http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (20): 5492-5496, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5492.5496
© 2013 Asian Network for Scientific Information

Efficient Feature Selection Methods in Chinese Spam Filtering

[1,2]Xu Yan
School of Information Science, Beijing Language and Culture University, Beijing, 100083, P.R. China
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, P.R. China

Abstract: This study, from the perspective of Chinese Spam Filtering, focuses on efficient feature selection methods. It expounds the traditional feature selection algorithms including Document Frequency (DF), Information Gain (IG), the Mutual Information (MI), Chi-square (CHI) and Knowledge Gain (KG) which is proposed in my previous study. Testing these methods on exposing Chinese spam data set, the results show that in Chinese spam corpus CHI and KG can efficiently extract valid features for spam classifications.

Key words: Chinese spam filtering, text classification, feature selection

INTRODUCTION

Referring to anti-spam technologies which analyses the content of an e-mail message (such as the body text) to identify spam, this content filter technology provides a more accurate method of filtering messages which can automatically get the spam feature and capture the changes of them in real time.

Content-based spam filtering techniques include two categories: rules-based methods and statistics-based methods. Rules-based methods can artificially set the spam keywords (keyword filtering), meta-information (such as the sender, sender addresses, IP addresses and so on). At present, many anti-spam tools on the market are using these technologies (De Capitani *et al*., 2004), or using black/white list to deal with spam. However, arbitrary settings will cause significant loss of normal mail. From a content point of view, spam filtering classifies mail messages into spam and legitimate classes. So, all text classification methods can be used for spam filtering. (Manber, 1994).

Since spam filtering issues are seen as classification problems in real terms, the classification with the body of the message as a feature is a text classification problem. One of the main challenges for text categorization is in high-dimensional feature of feature space, particularly in Chinese message classification. These words characteristic can have more features to meet thousands of and even tens of thousands of species. Therefore, feature selection has been a very important step for text categorization.

Feature selection method is a mapping from features to the real numbers. In practice, we will use feature selection equation to calculate values for each tag on the training set and remove these whose mark is less than the threshold value. Existing mainstream feature selection algorithms include Document Frequency (DF), Information Gain (IG), the Mutual Information (MI), Chi-square (CHI) and so on. They are all came from the area of machine learning. It was able to successfully address the many practical problems that occur in. However, this algorithm tended to focus on improving the overall classification feature selection algorithm. I thought of using rough sets to depict the knowledge and propose Knowledge Gain (KG) feature selection algorithm, then bring the idea to spam filters which have also achieved good results in the experiments.

Therefore, to study the efficient feature extraction in Chinese spam filtering issues so as to achieve accurate message classification purposes, comparing the five feature selection method, this article tested in a Chinese spam corpus and supported by a number of experimental findings.

RELATED WORK

Typical study on text classification consists of two classes, one is about the plane, combing the KNN and LLSF classification tools, analyses and compares the DF, IG, CHI, MI, as well as other feature selection methods. IG in many tests is the best feature extraction algorithm (Xu, 2011) (Androutsopoulos *et al*., 2000). Citation (Yu *et al*., 2003) is focused on non-balanced data set text classification, using Bayesian classification tool to analyze and compare the IG, Expectation Cross-Entropy (ECE), Weight of Evidence Text (WET), Odds Ratio (OR) and other methods. Results showed that binary advantage rate was the best feature selection method, while the IG is relatively poor. This conclusion is contradictory with the Dr. Yang. There are many contradictions in the results of these experiments and contradictions have the most direct reason that they use

different corpora. In other words, when you use different sets of corpus data, a feature selection method may perform very well in some corpus, but on another corpus, the method might not be so good. This is because text classification has its own characteristics of the data set. Yang Yiming and other experts found DF, IG and CHI tend to select high-frequency characteristics (Xu, 2011; Microsoft Corporation, 2008) and effect in traditional method is better. The relatively poor performance of MI method was because its feature selection tends to be rare words. DF has better performance in Chinese text classification (Xu, 2011), Min Zhang, found not only in the text category, but also on the network information retrieval, DF can have very good performance (Androutsopoulos *et al*., 2004).

In Chinese spam filtering, unbalances and skewed data sets are very new and important issues (Shi *et al*., 2002; Cai and Shi, 2003). Unbalance, that is, there is a great disparity in the number of samples in each category, resulting in a lot of feature selection and classification satisfactory. Because of imbalance problems, classification systems can easily be overwhelmed and ignored many categories. Most machine learning algorithms are based on training examples on balance data sets, so these algorithms on an unbalanced corpus often have poor performance in general categories with high correlation, but perform very well in rare categories with low correlation (categories with few samples) (Androutsopoulos *et al*., 2000). Especially when using binary classification strategy, the positive sample only has a small portion in all cases and then the classification on positive cases affects the results of negative cases (Shi *et al*., 2002; Cai and Shi, 2003). Unbalanced data sets are a common problem in many areas: text classification, anomaly detection, risk management, medical diagnostics and so on. Yan Xu proposed a solution to the unbalanced problem-constructs the form DFICF feature selection method which get significant improvement on the classification of rare class effect (Sun *et al*., 2006).

METHODS OF FEATURE SLESTION

Document frequency, DF: Document frequency DF (t) of feature t is the number of documents in which a term t occurs. At the time of feature selection, select features whose DF values are above a certain threshold and remove those whose DF values are below this threshold. The basic assumption is that rare terms are either non-informative for category predictions, or not influential in global performance. In either case removal of rare terms reduces the dimensionality of the feature space. Improvement in categorization accuracy is also possible if rare terms happen to be noise terms.

DF's biggest advantage is the speed and the linear time complexity with the document scale which is ideal for large scale document sets feature selection. It often used as a secondary method of feature selection to filter out some low-frequency characteristics.

Mutual information, MI: MI, used to represent the correlation between two variables, is mentioned in information theory. It is introduced into text classification to represent the relevance of feature and category. For a feature t and a class ci, we can construct an confusion matrix, as shown in Table 1. Where A presents the number of documents which belongs to cj and contains t, B presents the number of documents which doesn't belong to cj but contains t , C presents the number of documents which belongs to cj but doesn't contain t , D presents the number of documents which neither belongs to cj nor contains t. N = A+B+C+D, N presents the total number of documents concluded in the training set.

The MI between t and cj can be defined as:

$$MI(t,c_j) = \log\frac{P(t \wedge c_j)}{P(t)P(c_j)} = \log\frac{P(t \mid c_j)}{P(t)} = \log\frac{A \times N}{(A+C)(A+B)} \quad (1)$$

When t separates from cj, MI (t, cj) = 0. When the feature rarely appears in this category, their MI is negative which is called Negative correlation. Under the same conditional probability, rare characteristics will get higher MI, thus makes evaluation function prefer to the rare characteristics. MI isn't suitable for evaluating the features which own larger distribution differences in category.

Information gain, IG: IG is the most common ways to choose the greatest decision point in the decision tree. The theory of information gain bases on the conception of entropy which is defined in information theory. The information gain evaluates the average information which is used to judge the category of files, when you know whether the features are in files or not.

Information gain measures the amount of information obtained for category prediction by knowing the presence or absence of a term in a document. Let m $\{c_i\}_{i=1}^m$ denote the set of categories in the target space. The information gain of term t is defined to be:

$$\begin{aligned} G(t) &= -\sum_{i=1}^{m} p(c_i)\log p(c_i) \\ &+ p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) \\ &+ p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t}) \end{aligned} \quad (3)$$

Given a training corpus, for each unique term the information gain is computed and those terms whose information gain is less than some predetermined threshold are removed from the feature space.

statistic χ^2, CHI: The χ^2 statistic measures the lack of independence between t and c and can be compared to the χ^2 distribution with one degree of freedom to judge extremeness. Using the two-way contingency table of a term t and a category c, where A is the number of times t and c co-occur, B is the number of time the t occurs and N is the total number of documents, the term-goodness measure is defined to be:

$$\chi^2(t,c) = \frac{N*(AD-CB)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \quad (4)$$

A, B, C, D represent quantity of document, showing in the above table.

We computed for each category the χ^2 statistic between each unique term in a training corpus and that category and then combined the category-specific scores of each term into two scores:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P_r(c_i)\chi^2(t,c_i) \quad (5)$$

$$\chi^2_{max}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\} \quad (6)$$

The computation of CHI scores has a quadratic complexity, similar to IG and χ^2 values are comparable across terms for the same category.

Knowledge gain, KG: Based on the view that knowledge is a classification issue, knowledge gain can be thought as a prediction from a classification based on one attribute to a classification based on another attribute which means how helpful and contributive that knowing a classification based on one attribute is to classification based on another attribute.

In General, if the equivalence classes in P and D are the same, then they are of maximum knowledge gain, if the probability of P and D are independent, their mutual knowledge gain is minimal. U is the universe, P is a property collection, D is another collection of properties. When given P, its contribution to classification of D, called P-D knowledge gain, written KG(D|P), defined as:

$$KG(D|P) = W_D - W_{D/P} \quad (7)$$

EXPERIMENT VALIDATION

Experiment corpus: This experiment uses Chinese spam corpus published by China education and research computer network emergency response team, containing 25,088 spam emails and 9,272 normal mails. The database collection first use honeypot spam, such as each spam is attracted to an e-mail account called xxx@ccert.edu.CN. Normal messages come from the Chinese public forum. This article uses the text portions of the data set. Data set download address: http://www.ccert.edu.CN/spam/SA/datasets.htm#2

Experiment design: Spam filtering problem is regarded as a second class of text classification problem, it can be abstracted for the flow chart in Figure 1 which mainly related to text pre-processing, the text representation, feature selection, as well as several aspects, such as training of a classifier.

According to the various statistics of message training document sets, we can represent various features of the each document, then use specified classification algorithm to train out classifier (this article experimental used Naïve Bayes and SVM classifier), for new message documents to be classified, after converting into features represented by vector, use the classifier to classify them to get their category.

Experimental results and analysis: We selected 20-200 features for each method, a total of 20 groups and experimented on Naïve Bayes and SVM platforms.

By comparison of precision, DF performance is the most prominent. Its highest value has reached 0.93. Results from the recall rates showed that the performance of KG is significantly better than the other two methods. Additionally, KG has a more powerful effect with the

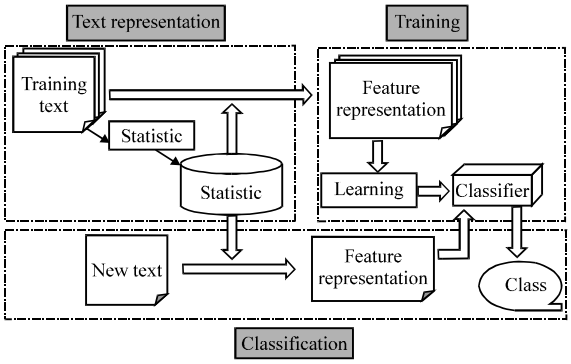

Fig. 1: Flowchart of the text classification in spam filter

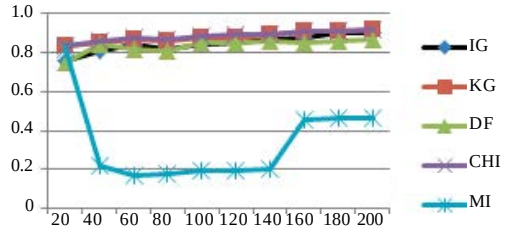

Fig. 2: F values under different feature dimensions

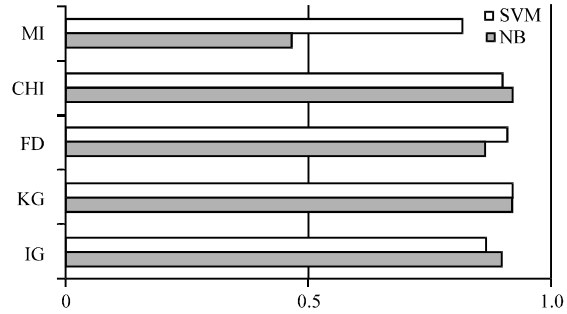

Fig. 3: F values under different classifiers

highest value of 0.92. Later, we will focus on the evaluation using P and R composite indicator. As we can see, while KG methods cannot be clearly superior to other methods on accuracy rate, in the other two indicators it is better than the other two methods.

By comparing the accuracies, KG algorithm has more advantages. Whether in the high-dimensional or a low-dimensional space, it had demonstrated its superiority, while the other two methods have not a big difference. Overall, the effects of Chi-square and KG are better. They can handle this data distribution well.

For more intuitive analysis of the relationship of experimental results, feature dimensions and classifiers, the following comes from the one-by-one introduction from the perspective of F value evaluation indicator.

Analysis on feature dimension effect on experimental results: Figure 2 shows the feature dimensions change (from 20 to 200), using different feature selection method for classification of F value changes. KG and CHI both have a high F value in high and low dimension, the combined effect is superior to others.

Analysis on classifier effect on experimental results: Overall, SVM classifier results superior to Naïve Bayes classifiers a little. Figure 3 shows the experiment results of two different classifiers using different feature selection methods.

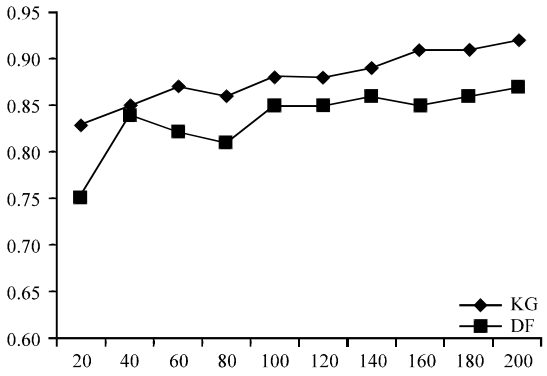

Fig. 4: F value of KG compared to the F values of DF

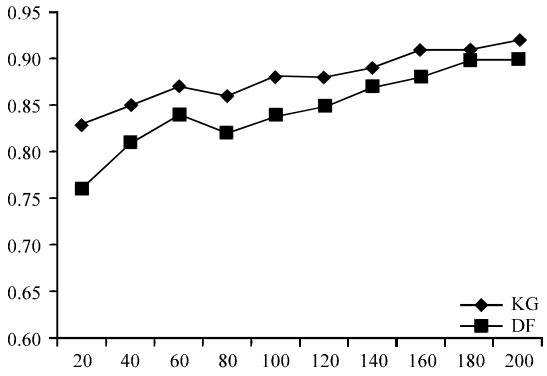

Fig. 5: F value of KG compared to the F values of IG

Comparison of KG and other methods (DF and IG): Lingdai Deng found that DF works best in a Chinese text classification. Yiming Yang's studys had concluded that the IG in classification task is the best method. While Knowledge Gain (KG) algorithm also performed good results in the experiments, conclusions are as follows: KG has more advantages than DF in II unbalanced classification. To describe the advantage of KG, we specifically set a couple of experiments on KG and IG for comparison. From Fig. 4 and 5 we can see, regardless of in high or low dimension space, KG is obviously better than DF and IG algorithm.

From the view of emergency Center of Chinese corpus experimental results, KG method show great superiority. This may be because KG method uses the theory of rough set to depict the benefits of knowledge gain.

CONCLUSIONS AND FUTURE WORK

This article was based on text classification technology to solve the problem of spam filtering on

Chinese spam data sets. We use classic methods of feature selection including DF, Chi-square, IG, as well as our KG method to carry out experiments, in order to explore the high-efficiency feature selection methods of Chinese spam filtering. Experimental results show that our knowledge Gain (KG), DF and chi-square feature selection methods can efficiently extract features, filtering spam in Chinese mails.

A problem currently exists in this area is that the spam is seen as a static set of data, but spam, in practice, exists as a kind of stream. Our next work is mainly focused on how to dig out the related features between spam and stream. Another worthwhile work is: there have been more and more new features of spam, such as adding some deformation characteristics, introducing pictures or falsifying returned letters. How to explore a wide variety of new features is also an important direction of spam filtering.

ACKNOWLEDGMENTS

This study is supported by National Key Technology Research and Development Program of China (2012BAH39B02), National Natural Science Foundation of Beijing. No: 4122076 and Fundamental Research Funds for the Central Universities(13YCX176). We would like to thank the reviewers for providing valuable comments and advices.

REFERENCES

Androutsopoulos, I., G. Paliouras and E. Michelakis, 2004. Learning to filter unsolicited commercial E-mail. Technical Report, No. 2004/2, NCSR Democritus, 2004. http://nlp.cs.aueb.gr/pubs/TR2004_updated.pdf

Androutsopoulos, I., J. Koutsias, K.V. Chandrinos and C.D. Spyropoulos, 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. July 24-28, 2000, ACM, Athens, Greece, pp: 160-167.

Androutsopoulos, I., J. Koutsias, K.V. Chandrinos, G. Paliouras and C.D. Spyropoulos, 2000. An evaluation of naive bayesian Anti-spam filtering. Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, May 2000, Barcelona, Spain, pp: 9-17.

Cai, L.J. and R.S. Shi, 2003. A High-performance intelligent content filtering model. Comput. Eng., 29: 146-148.

De Capitani, D., E. Damiani, S. De, C. Vimercati, S. Paraboschi and P. Samarati, 2004. An open digest-based technique for spam detection. Proceedings of the 2004 International Workshop on Security in Parallel and Distributed Systems, (PDCS`04), USA., pp: 15-17.

Manber, U., 1994. Finding similar files in a large file system. Proceedings of the Technical Conference on USENIX Winter, January 17-21, 1994, San Francisco, California, pp: 1-10.

Microsoft Corporation, 2008. Microsoft security intelligence report. Volume 6, July-December 2008. http://www.microsoft.com/en-us/download/details.aspx?id=22513.

Shi, X.J., R. Lin and Z.P. Chen, 2002. Bayesian spam filtering algorithm based on minimum risk. Comput. Sci., 29: 50-51.

Sun, C.M., X. Yan and B.Y. Lin, 2006. Improvement of a method based on term frequency inverse document frequency. China Electr. Power Educ. Suppl. (Second Series).

Xu, Y., 2011. Rough set and its application in Chinese spam filtering. Proceedings of the IEEE International Conference on Granular Computing, November 8-10, 2011, Kaohsiung, pp: 750-755.

Yu, H., Z. Li, H. Tang and Z. Wu, 2003. A rough set approach for analyzing E-mail filtering system. Comput. Eng. Appl., 15: 47-48, 67.