Comparing Feature Selection Methods in Spam Mail Filtering

Jong-Wan Kim*, Sin-Jae Kang*

*School of Computer and Information Technology, Daegu University
Gyeonsan. Gyeongbuk. 712-714 South Korea
{jwkim, sjkang}@daegu.ac.kr

Abstract - In this work, we compared several feature selection methods in the field of spam mail filtering. The proposed fuzzy inference method outperforms information gain and chi squared test methods as a feature selection method in terms of error rate. In the case of junk mails, since the mail body has little text information, it provides insufficient hints to distinguish spam mails from legitimate ones. To address this problem, we follow hyperlinks contained in the email body, fetch contents of a remote web page, and extract hints from both original email body and fetched web pages. A two-phase approach is applied to filter spam mails in which definite hint is used first, and then less definite textual information is used. In our experiment, the proposed two-phase method achieved an improvement of recall by 32.4% on the average over the 1st phase or the 2nd phase only works.

Keywords: artificial intelligence, feature selction, fuzzy inference, spam-mail filtering.

1 Introduction

With the popularization of the Internet, low cost, and fast delivery of message, email has become an indispensable method for people to communicate each other. Though email brought us such huge convenience, it also caused us trouble of managing the large quantities of spam mails received everyday. Spam mails, which are unsolicited commercial emails or junk mails, flood mailboxes, exposing young people to unsuitable content, and wasting network bandwidth [1]. Most software for email clients provides some automatic spam mail filtering mechanism, typically in the form of blacklists or keyword-based filters. Unfortunately constructing these lists and filters is manual time-consuming process, and is not perfect for a variety of cases in real situation.

The spam filtering problem can be seen as a particular case of the text categorization problem. Several information retrieval (IR) techniques are well suited for addressing this problem, and in addition it is a two-class problem: spam or non-spam. A variety of machine learning algorithms have been used for email categorization task on different metadata [2, 4, 5, 6]. Sahami et al. [2] focuses on the more specific problem of filtering spam mails using a Naïve Bayesian classifier and incorporating domain knowledge using manually constructed domain-specific attributes such as phrasal features and various non-textual features. In most cases, support vector machines (SVM), developed by Vapnik [3], outperforms conventional classifiers and therefore has been used for automatic filtering of spam mails as well as for classifying email text [4, 5]. Yang et al. [6] demonstrate that Naïve Bayesian and SVM classifier is by far superior to TFIDF. In particular, the best result was obtained when SVM was applied to the header with feature subset selection. Accordingly, we can conclude that SVM classifier is slightly better in distinguishing the two-class problem.

For selection of important features or terms representing documents such as mails or news well, assigning them weights are the same problem that the existing linear classifiers such as Rocchio and Widrow-Hoff algorithms [7] find centroid vectors of an example document collection. Both of these algorithms use TF (Term Frequency) and IDF (Inverse Document Frequency) for re-weighting terms but they do not consider term co-occurrence relationship within feedbacked documents. To resolve this problem, the computation of term co-occurrences between these representative keywords and candidate terms within each example document is required. Three factors of TF, DF (Document Frequency), and IDF have essentially ambiguous characteristics, which are used to calculate the importance of a specific term. Since fuzzy logic is more adequate to handle intuitive and uncertain knowledge, we combine the three factors by the use of

fuzzy inference. We calculate weights of candidate terms by using the method [8] that it is known to give superior performance to the existing representative keyword extraction methods and assign a priority to select representative keywords with the weights of candidate terms.

In this paper, we present the feature selection by fuzzy inference is a little superior to the conventional methods such as information gain and chi square in pornography or porn mails filtering. A two-phase filtering system for intercepting spam mails based on textual information and hyperlinks is also given. Our system relies on two basic ideas. First, to select features with high discriminating power, we compared the fuzzy inference method with information gain and chi square because information gain and chi squared test were known effective in text categorization [9]. Second, a spam mail is classified by using two-phase system. In the first phase, definite information such as sender's URL, email addresses, and spam keyword lists is applied. In the second phase, remaining, that is, unclassified emails are classified using less definite information, extracted not only from email header and body but also by fetching web pages.

2 Feature Selection in Training Phase

Since the body of a spam mail has little text information recently, it provides insufficient hints to distinguish spam mails from legitimate mails. To resolve this problem, we utilized hyperlinks contained in the email body and extracted all possible hints from original email body and the fetched webpage. These hints are used to construct SVM classifiers. We divided hints into two kinds of information: definite information and less definite textual information. Definite information for filtering spam mails is sender's information, such as email id and URL addresses, and definite spam keyword lists such as "porno," "big money" and "advertisement". There are many particular features of email, which provide evidence as to whether an email is spam or not. For example, the individual words in the text of an email, domain type of the sender, receiving time of an email, or the percentage of non-alphanumeric characters in the subject of an email are indicative of less definite textual information in a spam mail [10]. In the two-phase approach, we first classified the spam mail by using the definite information, and then used the less definite information.

Feature selection from less definite textual information involves searching through all possible combination of features in the candidate feature set to find which subset of features works best for prediction. A few of the mechanisms designed to find the optimum number of features are document frequency threshold, information gain, mutual information, term strength, and chi square. In comparing learning algorithms, Yang and Pedersen found that, except for mutual information, all these feature selection methods had similar performance and similar characteristics [9]. To select features having high discriminating power, we compared the fuzzy inference method [8] with information gain and chi square because information gain and chi square were known effective in text categorization. Information gain is frequently employed as a term goodness criterion in the field of machine learning. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. The chi square measures the lack of independence between a term t and a category c and can be compared to the chi squared distribution with one degree of freedom to judge extremeness. The chi square statistic has a natural value of zero if t and c are independent.

In the fuzzy inference, TF, DF, and IDF of each term are calculated from the preprocessed email documents and are normalized. The normalized term frequency NTF, the normalized document frequency NDF, and the normalized inverse document frequency NIDF are used as fuzzy input variables. Membership functions of the fuzzy input and output variables should have been fuzzified to the form suitable for fuzzy inference. First, we will define the membership functions (μ) of three fuzzy input variables NTF, NDF, and NIDF as the following expressions in which the meaning of each fuzzy term {S, M, L} is corresponding to Small, Middle, and Large, respectively:

$\mu_S(x) = \max(0, 1- x/0.75)$ and $\mu_L(x) = \max(0, 1+(x-1)/0.75)$ for x = NTF variable,

$\mu_S(y) = \max(0, 1- y/0.35)$,

$\mu_M(y) = \min\{\max(0, 1+(y-0.5)/0.35), \max(0, 1-(y-0.5)/0.35)\}$, and

$\mu_L(y) = \max(0, 1+(y-1)/0.35)$ for y = NDF variable or NIDF variable.

Similarly the fuzzy output variable TW (Term Weight) that represents the importance of each term has the following membership functions. At this time, the meaning of each fuzzy term {Z, S, M, L, X, XX} is corresponding to Zero, Small, Middle, Large, Xlarge, XXlarge, respectively.

$\mu_Z(TW) = \max(0, 1-TW/0.2)$, $\mu_{XX}(TW) = \max(0, 1+(TW-1)/0.2)$,

$\mu_S(TW) = \min\{\max(0, 1+(TW-0.2)/0.2), \max(0, 1-(TW-0.2)/0.2)\}$,

$\mu_M(TW) = \min\{\max(0, 1+(TW-0.4)/0.2), \max(0, 1-(TW-0.4)/0.2)\}$,

$\mu_L(TW) = \min\{\max(0, 1+(TW-0.6)/0.2), \max(0, 1-(TW-0.6)/0.2)\}$, and

$\mu_X(TW) = \min\{\max(0, 1+(TW-0.8)/0.2), \max(0, 1-(TW-0.8)/0.2)\}$.

Table 1 gives 18 fuzzy rules to inference the term weight TW, where NTF is considered as primary factor, NDF and NIDF as secondary ones. See in [11] to refer explanation in detail. Finally, the terms with higher TW

values are selected as feature vectors to classify mail messages by fuzzy inference.

Table 1. Fuzzy inference rules are composed of 2 groups according to NTF value

	NDF \ NIDF	S	M	L
NTF = S	S	Z	Z	S
	M	Z	M	L
	L	S	L	X
	NDF \ NIDF	S	M	L
NTF = L	S	Z	S	M
	M	S	L	X
	L	M	X	XX

3 Experiments

The email corpus used in the experimental evaluation contained a total of 4,792 emails and 4 categories: 2,218 for legitimate mail, 1,100 for porn spam, 1,077 for financing spam, and 397 for shopping spam. To select important features, we used the *weka.attributeSelection* package provided by WEKA [12]. WEKA is a workbench designed to aid in the application of machine learning techniques to real world data sets. WEKA contains a number of classification models. The SVM classifier used in this experiment was also provided by WEKA. SVM is tested with its default parameters settings within the WEKA.

In email filtering, it is extremely important that legitimate emails are not filtered out. In comparison, a user may be satisfied if some spam-email was not filtered, in order not to miss any good email. Error rate represents the ratio of the incorrect predictions over total mails [13, 14]. Thus, good email filtering should indicate low error rate. Error rate is defined as:

$$\text{Error rate} = \frac{number\ of\ incorrect\ predictons}{total\ number\ of\ emails} \quad (1)$$

We used ten-fold cross validation to reduce random variation in the experiments. E-mail corpus was randomly partitioned into ten parts, and each experiment was repeated ten times, each time reserving a different part for testing, and using the remaining nine parts for training. Results were then averaged over the ten runs. Figure 1 compared the performance of the fuzzy inference and the conventional ones such as information gain and chi square in selecting features for filtering pornography spam. Almost 7,600 morphemes were extracted by eliminating stop words and redundant words. These morphemes are the candidate features to be used training porn mails. In this work, we selected 200, 338, 485, 681, and 838 features for information gain and chi square by the WEKA and for the fuzzy inference by our system among these 7,600 morphemes. When compared with the experimental results by Yang [9], it gave almost same results. As you can see in Figure 1, the fuzzy inference method improved about 6% and 10% over information gain and chi squared test in terms of the average error rate, respectively. It is more important to reduce the average error rate than to increase other performance measures such as accuracy and F-measure. Therefore, the proposed fuzzy inference is regarded as a good and stable feature selection method regardless of the number of selected features.

To evaluate the filtering performance on the email document corpus, we use the recall (R) and precision (P) commonly employed in the information retrieval field. The 4,335 emails among 4,792 ones are used for training SVM classifiers and the remaining 457 are used for testing the proposed system's performance. Testing emails are used to determine whether the mails are spam or not using the information and classifiers constructed during the training phase. We already divided hints into two kinds of information: definite information and less definite textual information. In case that an email contains one of the definite information, there is no need to perform machine learning algorithms, since it has a very high probability of being spam mails. In other case that the email has no definite information, it is evaluated using the SVM classifiers. That is to say, if an email contains one of the definite information, it is regarded as a spam mail. Otherwise, it is passed to the next SVM applying phase. SVM classifier for porn spam mails is applied first. If an email is classified as a spam mail, the second applying phase is over. If not, it is passed to the next SVM classifier for financing spam. When the email is classified as a financing spam mail, the second applying phase is over too. Like the above two SVM classifiers, the last SVM classifier for shopping spam is performed in sequence if needed.

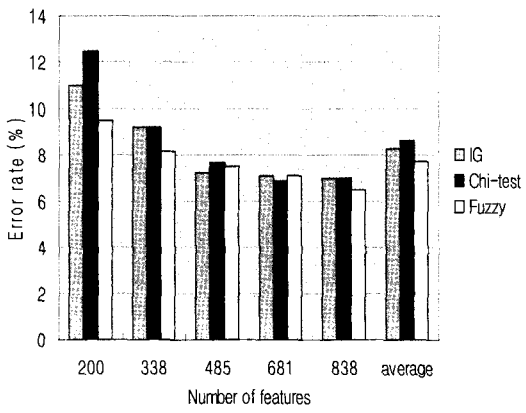

Fig. 1. Error rates of three feature selection methods

We found from Table 2 that the proposed two-phase method was more effective than the method applying each phase separately, since the 1st phase undertook some portion of the 2nd phase's workload with very high precision. Compared the 1st phase or the 2nd phase only works, there is little change of precision, but recall was improved a lot when using hyperlinks. That is, the two-phase method improved the performance of recall by 68.6% and 9.2% over the 1st phase or the 2nd phase only works, respectively. When we consider the average of the 1st phase only and the 2nd phase only work, the proposed two-phase method improved 32.4% in terms of recall. We can recognize from these results that fetching web pages plays an important role in collecting more features and then deciding ambiguous mails.

Table 2. Performance of the proposed method (%)

Applying phase	Recall	Precision
1st phase only	43.6	100
2nd phase only	67.3	99.4
1st + 2nd phase	73.5	99.5

4 Conclusion

In this paper, we performed a comparative experiment on feature selection in pornography mails categorization. As you can see in Figure 1, the proposed fuzzy inference method lowered the average error rate over information gain and chi-squared test by the 6% and 10%, respectively. In general, it is very important to reduce the average error rate than any other measures in spam filtering domain. Though the 6 or 10% improvement of error rate by the proposed fuzzy inference looks not significant, improving the error rate by only 1% is not easy and meaningful to email users. In also, we proposed a two-phase method for filtering spam mails based on textual information and hyperlinks. The proposed two-phase method achieved an improvement of recall by 32.4% on the average over the method of the 1st phase or the 2nd phase only work. We discovered that fetching hyperlinks is very useful in filtering spam mails, and the two-phase method is more effective than the method using machine learning algorithm only, blacklists, or keyword-based filters. This research is very important in that our system can prevent young people from accessing pornography materials on spam mails by chance, and save valuable time by lightening the email checking work. We will do further research on how to find more features by considering images in email messages and constructing ontology on spam keywords. Therefore we will improve the filtering performance.

References

1. Cranor, L. F. and LaMacchia, B. A., "Spam!," Communications of ACM, Vol.41, No.8 (1998) 74-83
2. Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E., "A bayesian approach to filtering junk e-mail," In AAAI-98 Workshop on Learning for Text Categorization (1998) 55-62
3. Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York (1995)
4. Drucker, H., Wu, D. and Vapnik, V., "Support Vector Machines for Spam Categorization," IEEE Trans. on Neural Networks, Vol.10(5) (1999) 1048-1054
5. Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," ECML, Claire Nédellec and Céline Rouveirol (ed.) (1998)
6. Yang, J., Chalasani, V., and Park, S., "Intelligent email categorization based on textual information and metadata," IEICE Transactions on Information and System, Vol.E86-D, No.7 (2003) 1280-1288
7. Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R., "Training algorithms for linear text classifier," Proc. of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (1996) 298-306
8. Kim, B. M., Li, Q., and Kim, J. W., "Extraction of User Preferences from a Few Positive Documents," Proceedings of The Sixth International Workshop on Information Retrieval with Asian Languages (2003) 124-131
9. Yang, Y, and Pedersen, J. P., "A comparative study on feature selection in text categorization," in Fourteenth International Conference on Machine Learning (1997) 412-420
10. Wolfe, P., Scott, C., and Erwin, M. W., Anti-spam toolkit, McGraw-Hill, 2004.
11. Kim, J. W., Kim, H. J., Kang, S. J., and Kim, B. M., "Determination of Usenet News Groups by Fuzzy Inference and Kohonen Network," Lecture Notes in Artificial Intelligence, Vol.3157, Springer-Verlag (2004) 654-663
12. Witten, I. H. and Frank, E., Data Mining: Practical machine learning tools and Techniques with java implementations, Morgan Kaufmann (2000)
13. Androutsopoulos, I., Koutsias, J., Chandrinos, V., Spyropoulos, D., "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages", Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (2000) 160-167
14. Resnick, P. J., Hansen, D. L., and Richardson, C. R., "Calculating Error Rates for Filtering Software," *Communications of ACM*, Vol.47, No.9, pp. 67-71, 2004.