features of HTML identifiers. It easy for users to more precise locate information on Internet through reduced the difficulty of Internet information retrieval.

INTRODUCTION

Now, there are tens of thousands of WWW Servers on the Internet, and it stores vast amounts of information resources. Searching for information across many different Web sites and search engines makes so much pages information return the users [1]. It is very important to let the user know which page information is useful and which are outmoded or less useful [2].To register information with an interest in this type in advance, we can learn the users' interests from the user's historic behavior on Web by used some methods [3]. And if somebody publishes related information on Web, the personal information will push to the user immediately. This is known as "On-demand information services". Every user has its own particular information needs [4-5].

Text classification determines the category taxonomy to the documents in the document collections by the predefined topic category [6]. On the one hand, text automatic classification could establish the database corresponding to page category, and query database with classification. It can improve the recall and precision of information. On the other hand, it also can set up automatic categorized information resource and offer the classified information catalogue to users.

TEXT CATEGORY

The common practice of text automatic classification is: assigns pre-defined text category, and provides pre-classified text for each category (called training documents). The classification system trains the text to learn website classified knowledge. When the text needs to be classified, the system will determine the category according to knowledge base contents. Automatic classification is to say text classification; it will divide the unclassified text into multiple categories by using classification methods. Automatic classification does not need to train the set of texts, the categories are not certain. The Automatic text classification which discussed in this paper is defined as the task to assign pre-defined category labels to documents.

the Process of Automatic Classification

The process of the Automatic text classification is: extract a set of keywords that characterize the document and represent the documents as a same standardization. It judges text respective characteristic category by use classifiers. The classifiers is the core of the classification system, and it will continuous improvement and perfection by studying. The classifier category includes: C as classes sets, T as the collection of all documents, D as the collection of train documents, S as the vector space of documents characteristics, d as documents, R as T maps to S, U as D maps to C.

Large numbers of text on Web page will be summarized, classification and clustering by text classification. It also could improve the performance of web text mining. Web text mining extracts the information to constitute feature set of web pages. And the feature set is classified and summarized to categorized view by some methods.

Text Classification Model

The text classification consist training module and classification module. In the training module, training documents are vectored to get a collection of features. Features subset extracting algorithm extracts an optimal features subset from the features collection. Whether the features subset is optimal is subject to verification by evaluation algorithm. The evaluation algorithm lets us sort the training text which represented by features subset according to classifier, and evaluates the performance of sorting performance. In the classification module, test text is expressed in optimal features subset, and then classifier by classifier. The text classification model is showed in figure 1.

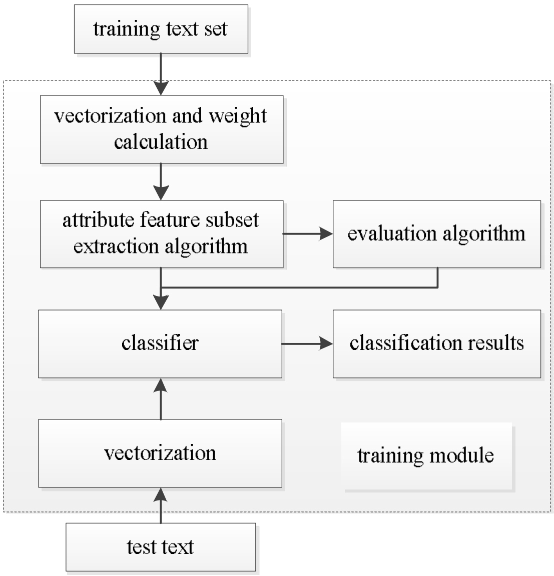

Fig. 1. The text classification model

STUDY OF TEXT CLASSIFICATION

the Method of Text Automatic Classification

The primary task of the research on page text classification is to determine the required web site, and collect a set of these sites. Then these site would be detect whether they have information labels and valid. The common use is collecting URL information which are login frequently. It can cover more sites. The another advantage of daily data collection is better reflecting real user's behavior. On this basis, this paper achieve the same effect as simulation through manually input the training set and test set. We should say, the label testing is only in the sites of training set.

Text Classification with Bayes Algorithm

The simple Bayes model assumes the components of feature vector are relatively independent relative to the decision variables. Each component independently plays a role to decision variables. Although this assume has somewhat limited the scope of the application of simple Bayes model. But in a practical application, not only it is less complex exponentially of setup of Bayes net, but also Bayes is strong and effective even violate the hypothetical constraints in many fields. Bayes says that how to forecast unknown samples with given training samples. This is a basis for forecasting: get the category with maximal posteriori probability.

The basic idea of Bayes algorithm is calculating the probability of web pages classification. The probability is a synthetic expression of the probabilities of every word that in the web pages classification. The process of Bayes algorithm is:

- The training pages and test pages separately use pre-process and separate words (using Automatic cutting word dictionary) to extract feature.

- To calculate the rate vector of feature words which belongs to a category . The formula is shown in formula(1).

$$w_k = P\left(w_k \mid C_j\right) = \frac{1+\sum_{i=1}^{|D|} N\left(w_s, d_i\right)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N\left(w_s, d_i\right)} \qquad (1)$$

In formula(1), $|D|$ is the number of training pages, $N\left(w_s, d_i\right)$ is the word w's frequency in "di", $|V|$ is the total number of words.

- Bayes algorithm calculate the page's "di" follow the formula(2) according to feature words When the new page comes. The formula is shown in formula(2).

$$P\left(C_j \mid d_i;q\right) = \frac{P\left(C_j \mid q\right)\Pi_{k=1}^{n} P\left(w_k \mid C_j;q\right)^{N(w_k,d_i)}}{\sum_{r=1}^{|C|} P\left(C_r \mid q\right)\Pi_{k=1}^{n} P\left(w_k \mid C_j;q\right)^{N(w_k,d_i)}} \qquad (2)$$

In formula(2), $P\left(C_j \mid q\right) = C_j$ is the number of training pages/the total number of training pages, $P\left(C_r \mid q\right)$ is similar meaning, $|C|$ is the total number of classify, $N\left(w_k, d_i\right)$ is w_k's word frequency in "di".

EFFECT OF SYSTEM

There are two important indexes to evaluate and test of the text automatic classification algorithm: recall ratio and precision ratio. The formula is shown in formula(3).

$$Recall\left(T\right) = CorrectTextsNum / AllTextsNum \qquad (3)$$

In formula(3), $CorrectTextsNum$ is the text number of class C which is correctly unclassified by classification algorithm. $AllTextsNum$ is the text number of class C which is unclassified. The formula is shown in formula(4).

$$Precision\left(T\right) = CorrectTextsNum / ClassifyTextsNum \qquad (4)$$

The TABLE I lists the effect of the description web pages on classification algorithm for selected page samples are trained.

TABLE I. RECOGNITION RESULTS

Content	Recognition Rate
Method	90.6%
Text Word Frequency	91.6%
Word Frequency + Heading	89.3%
Word Frequency + Description	90.7%
Word Frequency +Key Word	91.5%
Word Frequency + Hyperlink	86.2%
Word Frequency +all Description Info	87.3%

From TABLE I we can see that using the frequency and description of web pages through weighted adjustment. When training and recognizing the web text, increase as the heading, description, keywords and hyperlink for such a high weight to improve the recognition.

Lots of pages collect from Web-Searcher use 150 page samples of the class of computer, Economic and sport. In the course of the experiment, we can use the parameter of a=1, b=1, c=1 and d=1. To make classification experiments with all test samples, the performance test results of the text automatic classification as shown in the TABLE II.

TABLE II. THE RESULT OF EXPERIMENT

Cass	Computer	Economic	Sport
Correct	38	43	35
Classify	45	48	40
All Text	50	50	50
Recall	76%	86%	70%
Precision	84%	90%	88%

For some reason, the text doesn't be classified and the pages quantity is small, we can classify them by manual method. From the experimental results can be seen that the recall ratio and precision ratio of text automatic classification both have high value in terms of obvious features subject. It shows high performance. So the method of text automatic classification is feasible.

CONCLUSION

The text automatic classification demands an ordered organization with the text. It organizes similar and related text together to provide more efficient searching and more accurate search result. In this paper, based on the existed methods of text information and feature extraction, we design and realize a simple and practical algorithm for feature extraction based on word feasibility, and obtained better result. The experiment results show that the algorithm is reliable. In this paper, we focus on the influence of web pages and discuss the lack of existed methods of weight calculation for text. The opinions on and direction of improvements are given. This paper focuses on feature extraction algorithm and weights calculating method for web text. Finally, it should also be pointed out that set reasonable classification can be solved the problem of text features bottlenecks to some extent. But due to the ever-changing language, classification cross is inevitable. We can consider revealing the deep-seated features. It needs use the method of linguistics. It is becoming an attractive research area to combine linguistic knowledge with statistical methods.

References

[1] Zhao Da-peng, Research on the Vector Space Model Based Text Automatic Classification System, 7 (2013) 381-388.

[2] Elmarhoumy M, Abdel Fattah M, Suzuki M, Ren F, A new modified centroid classifier approach for automatic text classification, 8 (2013) 52-58.

[3] Luo Xi, Ohyama Wataru, Wakabayashi Tetsushi, Kimura Fumitaka, Automatic chinese text classification using character-based and word-based approach, Proc. Int. Conf. Doc. Anal. Recognit, 2013, pp. 329-333

[4] Nunez H, Ramos E, Automatic classification of academic documents using text mining techniques, 2012, pp. 7

[5] Dalal M.K, Zaveri M.A, Automatic Text Classification of sports blog data, ComComAp. Int. Conf. Los Alamitos, CA, USA, 2012 pp. 219-222

[6] García Adeva J.J, Pikatza Atxa J.M, Ubeda Carrillo M, Ansuategi Zengotitabengoa E, Automatic text classification to support systematic reviews in medicine, 41 (2014) 1498-1508.