Individualized Automatic Classification of Web Documents

Yihjia Tsai, Kaun-Yu Chen

Department of Computer Science and Information Engineering, Tamkang University, Taipei

Email:eplusplus@gmail.com, 795410249@mail.tku.edu.tw

Abstract: This paper applies Naïve Bayes classifier in designing customized automatic web document classification to systematically collecting massive news articles from the Internet. The proposed news classification system allows users to establish the necessary information classifications based on their own preferences. When the amount of daily news is increasing, this approach enables users to effectively filter through large amount of articles and more focused on interested articles. Performances of the proposed approach are characterized by the recall rate and precision. This system can achieve over 66% recall rate, and over 89% precision rate for a real-world Chinese test database.

Keywords: Naïve Bayes; web documents classification

I. Introduction

In the age of networking, news from the Internet has become an important source of information. Most major websites use their own classification method to display the news, but that usually is insufficient or produces too many classifications. When a user would like to collect different view points from different news websites and form a more diversified picture of the same event, a lot of clicking and manual works have to be done in order to form such view point.

For this reason, this research investigates a classification method based on the contents of news articles, and implements an automatic classifier according to personal preferences. After automatically collecting news articles from pre-specified news web sites, the system is able to analyze the contents of news according to user defined categories and display updated news articles.

At the heart of this classification scheme is a method based on the Naïve Bayes, one major motivation for adopting this method is that it is fast obtain prediction results and the accuracy is high in classify Chinese news articles. In this paper, we use news articles from the Chinese news database provided by Symposium of Search Engine and Web Mining (SEWM) 2006 in studying the performance of our proposed scheme.

II. Theoretical Background

Document classification refers to grouping document files in one or more previously defined categories based on the nature of the document. The classification model establishes the classification rules and sorts out other information through training on known history data mainly to filter spam e-mails and engage in enterprise knowledge document management. Commonly used document classification methods includes: the Naïve Bayes, decision tree, K-nearest neighbor, etc.

The Naïve Bayes classifier is based on the Bayes's Theorem in which known samples class distributions are used to predict the unknown samples. According to the Bayes's Theorem, which states that when X represents an unknown type of news and C represents a category, then the probability of X being under the category C is as shown in equation 2-1.

$$P(C \mid X) = \frac{P(C)P(X \mid C)}{P(X)} \tag{2-1}$$

In equation 2-1, we know that the right hand side consists of P(X|C) and prior probability P(C), which can both be estimated empirically from training data. When each document contains more than one features used in the classification task, a simple model is used to simplify the calculation of joint probability distribution. This simplified model is called the Naïve Bayes classifier. It has added one assumption, conditional independent among those features. Suppose that X includes the k attribute values and the attributes are independent of each other as shown in equation 2-2, where $x_1 \ldots x_k$ are the k attribute values of news X.

$$P(C,x_1,x_2,\cdots;x_k)=P(C)P(x_1 \mid C)P(x_2 \mid C)\cdots P(x_k \mid C) \tag{2-2}$$

Therefore according to the Naïve Bayes model, the conditional distribution over category C is expressed as in equation 2-3, where α is a scaling factor dependent only on $x_1 \ldots x_k$. This equation indicates that the computation cost is relatively low in classification task [1].

$$P(C|x_1,x_2,\cdots,x_k)=\frac{1}{\alpha}P(C)\prod_{i=1}^{k}P(x_i|C) \qquad (2\text{-}3)$$

In addition to the classification engine, there is one more issue that needs to be addressed in the proposed system, that is, Chinese words segmentation scheme. This is due to the fact that the basic unit in Chinese sentence is called "word" which is the smallest unit in lexical analysis instead of the "character" [2]. Unlike English sentence, where a white space is placed between two consecutive words, in a Chinese sentence, there is no such clear distinction between words. Thus, Chinese natural language processing systems, such as document retrieval, speech recognition, all require a word segmentation scheme before proceeding to the next step. The accuracy and completeness of the word segmentation outcome therefore is a crucial factor in the follow-up handling, making word segmentation of great importance. At present, the main Chinese word segmentation methods consist of three types: the heuristic rule-based, statistical, and hybrid schemes [3,4,5,6]. In this research we adopt the heuristic rule-based word segmentation scheme. This decision is based on its ease of implementation and its result accuracy.

III. Implementation

The proposed system works in two distinctive phases. The first is the training phase, where we use sample news articles to train the classifier. The second phase is then the application of the trained classifier to classify news articles. The training of classifier is started by importing training articles. After word segmentation and frequency counting, the final results are saved for the second phase.

The automatic classification module contains the first few steps of the training phase as stated above. At the end of the frequency counting, the Naïve Bayesian inference engine is invoked to calculate the likelihood of recurrent events based on the probability of events that have already taken place.

IV. Performance Analysis

In this paper, real-world test data are taken from the Chinese classification evaluation database used in the Symposium of Search Engine and Web Mining held in China in 2006. Its contents are from China's Xinhua Net and Sina Corp, including a total of eight news categories, 960 training web pages, and 240 trial web pages.

A higher recall rate signifies higher capability to classify and more accurately distinguish the articles' categories. The precision rate on the other hand is used to evaluate the correctness of the classification system and a higher precision rate in this case signifies higher accuracy.

The classification capabilities of the proposed system are characterized by two performance metrics, i.e. recall rates and precision rates. The recall rate for a category is defined as the number of correctly classified news articles divided by the total number of news articles that belongs to that category. It is used to evaluate the classification system's ability to classify. Precision for a category is defined as the number of correctly classified news articles divided by the total number of articles that was classified as that category.

Table 4-1 summarizes the performance of the proposed system. Since the precision rates exceeded 89%, but the recall rates were only 66.85% on average, there is a need to investigate further the differences between the precision rates and recall rates.

Taken for instance the "finance and economics" category which had the lowest recall rate, there were supposedly only 30 articles of the test data under each category, but the "finance and economics" category had 31. It is very likely that there was incorrect information in the Beijing University's research test data. Based on the classification results, a total of 11 news articles were classified under the "finance and economics" category and 10 of them were "finance and economics" news, thus the precision rate reached 90.91%. However, 21 of the news articles were incorrectly classified or they belonged

Table 4-1 Results of Classification

categories	articles	Automatic Classification	accuracy	Recall rate	Precision Rate
Education	30	27	27	90.00%	100%
Entertainment	29	24	22	75.86%	91.67%
Health	30	26	25	83.33%	96.15%
Tech	30	29	21	70.00%	72.41%
Military	30	13	12	40.00%	92.31%
Politics	30	16	14	46.67%	87.50%
Business	31	11	10	32.26%	90.91%
Sports	30	33	29	96.67%	87.88%
average				**66.85%**	**89.85%**

978-1-935068-15-0 © 2010 SciRes.

to several categories, thus the recall rate of 30.26% was the lowest. If we further analyze these 21 articles, these articles were supposed to fall under the "finance and economics" news category, but the system's analysis showed that 4 of these articles were incorrectly classified, 13 belonged to several categories, and 3 either could not be classified or belonged to the same category.

These findings showed that a news article may have similar possibilities to be in two or more categories, signifying that it may belong to several classifications.

V. Conclusion

This paper proposed a customized automatic news classification system based on Naïve Bayes model. Individuals first established their own categories, imported the training articles to strengthen the correctness of the classifier. Then, the test news through the classifier was redefined and classified to derive at a recall rate exceeding 66% and a precision rate exceeding 89% which is a realization of the automatic classification of customized news.

The experimental results show that by using the text information in the news as the characteristic information of the training classifier and executing the fast Naïve Bayes Classifier to construct the news classifier, the classification results have been quite remarkable.

Furthermore, the experimental results also show that a news article may belong to several news categories at the same time. So, as we emphasize customized classifications, we ought to conduct in-depth analysis again on the news articles under different categories that have obtained the same score to minimize the problem of news repetition. If we can add special terminology in the database, we may achieve more significant results in analyzing article classification.

Reference

[1] Lewis, D. D., "Naive (Bayes) at forty: The independence assumption in information retrieval," Proceedings of ECML '98., 1998.

[2] Liang, N.Y., "CDWS: An Automatic Word Segmentation System for Written Chinese Texts," Journal of Chinese Information Processing, Vol. 1, No.2, 1987.

[3] Chen, K.J. and S.H. Liu, "Word Identification for Mandarin Chinese Sentences," Proceeding of COLING-92, 14th Int. Conf. on Computational Linguistics, pp. 101-107, 1992.

[4] Li, G.C., K.Y. Liu and Y K. Zhang, "Identifying Chinese Word and Processing Different Meaning Structures," Journal of Chinese Information Processing, Vol. 2, pp. 45-53, 1988.

[5] Liang, N.Y. "Knowledge of Chinese Word Segmentation," Journal of Chinese Information Processing, Vol. 4, pp. 42-49, 1990.

[6] Fung, P. and D. Wu, "Statistical Augmentation of a Chinese Machine-Readable Dictionary," Proceedings of Second Annual Workshop on Very Large Corpora, pp. 33-56, 1994.