An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization

Jyh-Jian Sheu
Department of Information Management, National Dong Hwa University
1, Sec 2, Dahsueh Road, Shoufeng, Hualien, 97401, Taiwan, R.O.C. (Email: jjsheu@mail.ndhu.edu.tw)
(Received May 9, 2008; revised and accepted July 4, 2008)

Abstract

The e-mail's header session usually contains important attributes such as e-mail title, sender's name, sender's e-mail address, sending date, which are helpful to classification of e-mails. In this paper, we apply decision tree data mining technique to header's basic attributes to analyze the association rules of spam e-mails and propose an efficient spam filtering method to accurately identify spam and legitimate e-mails. According to the experiment of applying numerous Chinese e-mails to our spam filtering method, we obtain the following excellent datums: the Accuracy is 96.5%, the Precision is 96.67%, and the Recall is 96.3%. Thus, the method proposed in this paper can efficiently identify the spam e-mails by checking only the header sessions, which can reduce the cost for calculation.

Keywords: Data mining, decision tree, security, spam filtering

1 Introduction

In recent years, with the increase of computer use and prevalence of Internet, the content of e-mails increases from characters to pictures, forms, video messages and varied files. Convenience, immediateness, low cost and extensive transmission scope allow e-mail to replace traditional mails and become one of the critical tools for exchanging messages among firms or organizations and people. E-mail results in revolutionary change with respect to communication and business operation. As internet advances, e-mail has turned into the principal tool of mass marketing. Many firm or website managers use e-mail as marketing tool because of its convenience and speediness. For instance, e-mail marketing is used to replace mail-order catalogues or magazines. Moreover, it can even replace part of customer services and increase customer satisfaction.

However, excessive application of e-mail has brought the problem of spams. According to the latest report of spam reorganization proposed by Symantec, the international firm of information security [22], by the end of May, 2007, spam has been 75% of e-mails in the world. Mail boxes are filled with plenty of spam which has become the annoying problem of many people. On the other hand, from the views of firms and governments, plenty of spams resulted in the serious burden of e-mail server. Thus, spam filtering method becomes an important research subject of Internet application [3, 5].

Currently, many spam filtering methods [1, 2, 4, 6, 7, 8, 9, 10, 12, 13, 15, 18, 19, 20, 22] have been proposed. For instance, comparing Blackhole and Whitehole Lists, using IP address detection, judging the abnormality of number of connections, filtering by key words, automatic semantic analysis and varied experience rules. However, these methods tend to include some restrictions. For example, the methods of filtering by keywords and automatic semantic analysis require the long-term operation for checking the content of e-mail which will consume plenty of time and costs; The methods to judge abnormality of the number of connections and IP address detection require the cooperation of ISP firms. How to more efficiently and accurately filtrate spam becomes the major motive of this research.

The purpose of this research is to construct an efficient spam filtering method. Different from many other methods at present which have to investigate the complete content of e-mails, this research doesn't check the content of e-mails to avoid the complexity and executive efficiency. We will concentrate our attention on e-mail's header session instead of scanning e-mail's content.

In this paper, we classify e-mails into several categorizations as follows: (1) sexual, (2) finance and job-hunting, (3) marketing and advertising, and (4) total. Then we analyze and record the basic attributes of e-mail's header session such as e-mail title, sender's name, sender's e-mail address, sending date. Subsequently, we apply decision tree data mining technique for each categorization to find the association rules of spam e-mails in this categorization. Based on these association rules, we can build a systematic method to accurately identify spam and legitimate e-mails.

Our method consists of the following two phases: (1) Rule training and (2) Classification of unknown e-mails. The purpose of training phase is to seek for association

rules between training e-mail's attributes in header session and e-mail's type (spam or legitimate). For each categorization of training e-mails, we apply the decision tree data mining algorithm to find the association rules between training e-mail's attributes in header session and e-mail type. These rules will be useful for classifying unknown e-mails in next phase.

The second phase is to judge the unknown e-mail to be either a legitimate mail or spam. Moreover, we establish a reversing mechanism such that the misjudgment (identifying a spam as a legitimate e-mail) could be reversed to elevate the overall accuracy.

The remainder of this paper is organized as follows. In Section 2, we discuss the related background of this paper. And we describe our new spam filtering method based on decision tree data mining technique in Section 3. The experimental results of our method are shown in Section 4. In Section 5, we conclude this paper.

2 Related Background

2.1 Various Spam Filtering Methods in Present Literatures

The filtering of spam is an important subject in the research area of Internet applications. Various methods or algorithms with respect to classification or filtering of spam had been proposed in present research literatures. For example, Cohen [2] proposed a set of RIPPER learning algorithm to classify the mails in 1996; Sahami et al. [19][19] screened the spams by applying Bayes classifier in 1998; Drucker et al. [6] classified spam and non-spam by Support Vector Machines (SVMs) in 1999; Carreras and Marquez [1] proposed a new method by applying AdaBoost algorithm in 2001; Manco et al. [13] introduced a filtering method based on clustering algorithms of data mining in 2002; In 2007, He and Bo [10] developed a new asymmetric boosting method; In 2006, Delany and Cunningham [4] proposed a kind of KNN (the k nearest neighbors) approach; And in 2008, Hsiao and Chang proposed [9] an incremental clustering-based classification method.

However, most of these filtering methods mentioned above had to consume a lot of time to scan fully the content of e-mail. In this paper, we expect to apply data mining technique to dig out the association rules of characteristics of spams. We check only the basic attributes of e-mail's header session such as e-mail titles, sender's name, sender's e-mail address, receiver's name, sending date, and receiving date. Then, these attributes are treated as input data for further analysis of decision tree data mining. We apply decision tree data mining algorithm to analyze the potential association rules among these attributes of e-mail's header session. These association rules will be effective to distinguish between spams and legitimate mails. Based on the rules, we can propose a systematic and efficient method to correctly filtrate spam e-mails.

2.2 Decision Tree Data Mining Algorithm

In the related techniques of data mining, decision tree is one of the most popular methods. Decision tree is the data mining method upon the tree data structure. The general statistical method usually can only calculate the distribution of the surface of data whereas decision tree can analyze the potential association rules among the critical attributes from the data. Moreover, the class prediction of the unknown data samples can be further acquired by testing the related attributes' values according to these association rules.

Among various decision tree algorithms, Iterative Dichotmiser 3 [16, 17], called ID3 for short, is one of the most well-known and effective decision tree algorithms. In 1999 [11], Katharina et al. studied the behaviour of ID3 and pointed out that ID3 is better than other decision tree methods, such as C4.5, CHAID, and CART. As compared with the improved methods (for example, C4.5) based on ID3, Ohmann et al. in 1996 indicated that the number of association rules worked out by ID3 is not as numerous as that of C4.5 [14]. Thus, considering the simplicity of rule number, ID3 algorithm possesses the superior characteristic. Hence, we choose ID3 as the data mining technique for this study.

Assumed that a certain attribute is selected out as the target attribute, and this attribute has t kinds of different values. Then the ID3 algorithm will classify all data patterns into t "classes" according to t values of the attribute, which will be labeled in leaf nodes. While the computation is finished, the path constructed from root node to each leaf node forms an association rule. In other words, all of the internal nodes on the path construct a row of "if" judgment of several attributes. With the "then" results presented by the classes signified by the leaf nodes, there are the association rules of "if-then" pattern constructed. The ID3's detailed computation process [16, 17] is described in appendix at the end of this paper.

3 A New Filtering Method Based on Categorized Decision Tree Data Mining

In 2007, Wang and Chen [23] showed that the header session messages of e-mail could be utilized to efficiently filter out spam e-mails. In this section, we propose a two-phase spam filtering method, which will apply the basic information contained in header session of e-mail to effectively identify the unknown e-mail as either a spam or a legitimate mail.

The concept of our method is illustrated in Figure 1. The two-phase method possesses two phases, which are described in following two subsections

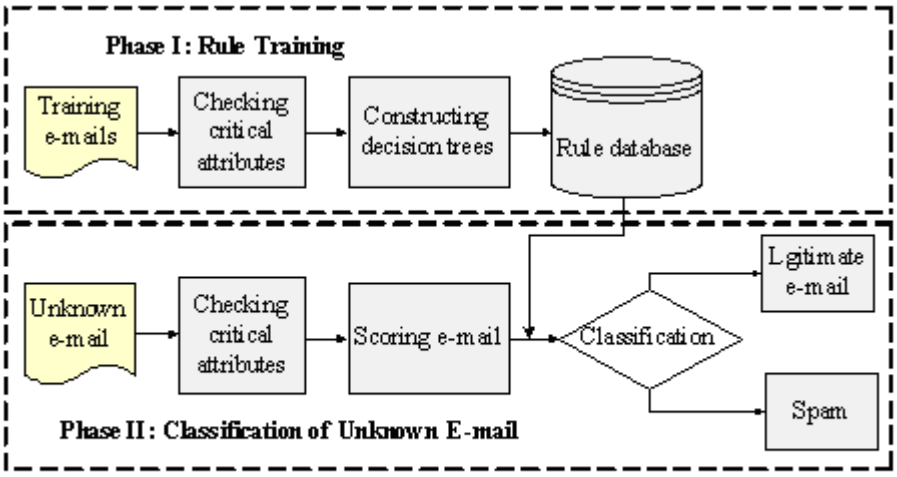

Figure 1: Concept of the two phase spam filtering system

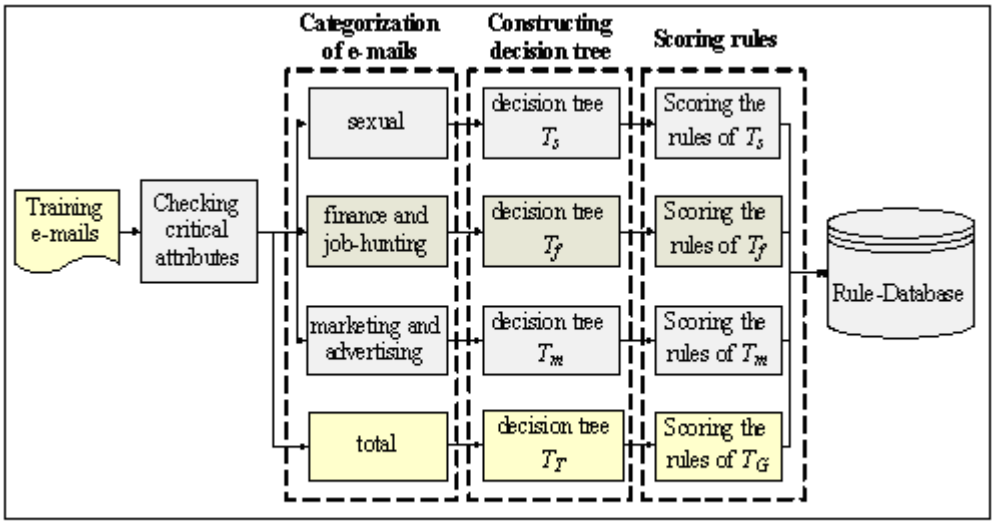

Figure 2: Phase I: Rule training

3.1 Phase I: Rule Training

3.1.1 The Process of Rule Training

The process of this first phase is shown in Figure 2. In this phase, we take numerous e-mails collected in advance as the training data. The attribute "e-mail's type" (spam or legitimate) is designated as the target attribute. By checking the basic attributes (such as title, sender, and date) in the header session of e-mail, we define 9 critical attributes as described in Table 1. While checking each training e-mail, its basic information in header session should be recorded into corresponding attributes (expressed in binary value) respectively. Note that a keyword table is necessary for checking critical attributes of e-mail. We construct the keyword table in advance, which is composed of three kinds of keywords: (1) sexual, (2) finance and job-hunting, and (3) marketing and advertising. Here we omit the detail of the keyword table for simplicity.

While checking the header session, each e-mail should be assigned to the category that there exist the most associated keywords in its title. Thus, all training e-mails can be classified into one or more of the following three categories: (1) "sexual", (2) "finance and job-hunting", and (3) "marketing and advertising". Note that the fourth category, "total", is constructed to collect all training e-mail. Thus, each training e-mail will be copied into the category "total", and probably be additionally classified into one of those three categories if there exists corresponding keyword in its title.

Then we apply the decision tree data mining algorithm ID3 [16, 17] to each of the four categories to find the association rules between the target attribute and those 9 critical attributes of training e-mails. The potential association rules would have been constructed in the resulted decision tree after the ID3 algorithm terminated. Then we score all the rules of each category by the formulas

Table 1: The critical attributes of e-mail

Attribute No.	Critical attributes	Value
1.	Length of senders' field	**1**: the length is more than 9 characters; **0**: the opposite situation.
2.	Field of sender	**1**: having abnormal symbols; **0**: the opposite situation.
3.	Abnormal keywords in sender's field	**1**: sender's field includes abnormal keywords; **0**: the opposite situation.
4.	The title is in abnormal format	**1**: the title includes abnormal symbols; **0**: the opposite situation.
5.	The title includes sexual keywords	**1**: the title includes sexual keywords; **0**: the opposite situation.
6.	The title includes keywords of finance and job-hunting	**1**: the title includes keywords of finance and job-hunting; **0**: the opposite situation.
7.	The title includes keywords of marketing and advertising	**1**: the title includes keywords of marketing and advertising; **0**: the opposite situation.
8.	Time field is abnormal	**1**: the sending date is abnormal; **0**: the opposite situation.
9.	Size of e-mail	**1**: e-mail's size is no more than 6KB; **0**: the opposite situation.
Target attribute: e-mail's type		1: this e-mail is a spam; 0: this e-mail is legitimate.

described as follows.

3.1.2 Scoring the Rules and Building Rule Database

In the following paragraph, we only consider the rules of one category for simplicity. Note that the rules of each category will be scored in the same manner mentioned later.

Given an arbitrary rule R, we suppose that C is the leaf node associated with R. Then the values of C's degree of purity (denoted as ($Purity(C)$)) and degree of support (denoted as ($Support(C)$)) should have been computed when the ID3 algorithm terminated. Based on the values, we now propose some formulas to score each rule R.

Given a rule R whose leaf node is C. Let γ be the number of e-mails whose the target attribute's value was "spam" in C. We first introduce $Spam\_Degree(R)$, which implies R' s "intensity" to identify e-mails to be spams, and can be defined as follows:

$$Spam\_Degree(R) = Purity(C)\ if\ Class(C) = \text{“}spam\text{”},$$
$$and\ Spam\_Degree(R) = (\frac{\gamma}{|C|}) \times 100\%\ otherwise,$$

where $|C|$ is the total number of e-mails in node C. Moreover, we set the degree of support of rule R, $Rule\_Support(R)$, to be the degree of support of leaf node C, that is, $Rule\_Support(R) = Support(C)$. Thus the degrees of support of all rules of the same associated category can be computed. We assumed that $Support_{MAX}$ was the maximum one and $Support_{MIN}$ was the minimum one of the associated category. Then we proposed the weighting function $W(Rule\_Support(R))$. Given a rule R, the weighted value of $Rule\_Support(R)$ is described in following formula:

$$W(Rule\_Support(R)) = \frac{Rule\_Support(R)}{Support_{MAX} + Support_{MIN}} \times 100\%. \quad (1)$$

The weighted values of rule's support for all rules of the same associated category then can be computed by Equation (1). Assumed that W_{MAX} is the maximum one and W_{MIN} is the minimum one among all weight values of the associated category. Given a rule R, the function $S(Rule\_Support(R))$ is to calculate the score for $Rule\_Support(R)$ as follows :

$$S(Rule\_Support(R)) = \frac{W(Rule\_Support(R) - W_{MIN})}{W_{MAX} - W_{MIN}} \times 100\%$$

Finally, the score of rule R can be computed by the following formula:

$$Score(R) = (0.7 \times Spam\_Degree(R) + 0.3 \times S(Rule\_Support(R))) \times 100.$$

Together with the score data, all of these rules of the same associated category will be stored into the Rule-Database after computing their scores. Thus, each category of rules will be scored and recorded into the Rule-Database, which should be accessed in the next phase to classify the unknown e-mails. For each category of rules, we choose out the minimum rule's score among the

rules with *Spam_Degree* more than 80%, and set it as the *threshold*, denoted as θ, for judging whether the unknown e-mail is spam.

3.2 Phase II: Classification of Unknown E-mail

The second phase is to judge the unknown e-mail to be either a legitimate mail or a spam mail. When an unknown e-mail arrives in our system, the basic information contained in its header session will be checked and recorded as the attributes in Table 1. Each e-mail will be classified into one category that there exist the most associated keywords in its title. If there is no any keyword found in e-mail's title, we will classify it into the fourth category "total".

Then, according to the critical attributes' values in the header session, the unknown e-mail will dovetail with some association rule of the associated category in Rule-Database. Thus, we define the score of this unknown e-mail as the score of the dovetailed association rule. Then we can judge whether the unknown e-mail is spam by its score. If the score of any unknown e-mail is higher than the threshold θ of the associated category, it should be regarded as a spam e-mail.

However, a reversing mechanism is necessary for our system to avoid misadjudging the spam as a legitimate mail. We build a Critical-Reversing-Item-Table whose each item is defined by an individual score as recorded in Table 2. If an unknown e-mail is judged as a legitimate e-mail (i.e., the score of this unknown e-mail is lower than the threshold θ), this unknown e-mail must be re-examined according to the items of Critical-Reversing-Item-Table to accumulate its additional scores. It should be judged as a legitimate e-mail if its final total score is less than the threshold, and a spam e-mail otherwise.

Table 2: Items of reversing mechanism

Critical-Reversing-Item-Table	
Critical items	Scores
Senders' field is abnormal	+20
Senders' field include abnormal key words	+20
Title is not in legitimate format	+20
Title includes abnormal key words	+20
Time field is abnormal	+20

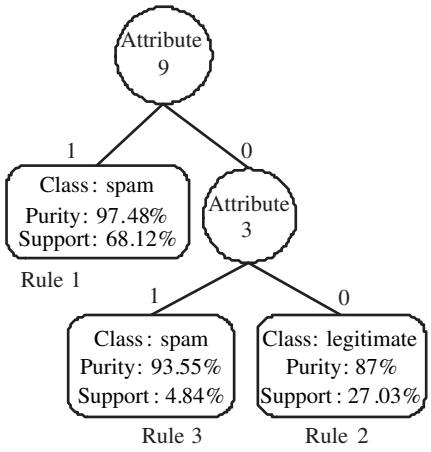

Figure 3: The resulted decision tree for the category "sexual class"

4 Experimental Results

In this section, we arrange the experiment to confirm the accuracy and efficiency of our method. We collect 2000 Chinese e-mails as the training data for the first phase. Then, we arrange 600 unknown Chinese e-mails for the experiment of the second phase.

4.1 Categorization of Training E-mails

The 2000 Chinese training e-mails include 1000 spam e-mails and 1000 legitimate e-mails. According to the data in header session, each training e-mail is recorded into 9 critical attributes as defined in Table 1. By checking the header session, each of the 2000 e-mail is copied into one category with the most associated keywords in its title. As shown in Table 3, we obtain the following four categories of training e-mails: (1) the category "sexual" has 1280 e-mails, (2) the category "finance and job-hunting" has 237 e-mails, (3) the category "marketing and advertising" has 438 e-mails, and (4) the category "total" has 2000 e-mails. Moreover, we construct a Chinese keyword table, which was composed of three kinds of keywords: (1) sexual (142 keywords), (2) finance and job-hunting (93 keywords), and (3) marketing and advertising (96 keywords). Here we will omit the detail for simplicity.

4.2 Construction of Decision Tree and Training of Rules

By applying the ID3 algorithm to each category of training e-mails, we construct a decision tree and obtain several rules. We discuss the details in the following four cases.

1) **The resulted rules of the category "sexual":** The resulted decision tree of the category "sexual" is illustrated in Figure 3. There we find totally 3 association rules among the target attribute and 9 critical attributes. By executing the formulas of scoring rules, we acquire the scores of all rules, which are shown in Table 4. Then the minimum rule's score 65.49 is sought out from the rules with Spam_Degree more than 80%. And we set it as the threshold θ of this category.

2) **The resulted rules of the category "finance and job-hunting":** The resulted decision tree of

Table 3: The four categories of training e-mails

category	amount	legitimate e-mails	spam
"sexual"	1280	327	953
"finance and job-hunting"	237	23	214
"marketing and advertising"	438	72	336
"total"	2000	1000	1000

the category "finance and job-hunting" is shown in Figure 4. There exist 5 association rules among the target attribute and 9 critical attributes. The scores of all rules are recorded in Table 5. Then the minimum rule's score 70.00 is sought out from the rules with Spam_Degree more than 80%. We set it as the threshold θ of this category.

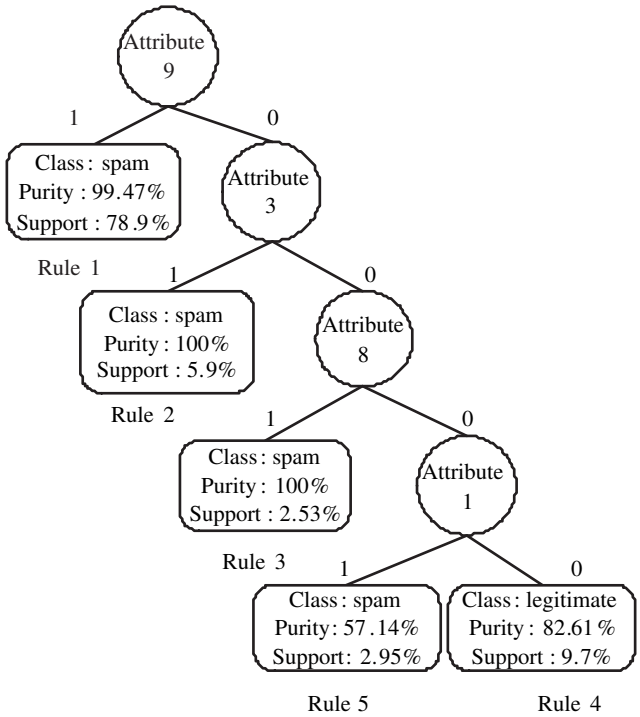

Figure 4: The resulted decision tree for the category "finance and job-hunting"

Table 4: The scores of rules for the category "sexual class"

Rule	Purity	Support	Spam_Degree	Rule's score
1	97.48%	68.12%	97.48%	94.26
2	87%	27.03%	13%	18.23
3	93.55%	4.84%	93.55%	65.49

3) **The resulted rules of the category "marketing and advertising":** As shown in Figure 5, there exist 5 association rules among the target attribute and 9 critical attributes in the category "marketing and advertising". The scores of all rules are recorded in Table 6. The minimum rule's score 71.57 is chosen from the rules with Spam_Degree more than 80%. We set it as the threshold θ of this category.

Table 5: The scores of rules for the category "finance and job-hunting"

Rule	Purity	Support	Spam_Degree	Rule's score
1	99.47%	78.9%	99.47%	97.76
2	100%	5.9%	100%	71.24
3	100%	2.53%	100%	70.00
4	82.61%	9.7%	17.39%	14.81
5	57.14%	2.95%	57.14%	40.15

Table 6: The score of rules for the category "marketing and advertising"

Rule	Purity	Support	Spam_Degr.	Rule's score
1	94.77%	74.2%	99.47%	95.43
2	86.27%	11.64%	13.73%	13.79
3	62.5%	3.65%	62.5%	44.75
4	80%	1.14%	20%	14.00
5	97.56%	9.36%	97.56%	71.57

4) **The resulted rules of the category "total":** As illustrated in Figure 6, we find 4 association rules among the target attribute and 9 critical attributes in the category "finance and job-hunting". The scores of all rules are recorded in Table 7. The mini-

Table 7: The scores of rules for the category "total"

Rule	Purity	Support	Spam_Degr.	Rule's score
1	93.47%	36%	6.53%	25.71
2	97.48%	43.6%	97.48%	94.25
3	86.99%	17.35%	13.01%	18.23
4	93.55%	3.1%	93.55%	65.49

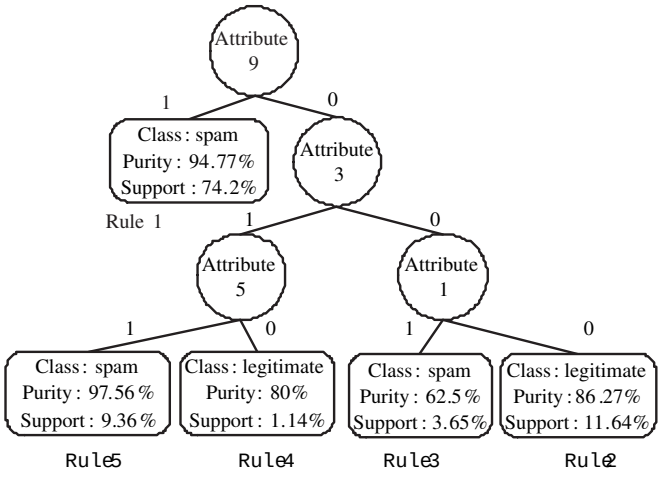

Figure 5: The resulted decision tree for the category "marketing and advertising"

mum rule's score 65.49 is chosen from the rules with Spam_Degree more than 80%, and we set it as the threshold θ of this category.

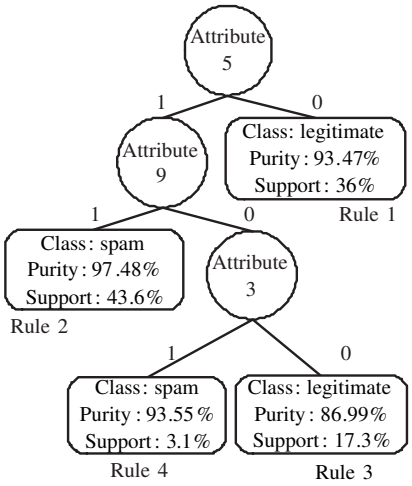

Figure 6: The resulted decision tree for the category "total"

4.3 The Analysis about Accuracy and Performance

Then, we arrange 600 Chinese e-mails, which are composed of 300 spam e-mails and 300 legitimate mails, as the "unknown e-mails" for the experiment of the second phase. The critical attributes in the header session of each unknown e-mail are examined first. Then, as shown in Table 8, all the 600 e-mails are arranged into four categories according to keyword of the associated category in its title. Moreover, each unknown e-mail's score is then decided according to the principles proposed in Section 3.2. It should be judged as a legal e-mail if its final total score was less than the threshold θ, and the spam otherwise.

The following efficacy assessment indexes [2, 9] are constructed based on four items in Table 9:

- *Recall*: The index refers to the ability of filtering mechanism to correctly find spam. It is defined by the following formula:

$$Recall = (C_A/(C_A + C_C)) \times 100\%.$$

- *Precision*: When precision is higher, it is less likely to treat legitimate mails as spam. It is defined by the following formula:

$$Precision = (C_A/(C_A + C_B)) \times 100\%.$$

- *Accuracy*: the percentage of total e-mails that are correctly recognized. It is defined by the following formula:

$$Accuracy = ((C_A + C_D)/n) \times 100\%.$$

The experimental result of this research is recorded in Table 10. Obviously, our method is better than others in the Accuracy. Although the method based on Bayes classifier proposed by Tretyakov [22] has better Precision, the spam filtering method of this paper has better effect with regard to Recall and Accuracy. Moreover, although the Recall of this method KNN (the k nearest neighbors approach proposed by Delany et al. [4]) is better than that of our method, the spam filtering method proposed in this paper has better effect with regard to both Precision and Accuracy. Note that the method proposed in this paper check only the header session of e-mails. Obviously, according to the datum of experiment, we have confirmed that the method proposed by this research can effectively recognize spam and filter more spam without analyzing the full text of e-mail.

5 Conclusion

In this paper, we propose an efficient spam filtering method based on decision tree data mining technique. We dig out the association rules among basic attributes in the header session of e-mails, and apply these rules to develop an efficient two-phase spam filtering method. Different from those methods of checking the complete content of e-mail, we judge e-mail simply by its basic header data without screening overall content of e-mail and analyzing the key words.

According to the experimental results, the efficiency of the spam filtering method proposed in this paper can be evaluated in the following datums: the rate of accuracy is 96.5%, the rate of precision is 96.67%, and the rate of recall is 96.35%. Obviously, these datums are not lower than that of other present filtering methods of checking e-mail content. It shows that our method can efficiently filtrate the spam e-mails by checking only the header session of e-mails, which will reduce the cost of computation and do not consume too many system resources.

Table 8: The categorization and classification of unknown e-mails

category	Total number	legitimate mail	spam
"sexual"	122	3	119
"finance and job-hunting"	54	5	49
"marketing and advertising"	123	22	101
"total"	600	300	300

Table 9: Four cases of judgment

	e-mail's type in reality	
	Spam	Legitimate mail
To be judged as spam	C_A	C_B
To be judged as legitimate mail	C_C	C_D

References

[1] X. Carreras, and L. Marquez, "Boosting tree for anti-spam email filtering," *Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, BG, 2001.

[2] W. W. Cohen, "Learning rules that classify e-mail," *Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

[3] L. F. Cranor, and B. A. LaMacchia, "Spam!," *Communications of the ACM*, vol. 41, no. 8, pp. 74-83, 1998.

[4] S. J. Delany, P. Cunningham, and B. Smyth, "ECUE: A spam filter that uses machine learning to track concept drift," *Proceedings of the 17th European Conference on Artificial Intelligence (PAIS stream)*, pp. 627-631, 2006.

[5] P. J. Denning, "ACM president's letter: electronic junk," *Communications of the ACM*, vol. 25, no. 3, pp. 163-165, 1982.

[6] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048-1054, 1999.

[7] F. F. Riverola, E. L. Iglesias, F. Diaz, J. R. Mendez, and J. M. Corchado, "Applying lazy learning algorithms to tackle concept drift in spam filtering," *Expert Systems with Applications*, vol. 33, pp. 36-48, 2007.

[8] F. Fdez-Riverola, E. L. Iglesias, F. Diaz, J. R. Mendez, and J. M. Corchado, "SpamHunting: an instance-based reasoning system for spam labeling and filtering," *Decision Support Systems*, vol. 43, pp. 722-736, 2007.

[9] W. F. Hsiao, and T. M. Chang, "An incremental cluster-based approach to spam filtering," *Expert Systems with Applications*, vol. 34, pp. 1599-1608, 2008.

[10] J. He, and T. Bo, "Asymmetric gradient boosting with application to spam filtering," *Proceedings of Fourth Conference on Email and Anti-Spam CEAS*, 2007.

[11] I. Koprinska, J. Poon, J. Clark, and J. Chan, "Learning to classify e-mail," *Information Sciences*, vol. 177, pp. 2167-2187, 2007.

[12] P. Lieven, "Pre-MX spam filtering with adaptive greylisting based on retry patterns," *Heinrich Heine Universitat Dusseldorf*, pp. 5-8, Germany, 2006.

[13] G. Manco, E. Masciari, M. Ruffolo, and A. Tagarelli, "Towards an adaptive mail classifier," *Proceedings of Conference of the Italian Association of Artificial Intelligence*, 2002.

[14] C. Ohmann, V. Moustakis, Q. Yang, and K. Lang, "Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain," *Artificial Intelligence in Medicine*, vol. 8, no. 1, pp. 23-36, 1996.

[15] T. R. Payne and P. Edward, "Interface agents that learn: an investigation of learning issues in a mail agent interface," *Applied Artificial Intelligence*, pp. 1-32, 1997.

[16] J. R. Quinlan, "Discovering rules from large collections of examples: A case study," *Expert Systems in the Microelectronic Age*, D. Michie, Editor, Edinburgh, Scotland: Edinburgh University Press, 1979.

[17] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[18] P. J. Resnick, D. L. Hansen, and C. R. Richardson, "Calculating error rates for filtering software," *Communication of the ACM*, vol. 47, no. 9, pp. 67-71, 2004.

[19] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[20] S. Sinclair, "Adapting Bayesian statistical spam filters to the server side," *Journal of Computing Sciences in Colleges*, vol. 19, no. 5, pp. 344-346, 2004.

[21] K. D. Stark, and D. U. Pfeiffer, "The application of non-parametric techniques to solve classification

Table 10: The comparison of different spam filtering methods

Methods	Checking the complete content of e-mail	Recall(%)	Precision(%)	Accuracy(%)
The method based on Bayes classifier proposed by Tretyakov [22]	Yes	87.44	100	94.49
Incremental clustering-based classification (ICBC) proposed by Hsiao and Chang [9]	Yes	96.1	95.76	94.73
The k nearest neighbors (KNN) approach proposed by Delany et al. [4]	Yes	98.66	95.93	96.42
The spam filtering method proposed in this research	**No**	**96.35%**	**96.67%**	**96.5%**

problems in complex data sets in veterinary epidemiology - an example," *Intelligent Data Analysis*, vol. 3, no. 1, pp. 23-35, 1999.

[22] K. Tretyakov, *Machine Learning Techniques in Spam Filtering*, Technical report, Institute of Computer Science, University of Tartu, 2004.

[23] C. C. Wang, and S. Y. Chen, "Using header session message to anti-spamming," *Computers & Security*, vol. 26, pp. 381-390, 2007.

Appendix

The construction process of decision tree starts from root node and all of the data patterns are initially included in the root [16, 17]. ID3 algorithm will select an unselected attribute with the maximum Information Gain. According to the value of this attribute, ID3 algorithm divides all the data patterns contained in this node into children nodes. Subsequently, each children node respectively repeats the same process for its own data patterns.

There are two conditions to end the computation: (1) all of the attributes are selected; (2) all of the data patterns contained in the node are of the same class. If any of the two conditions is satisfied, the current node will be signified as a leaf node. Given a leaf node C, we assign the class of C, denoted as $Class(C)$, to be the class with the most data patterns in C. And then we calculate Purity (denoted as $Purity(C)$) and Support (denoted as $Support(C)$) for this leaf node to end this node's execution of algorithm. The formulas of $Purity(C)$ and $Support(C)$ are defined as:

$$Purity(C) = (|Class(C)|/|C|) \times 100\%;$$
$$Support(C) = (|C|/N) \times 100\%.$$

Where the number of data patterns in node C is denoted by $|C|$, and N is the number of total data patterns.

The ID3 algorithm is summarized as follows.

Step 1. If all data patterns associated with node C belong to the same class, then label node C as this class, set C to be a leaf node, compute $Purity(C)$ and $Support(C)$, and stop;

Step 2. If all attributes are "selected" then check all classes of the data patterns contained in node C, and select the class with maximum data patterns. Label node C as this class, set C to be a leaf node, compute $Purity(C)$ and $Supoort(C)$, and stop;

Step 3. Compute the Information Gain $G(A)$ for each unselected attribute A, and select the one with maximum Information Gain. Divide all data contained in node C into disjoint children nodes according their values of the select attribute;

Step 4. Treat each children node as node C, and continue the algorithm recursively from Step 1.

Considering a certain attribute A on node C, the Information Gain $G(A)$ of attribute A will concern the Entropy $E(C)$ of node C, which is calculated by using the following formula:

$$E(C) = -\sum_{i=1}^{t} \frac{p_i}{n} \times \log_2 \frac{p_i}{n},$$

where t was the total number of classes associated with C, $p_i$was the total number of data patterns corresponding to the i-th class in C, and n is the total number of data patterns in C.

The Information Gain $G(A)$ of attribute A is calculated by using the following formulas:

$$G(A) = E(C) - E^{+}(A);$$
$$E^{+}(A) = \sum_{j=1}^{k} (n_j/n) \times E(C_j),$$

where k is the number of possible attribute values of A, C_j (for $1 \leq j \leq k$) is a subset of C including the data patterns corresponding to the j-th possible attribute value of A, and n_j is the total number of data patterns contained in C_j.

Jyh-Jian Sheu is currently an Assistant Professor in the Department of Information Management, National Dong Hwa University, Taiwan. He received BS degree in Management Information Systems from National Chengchi University, Taiwan. And he received his MS and PhD degrees in Computer and Information Science from National Chiao Tung University, Taiwan. Sheu's primary research interests include data mining, Internet security, and interconnection networks.