Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2014, *8,* 435-441

Open Access

Bayesian Spam Filtering Mechanism Based on Decision Tree of Attribute Set Dependence in the MapReduce Framework

Yanyan Guo[1], Lei Zhou[2], Kemeng He[1], Yuwan Gu[1] and Yuqiang Sun[1,*]

[1]*School of Information Science & Engineering, Changzhou University, Jiangsu, Changzhou, 213164, China;* [2]*Department of Computer, Henan Institute of Engineering, Henan, Zhengzhou, 451191, China*

Abstract: Bayesian spam filtering is a classification method based on the theory of probability and statistics, and the Bayesian spam filtering based on Mapreduce can solve the defect of the traditional Bayesian spam filtering that consumes large amounts of system resources and network resources when the mail set is pre-training. It needs to classify mails manually in the pre-training phase of mail set, which consumes a lot of human and financial resources and affects the efficiency of the system. Bayesian spam filtering mechanism based on decision tree of the attribute sets dependence in the MapReduce framework which is presented in this paper. And the decision tree of attribute sets dependence is used in the training stage of the mail set, which improves execution efficiency of the system by lowering the time complexity.

Keywords: Bayesian spam filtering, decision tree of attribute sets dependence, MapReduce.

1. INTRODUCTION

People have suffered unsolicited e-mail for years, they have spent a lot of time and energy clearing fraudulent or unwanted messages from their mailbox which requires large amounts of network resources and system resources. These messages bring a number of security issues, so spam filtering technology has become a hot spot in network information security research gradually.

There are many anti-spam filtering technologies now, such as black-and-white list technology, Content-based filtering technology, host name of the reverse authentication technology, and so on. The Bayesian mail filtering technology [1] in this paper is content filtering technology based on Statistics. It has strong classification ability and can ensure high accuracy, so it has been widely used. Bayesian classification algorithm does not have to set rules in advance, and does not need to analyze content of mail. Through the analysis of characteristic words and the category of text, we get the statistical models. But the early stage of email set training with the Bayesian spam filters consume a lot of system resources and network resources. Now the data we face is growing in TB and even more and this leads to the low-efficiency by the traditional single-point operation method, and high resource consumption that can't be supported by the current generation of computers.

With the rapid development of computer technology, cloud computing has become the future trend of distributed computing. Google's MapReduce programming framework [2] is a representative technology of the cloud computing technology for large-scale data sets in parallel computing. Compared to the traditional parallel computing models, it has higher computing efficiency, and better performance. The Bayesian spam filtering mechanism is based on MapReduce that are presented in this paper and that solves the defect that traditional algorithms can't handle; which is, to handle huge amounts of data efficiently. It also can shorten the process of learning and classification of mail.

Spam filtering mechanism based on Bayesian has two phases: fist, the training of data set; the second is the test of data set. Currently, the training of data set is processed manually, sort out spam and legitimate mails, and then form a knowledge base. However, different users have different mail sets, and the format of mails are also different, so different user have different divisions of mails, and it will waste a lot of time if filtered manually. So the decision tree algorithm based on attribute sets dependence is used in the training stage of the mail set. The algorithm based on the mail header information is used to reduce the attributes by considering the interdependence between attributes, thus it can accurately remove redundant attributes. So Bayesian spam filtering mechanism based on the decision tree of attribute sets dependence in Mapreduce framework; and can reduce the cost of learning and classification of mail, and increase the system's performance.

2. RESEARCH BACKGROUND

2.1. Mapreduce

MapReduce is a distributed parallel programming model for parallel processing of large data. Data is stored in a distributed file system (DFS) and is represent with key - value pairs (key, value), MapReduce involves two steps. "Map" step: The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller

1874-110X/14

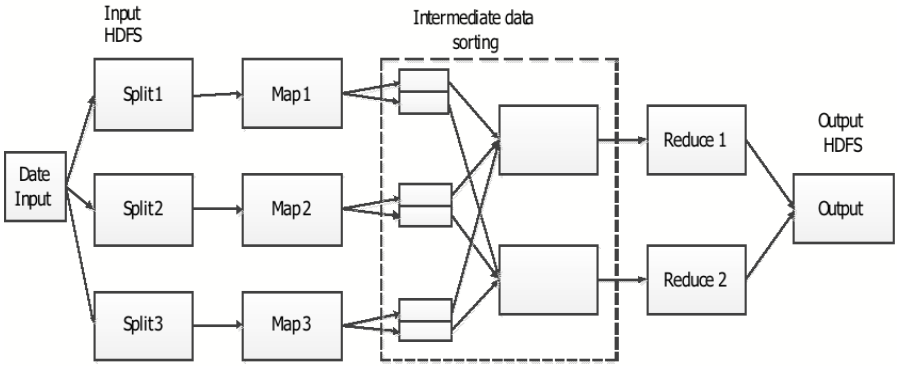

Fig. (1). Operation process of Mapreduce.

problem, and passes the answer back to its master node; Reduce step: The master node then collects the answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve. The detailed steps are that divide data into uncorrelated splits at first, then divide the splits into a number of key/value pairs like < Key1, Value1 > then pass it to the Map function, which is used to process each key-value pair to form a new key-value pair <Key2, Value2>, and then summarize the data of the same Key2, at the last output the intermediate result <Key2, list (Value2) >. The middle output results of the Map as input is to Reduce phase, and after they are handled accordingly by the Reduce function forming key-value pairs <Key3, Value3> which output to the HDFS、HBase or specified the location of the database according to user's need. The MapReduce implementation of operational processes as Fig. (**1**) shows [3].

2.2. Attribute Set Dependence

Decision attribute depends on an individual condition attribute is called attribute dependence, and decision attribute depends on two or more conditions is called attribute set dependence. Sometimes, multiple property dependence is exactly the same, so we cannot distinguish which property is the main and which is a secondary. It is not enough to accurately identify which property is most important only by attribute dependency. So we also need to look at dependencies between it and other properties, that is, the attribute set dependence must be considered [4, 5].

Definition 1: Let an information system S = (U, A, V, f) is an ordered quaternion, U is a training set and it is composed of objects; A is the property set, which can be divided into the condition attribute set C and the decision attribute set D, that is C ∪ D = A, C∩ D = Á;

V is the value of all properties; f is the corresponding rule: Specify the value of each object in the training set.

Definition 2: For a given information system S = (U, A, V, f), we make RA, x and y∈ U. If and only if there exists a∈A,f(x,a)=f(y,a) then x and y on R is not resolved. So for RA, defines A's equivalence relation on U (R is actually an attribute subset, also known as knowledge), notes for the IND (R) and without causing confusion, also shorthand for R.

Exists x ∈ U, the equivalence class of x is generated by the equivalence relation R is denoted by $[x]_R$. Thus, the domain U is divided by the corresponding equivalence relation R and is recorded as U／R={U_1, U_2,…,U_n).

Definition 3: For a given information system S = (U, A, V, f), then there exists X ∈ U and an equivalence relation RA of domain U. Defined the lower approximation of subsets x about knowledge R as follows:

$$\underline{R}(x)=\{x\mid(\forall x\in U)\wedge([x]_R\subseteq X)\} \tag{1}$$

The set $posR(X)=\underline{R}(x)$ is called R positive region of x

Definition 4: Let K = (U, R) is an approximation space, R is the attribute set, P and Q are the properties on R, P is the condition attribute and Q is the decision attribute. If P can launch Q, in other words Q is dependent on the P that we denoted by P => Q, or called Knowledge Q depends on the knowledge P with the dependence k ($0 \le k \le 1$). If and only if :

$$\gamma_p(Q)=\frac{card(POS_p(Q))}{card(U)}=\frac{Card(\bigcup_{x\in U/D}\underline{P}(x))}{Card(U)} \tag{2}$$

In the formula (2) for any set S and the card (S) represent the cardinality of set S, that is, the number of elements that are contained in the collection S.

2.3. Bayesian Filtering Technology

Bayesian filtering technology based on Bayes theorem, which using Bayesian formula to calculate the posterior probability through the prior probability of the data and selecting the class with largest posteriori probability as the class of the object. There is a derivation based on the Bayesian formula which is as following:

$$P(y_i\mid x)=\frac{P(x\mid y_i)P(y_i)}{P(x)} \tag{3}$$

Among 1: Let $x=\{a_1,a_2,...,a_m\}$ be a term to be classified, and each a is a characteristic property of x.

2: There are a set of classes $C=\{y_1,y_2,...,y_n\}$.

3: Calculating

$$P(y_1\mid x),P(y_2\mid x),...,P(y_n\mid x)$$

Table 1. Description for comprehensive mail header and the feature set of main image features in the mail content.

Features	Value	Feature Descriptions
A_1	int	Number of recipients
A_2	int	Mail relay times, namely the tag number of "received" in the mail header
A_3	int	Number of the domain name In the "received" and IP does not match
A_4	int	Number of mail routing information interrupt
A_5	int	Number of the domain name is missing in the "received" send site
A_6	0,1	If the sender's domain along the line in the mail delivery process that the value is 0, otherwise 1
A_7	0,1	If the original sending address in the "from" consistent with original sending address in the "received" the value is 1, otherwise 0
A_8	0,1	If the destination address in the "to" consistent with the actual recipient address in the "received" the value is 1, otherwise 0
A_9	0,1	If "delivered - to" and "to" is consistent the value is 1, otherwise 0, and the default value is 1
A_{10}	0,1	If "return path" and "from" is consistent the value is 1, otherwise 0, and the default value is 1
A_{11}	0,1,2	Mail types, sent directly is 0, reply is 1, transfer is 2
A_{12}	int	Number of attachments
A_{13}	0,1	The title exists or not, no title is 0, otherwise 1
A_{14}	int	Number of http links in the body
A_{15}	0,1	Whether the mail body contains tables, there is a table is1, otherwise 0
A_{16}	0,1	Whether the mail contains pictures, there is a picture is1, otherwise 0
A_{17}	0,1	Mail format (text or HTML), text is1, HTML is 0
A_{18}	int	Mail encoding type
A_{19}	int	Frequency of the character "!" in the mail
A_{20}	int	Frequency of the character "$" in the mail
A_{21}	0,1	The case of carbon copy, there is carbon copy 1, otherwise 0
A_{22}	0,1	Delivery time 8：00-23：00 is 0,23：00-8：00 is 1
A_{23}	0,1,2	Mail length is less than 1M is 0, more than 5M is 2, the others is1

4: If $P(y_k \mid x) = \max\{P(y_1 \mid x), P(y_2 \mid x), ..., P(y_n \mid x)\}$

that $x \in y_k$ s.

3. SPAM FILTERING OF DECISION TREE BASED ON ATTRIBUTE SET DEPENDENCE

3.1. Feature Selection Based on Mail Header

Users usually make judgment from the main image features first, without reading the content of the mail carefully when determining whether a mail is spam or not. Email format is defined by the RFC822, which is a semi-structured text files, including header and body. The Header contains key information such as From, Subject, Date and so on. Although it contains a large amount of mail features, it is not enough to judge the mail only by the characteristics of the mail header. Thus we need to take into account, the mail format, attachments, and other characteristics into account, we gives a comprehensive mail header of 23 properties. And the feature set of main features in the mail content as listed in Table **1**.

According to the definition of information systems based on rough sets, set the set of mails as the domain, and make each mail as an object in the domain. Make the feature set as the condition attribute set, and the weight of the feature item in the mail text as an attribute value of the attribute on the object, the type of a mail as decision attribute. Such a decision table is a decision-making system of unspecified emails, where each row represents a mail and each column represents an attribute. The values of A_1 to A_{23} represent the mail's condition attributes, and D represents the decision attributes. The value of V_D is V_D={d}={0,1,2} which represents the normal mail, suspicious mail and spam. We discrete the data sets of above table into decision-making information table as listed in Table **2**.

Table 2. Decision-making information table.

U	A_1	……	A_{23}	D
U_1	1	……	0	0
U_2	5	……	2	2
……	……	……	……	……
U_n	3	……	1	1

3.2. Spam Filtering Algorithm of Decision Tree Based on Attribute Set Dependence

Input: Decision-making table (U, CD, V, f)

Output: Decision tree

Algorithm steps:

Step 1: Calculating the given decision table (U, CD, V, f) about the division of property:

$U/D=\{X_1,X_2,\ldots,X_n\}$ and $X_i=[x]_{di}$,i=1,2,……n

Step 2: for(i=1;i<24;i++)

for(j=1;j<n+1;j++)

{Calculating $_i(X_j)$ according to formula (1) ;

Calculating $\gamma_{Ai}(d)$ according to formula (2) }

Step 3: Calculating $A_{i0}=\text{argmax}\{\gamma_{Ai}(d)\}$ and 1<=i<=23

Step 4: Calculating $U/A_{i0}=\{U_1,U_2,\ldots\ldots U_k\}$

Step 5: for (i=1;i<=k+1;i++)

{If the samples of the U_i belong to the same class, then the end;

Otherwise, repeating Step 2 to Step 4 for U_i }

Step 6: The attribute which has larger value of $\gamma_{Ai}(d)$ and contains the same condition attribute is split attribute. The attribute which leads to the value of $\gamma_{Ai}(d)$ is minimum can be discarded as a redundant attribute until all importance of the attributes is distinguished. Such a decision tree is constructed successfully.

4. BAYESIAN SPAM FILTERING BASED ON MAPREDUCE

4.1. Spam Filtering Based on Bayesian

After the decision tree is built, we can export the classification rules which are called the rule extraction. In this procedure the training email set is classified according to the rules, and then the suspicious mails are classified into regular mail and spam manually according to the needs of the user. At last, the strings W_1, W_2, ... , Wm of each mail is extracted in the mail set as feature words, and the frequency of feature words f_1, f_2, …, f_n, are counted, forming a knowledge base.

A known mail document d with its vocabulary set $W=\{w_1, w_2, \ldots, w_m\}$, $C_i=\{c_1, c_2\}$ is class variable, in this $c_1=\{$ Ham(Normal mail) $\}$, $c_2=\{$ Spam $\}$. The target of classification is to predict the category of unknown category mails: $c=\max\{p(c_i/d)\}$. According to Naïve Bayesian hypothesis, each featured item in the email is independent in comparison with the class attribute. So from the Bayesian formula (3) we get that

$$P(c_i/d)=\frac{P(c_i)\prod_{w_i\in di}P(w_i/c_i)}{P(d)} \quad (4)$$

For the same mail text, P(d) does not change, so we only need is to calculate $P(c_i)\prod_{w_i\in d}P(w_i/c_i)$. $P(c_i)$ is the prior probability of class, which is estimated by the training set, and $P(w_i/c_i)$ is the probability of the feature w_i , occurs in category c_i.

The process of spam filtering based on Bayesian is shown in Fig. (**2**) which Shows the Learning module, based on the contents in the figure includes extracting strings W_1, W_2, ... , W_m of each mail in the mail set as feature words, counting the feature words corresponding frequency f_1, f_2, …, f_n, and forming a knowledge base. The classification module based on probability is to calculate probability of class according to Bayesian formula.

4.2. Bayesian Spam Filtering Mechanism of the Decision Tree Based on Attribute Sets Dependence Under the MapReduce Framework

The literature [6] proved that there is no necessary link between the performance of the Bayesian classification model and whether they meet the independence hypothesis or not. So there is no significant effect that improving independence between the attributes of feature items improves the Bayesian classification model. The model of Bayesian spam filtering, based on MapReduce that is presented in this paper is still based on the "the independence hypothesis" in order to prevent increase in the complexity of the algorithm. The process of Bayesian spam filtering mechanism based on the decision tree of attribute sets dependence in the Mapreduce framework as Fig. (**3**) shows.

The description of four MapReduces in the Fig. (**3**) is shown in Table **3**.

5. EXPERIMENT AND THE RESULT ANALYSIS

5.1. Experiment Environment

To verify the performance and efficiency of the proposed algorithm, the experiment was performed in two environments: Experiment A was to learn and classify the mail with Bayesian spam filtering mechanism based on MapReduce on a Hadoop cluster (calculating the priority probability with

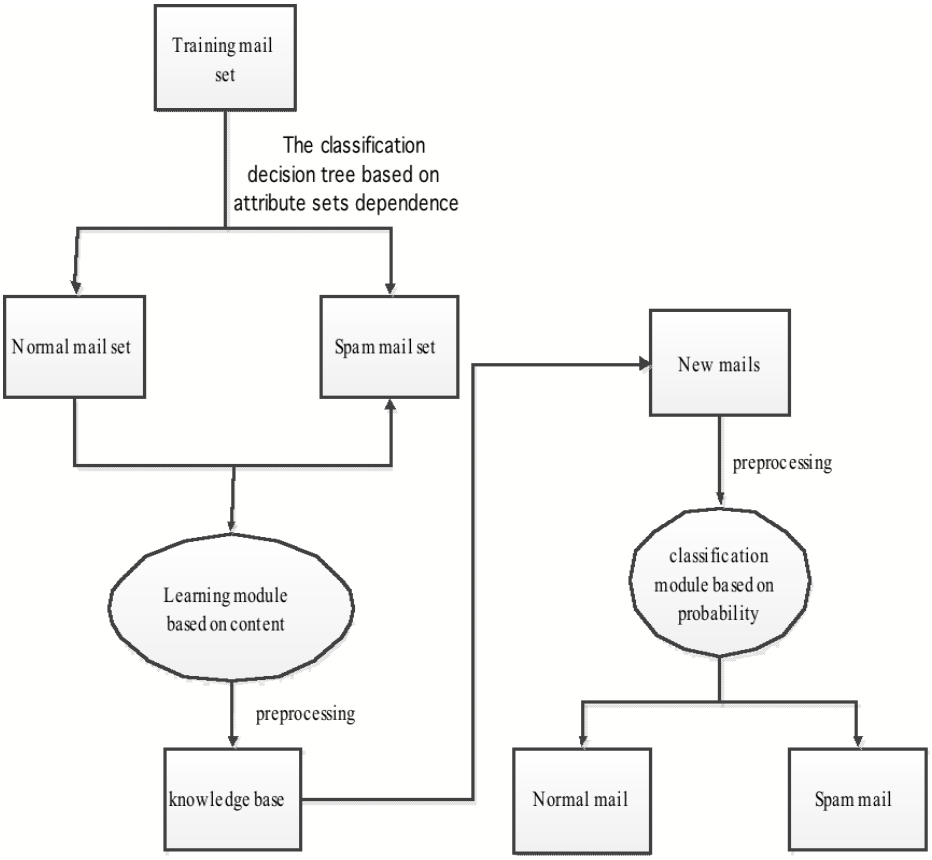

Fig. (2). Process of spam filtering based on Bayesian.

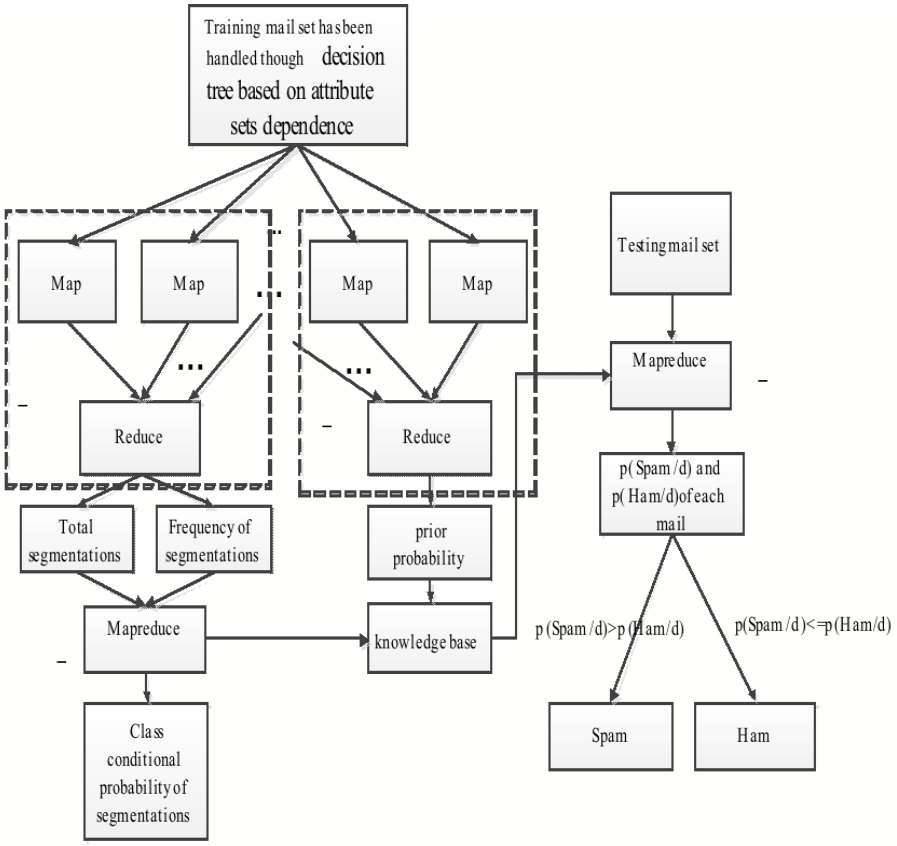

Fig. (3). Bayesian spam filtering mechanism of the decision tree is based on the attribute sets dependence.

Table 3. Process of MapReduce.

Process of MapReduce	Input	Output
(1)	Training mail set has been handled though decision tree based on attribute sets dependence	The feature words and their frequency of each mail class; Class name of mail and the total number of feature words
(2)	Training mail set has been handled though decision tree based on attribute sets dependence	Class name of mail and the corresponding prior probability
(3)	Output of (1) as input	Feature words and the corresponding class conditional probability
(4)	The pre-formed knowledge and the mail set which need to test	Mail name and the corresponding p(Spam/d), p(Ham/d)

Table 4. Experimental results.

Experimental Group	Mail Type	Distinguish the Number of Normal Mails	Distinguish the Number of Spam
A	Normal mails	146	4
	spam	8	142
B	Normal mails	145	5
	spam	8	142

Table 5. Performance indicators of experiments.

Experimental Group	Accuracy Rate /%	Recall Rate /%	Precision Rate /%
A	97.33	94.67	96.00
B	96.67	94.67	95.67

manual classification), and each node is configured with CPU Intel Xeon E1230,

The dominant frequency is 3.0 GHz and the memory is 4 Gbytes; experiment B had the same conditions as experiment A, except that it depended on Bayesian spam filtering mechanism based on the decision tree of attribute sets dependence in the MapReduce framework.

The experiments used 3,000 mails which came from a standard data set SEWM200 [7], including 1500 spam and 1500 normal mails. The mails were divided into 10 parts, each of them with 150 spam and 150 normal mails, we in turn selected a part as testing mail set and the other nine parts as training mail sets. At last we take the average value of 10 results as the final result. In experiment B, the suspicious mails was obtained according to the algorithm based on.

5.2. Evaluation Index

(1) Accuracy rate: The rate of judging the normal mails correctly, that is P = distinguish the number of normal mails／the total numbers of actual mails are normal.

(2) Recall rate: Spam detection rate R= distinguish the number of spam／the total number of actual mail that are spam, which reflects the capacity of the system checks out spam.

(3) Precision rate: A = number of mails that were correctly classified／total number of mails, which reflects the capacity of the system classify the mail correctly.

5.3. Experimental Results

The experimental results as shown in Table **3**, we calculate performance indicators according to the e data of Table **4** such as accuracy, recall rate and precision rate, the results are shown in Table **5**.

According to the data in Table **5**, the following conclusions can be drawn:

(1) The performances of three the experiments are almost the same in accuracy, recall rate, and precision rate, which shows that the introduction of the algorithm of the decision tree based on attribute sets dependence have no effect on the performances of Bayesian spam filtering.

(2) The experiments of literature [8] have shown that there are no more than 20% mails that are sentenced as suspicious mails, if the mails have been handled by the algorithm of the decision tree based on attribute sets dependence. The time for classifying suspicious mails artificially and the

time of the classification algorithm for decision tree based on attribute sets dependence must be less than the time for complete artificial classification. Therefore, the algorithm was proposed not only for reducing the cost of mail leaning and classification, but also it improves the efficiency of the system.

CONCLUSION

Spam filtering has become a global hot issue, and the mechanism of Bayesian spam filtering is currently the most widely used method. However, a large number of mails sets need to be trained prior, which consumes a lot of system resources and network resources which affect the executing efficiency of the system. The mechanism of Bayesian spam filtering based on MapReduce that was proposed in this paper can handle huge amounts of data, shorten the process of mail classification, and improve and the efficiency of the system. The traditional mechanism of Bayesian spam filtering needs to classify mails manually in the pre-training phase of mail set, which not only wastes a lot of time but also consumes a lot of human and financial resources. This paper presents the algorithm of the decision tree based on the attribute sets' dependence in the pre-training phase of mail set in order to sort out the normal mails, spam and suspicious mails, then to classify suspicious mails manually. It solves the problem of traditional filtering mechanisms. The experiments proved that the proposed algorithm kept good expression in performance of mail filtering, meanwhile, reducing the cost of the mail classification and increasing the efficiency of the system.

According to the proposed algorithm, there are still almost 20% mails that require manual classification in the pre-training phase of mail set, so the future aim is to achieve automatic mail classification under the premise of system performance so that it enhances the executing efficiency of the system.

CONFLICT OF INTEREST

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

ACKNOWLEDGEMENTS

Supported by The National Natural Science Fund (11271057, 51176016) and the project of general office of Broadcasting and Television (GD10101) and Natural Science Fund in JiangSu (BK2009535) and Natural Science Fund in ZheJiang (Y1100314) and Jiangsu Province ordinary university innovative research project (CXZZ13_0691).

REFERENCES

[1] M. Sahami, S. Dumais, and D. Heckerman, "A Bayesian Approach to Filtering Junk E-Mail", *Papers of AAAI Workshop on Learning for Text Categorization*, 1998, pp. 55-62.

[2] J. Dean, and S Ghemawat, "MapReduce: simplified data processing on large clusters", *Comm. ACM*, vol. 51, no.1, pp. 107-113, 2008.

[3] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: Distributed data-parallel programs from sequential building blocks", In: *Proc. 2nd Eur. Conf. Com. Syst., (EuroSys)*, 2007, pp. 59-72.

[4] Z. Pawlak, and A. Skowron "Rudiments of rough sets", *Inform. Sci.*, vol. 177, no. 1, pp. 3-27, 2007.

[5] Z. Pawlak, "Rough set theory and its applications to data analysis", *Cybern. Syst.*, vol. 29, pp. 661-688, 1988.

[6] I. Rish, "An empirical study of the Naïve Bayesian classifier", In: *Proc. IJCAI Workshop Empir. Methods Artif. Intell., [S.1.]*: IJCAI, 2001 pp. 41-46.

[7] China Education and Research Network [EB/OL]. [2008-07-17]. http://www.edu.cn/ji_shu_ju_le_bu_1640/20080717/t20080717_310253.shtml

[8] W. B. Deng, and Z. Hong, "Two-stage spam filtering method based on rough sets", *Comput. Appl.*, vol. 30, no. 8, pp. 2006-2009, 2010.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014