The study on the spam filtering technology based on Bayesian algorithm

WANG Chunping

Mathematics and Computer Science College, Xinyu University
Jiangxi, Xinyu 338000, China

Abstract

This paper analyzed spam filtering technology, carried out a detailed study of Naive Bayes algorithm, and proposed the improved Naive Bayesian mail filtering technology. Improvement can be seen in text selection as well as feature extraction. The general Bayesian text classification algorithm mostly takes information gain and cross-entropy algorithm in feature selection. Through the principle of Bayesian analysis, it was found that the characteristics distribution is closely related to the ability of the feature representing class, so this paper proposes a new feature selection method based on class conditional distribution algorithm. Finally, the experiments show that the proposed algorithm can effectively filter spam.

Keywords: *Naive Bayes, minimum risk Bayesian, active learning Bayesian, feature selection, email filtering*

1 Introduction

With the rapid development of Internet, interaction between people become more convenient, e-mail, with its quick and low-cost features, gradually become an important tool for interaction. People use it to exchange ideas, transfer files, and express their views, so it has become an indispensable communication tool in daily life. But at the same time it also brings some negative effects and a large part of the mail we receive each day are unsolicited. Some of them are commercials, some political propaganda, some pornographic advertising, there are even viruses. These are what we commonly known as spams.

The economic loss caused by spam to Internet users is quite staggering: According to statistics, only download Internet access fees and phone charges and other expenses a year will cost $ 94 of the world's Internet users. As the sender of the spam, the price is low, usually through mass email in a variety of ways. For e-mail service providers and users, spam gave them a lot of damages and losses, and the losses caused by pornography, computer viruses, and load fraudulent letter are inestimable. Despite some disputes on Bayesians philosophical view, it is undoubted that the ideas and methods are widely used in the social life and production practice. In particular, in recent years, the Bayesian approach, with its uncertain knowledge for its unique form of expression, the probability of rich expressive power and the priori knowledge incremental learning characteristics become the focus of many ways the most compelling data mining.

In 1996, Rvennie set up ifile, a machine learning applications for email filtering system based on Bayesian algorithm, which can use Bayesian algorithm to categorize messages. In the process of establishing ifile system, Rennie noted that each user has a different set of e-mail, and different way to organize messages, thus allowing the user to manually adjust the false positive mail. In 1998, Sahami, in using Bayesian algorithm to filter messages, noted that spam has unique properties different from the legitimate mail: For example, in the class of to get rich quickly spam, in addition to text messages like “free” and “money”, there will be a large number of stressed symbols like "!" and the representative symbol "$". Using Naive Bayes algorithm to filter mail, Sahamiliy hand-joined the domain information for these specific tasks phrase as well as spam features to filter, improving the accuracy of filtering spam; in addition, he is also using a characterization loss rate threshold to reduce false positives of legitimate mail. In 2001, Matthew and others developed a spam filter MEF. MEF can filter out virus e-mail whose attachments have the executable program in UNIX. The mail filter first decodes the binary code of the executable program, compares it with the existing binary code of the virus, uses Naive Bayes algorithm to calculate the probability that it belongs to spam, and makes decisions accordingly.

2. Naive Bayesian spam filtering basic theory

2.1 Naive Bayesian principle

Bayesian approach is an important method of spam filtering, the essence of the method is to identify messages as junk mail or regular mail, which is a classification problem.

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

Suppose that there are m sample spaces $\{c_1, c_2, ..., c_n\}$, and the mail d has n feature items $(w_1, w_2, ..., w_n)$. The probability of d belonging to the class c_k for a given class c_k (k = 1,2, ..., m) is

$$p(c_k \mid d) = Max\{p(c_1 \mid d), p(c_2 \mid d), ..., p(c_n \mid d)\}$$

By Bayesian probability formula we can get:

$$p(c_k \mid d) = \frac{p(d \mid c_k) p(c_k)}{p(d)} \quad \text{(k=1,2,…, m)}$$

In which:

$$p(d \mid c_k) = p(w_1, w_2, ..., w_n \mid c_k)$$

The denominator p (d) in formula (3) has nothing to do with the class, so it can be ignored when comparing maximum value in the equation (3). So we only need to calculate the probability $p(c_k)$ and $p(d \mid c_k)$ to categorize mail d.

In equation (4), $p(c_k)$ is a priori probability and easy to calculate, but the calculation of $p(d \mid c_k)$ is more difficult, particularly when the number of feature items is large and the dependence between the feature item is high, so the calculation would take a lot of time. In order to simplify the calculation, we introduced the conditional probability independence assumption, that is assuming that each feature items are independent of each other—the naive Bayes filter, then the formula (2-5) can be converted to:

$$p(d \mid c_k) = p(w_1, w_2, ..., w_n \mid c_k) = \prod_{i=1}^{n} p(w_i \mid c_k)$$

Naive Bayesian filter structure is shown in the following figure:

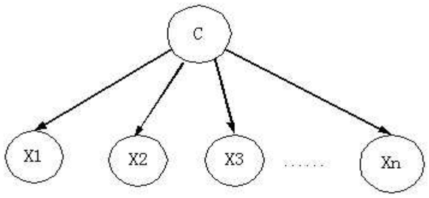

Figure 1 Naive Bayesian filter structure

Naive Bayesian filter uses priori probability to obtain the posterior probability, sets filters according to the training sample, and classifies emails according to the posterior probability of the message text.

2.2 Naive Bayesian mail filtering technology

Literature used Naive Bayes algorithm to design spam filtering system SpamCop [2]. The system is able to identify about 92% of spams with a fault correction rate of 1.16%. SpamCop system made improvements in keyword selection principle, including ignoring spaces, continuous sequence of letters and numbers, as well as getting rid continuous sequence of characters less than three characters apart from above mentioned characters. And the system used m in the calculation of the probability of the feature classes to estimate, and used the following formula:

$$P(Token / Ham) = \frac{N(Token, Ham) + \frac{1}{K}}{N(Ham) + 1}$$

$$P(Token \mid Spam) = \frac{N(Token, Spam) + \frac{1}{K}}{N(Spam) + 1}$$

$$P(Spam \mid Token) = \frac{P(Token \mid Spam)}{P(Token \mid Han) + P(Token \mid Spam)}$$

In these formulas, $P(Spam \mid Token)$ represents the posterior probability of feature Token belonging to the garbage category. $N(Token, Spam)$ represents the number of occurrences of the keyword in the spam, $N(Token, ham)$ represents occurrences keywords in the normal circumstances. $N(Ham)$ is the number of normal mail, and $N(Spam)$ the number of spam. K stands for the number of different keywords in the mail, solving the problem of zero possibility.

Literature provides an effective Bayesian spam filtering method [3]. The filter captures 99.5% of spam with a fault correction rate of less than 0.03%. The filter sets up two hash tables for spam and normal mail to calculate the occurrence of keywords of corresponding Corpus. To calculate the probability of each keyword, we use the following formula:

$$p(W \mid C) = \frac{b / nbad}{2 * g / ngood + b / nbad}$$

In the formula, b represents the number of occurrences of keywords in the spam, g represents the the number of occurrences of the keywords in the regular mail, nbad the total number of spam, ngood the total number of normal mail. A factor 2 in the denominator is a recommended empirical value, used to reduce the probability of normal mail as spam. In the calculation of the joint probability filter uses the following formula:

$$p(d_x \mid c_{spam}) = \frac{p_1 p_2 \cdots p_n}{p_1 p_2 \cdots p_n + (1 - p_1)(1 - p_2) \ldots (1 - p_n)}$$

Wherein p_i (i = 1,2, ..., N) represents the probability of the I keyword being calculated.

Data sparseness problem is often encountered when

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

using the formula (1-5) a mail to calculate spam probability, that is, if the message contains a new feature, no matter how high the possibility of this message containing other features items, it will cause zero conditional probability. This is not to ignore the problem. In literature [3], it is given a zero probability smoothing formula, a better solution of zero probability problems. The formula is as follows:

$$f(w) = \frac{a * x + (n * p(w))}{a + n}$$

In this formula, a is an adjustable constant, and n is spam and normal mail aggregate that contains the characteristics of w, x is the initial probability, and when n = 0, f (w) equals to the initial probability, as n increases , f (w) becomes more and more close to the p (w). Based on experience, the initial probability is generally set as 0.52, a is 0.0178. Literature [4] improved formula 1-5 and provided new method in calculating Keywords joint probability.

$$P = 1 - \sqrt[n]{(1-p_1)*(1-p_2)*...*(1-p_n)}$$

$$Q = 1 - \sqrt[n]{p_1 * p_2 * ... * p_n}$$

$$S = \frac{P - Q}{P + Q}$$

S value is between -1 and 1. The high value means spam and low means regular mail. 0 means between the two.

Most of the filter built using naive Bayes got improved on the basis of Bayes formula, and did not take in to account the difference between mail filtering and ordinary text classification. On the other hand, these traditional Bayesian methods are based on the minimum error rate of the decision-making methods, not taking into account different characteristics between the legitimate mail and spam, which is that legitimate e-mail misidentified as spam may give users a greater loss. In addition, traditional Bayesian learning algorithm used given the training sample to learn classification parameters. The training samples it dealt with must be with a category label, and be randomly selected with passive acceptance of these samples. In this paper, the Bayesian spam filtering process is studied, the improved method of feature selection process is proposed with two naive Bayes extension models: minimum risk Bayes and active learning Bayes.

4 Improved Naive Bayesian filter design

To better use Bayesian algorithm in spam filtering, this paper improved Bayesian algorithm in the following aspects:

(1) Text representation
In ordinary text classification Bayes algorithm, the text was represented by a word or phrase. Words and phrases are the smallest unit that can represent semantics. In spam, in order to avoid being filtered the spammers use variants of junk words instead of junk words.

(2) Feature selection
The ordinary Bayesian text classification algorithm feature selection mostly take the information gain, and expected cross entropy algorithm. Through analysis of Bayesian principle, it was found that the characteristics distribution is closely related to the ability of the feature representing class, therefore a new feature selection method based on class conditional distribution algorithm is proposed.

4.1 Text representation

In text classification, usually we usually use the vector model (VSM) to represent text, which can be represented as an n-dimensional vector $(t_1, t_{2,} t_3, ..., t_n)$, in which $t_i(i = 1, . . . , n)$ represents the weight of the i-th feature items.

The feature item is usually defined as a series of consecutive character string separated by spacebar, tabs or various punctuation marks and accents in the English text. Under normal circumstances, the feature item is a meaningful word or phrase. In character handling, all uppercase letters are converted to lowercase. All spacebar, tabs, line breaks, and various punctuation marks and accents are removed.

In Chinese, text feature item is a character, word, phrase, or some kind of concept. In the Chinese text, they mainly refer to vocabulary after word processing. But in comparison of several spam messages similarity, we found that block phrases appear more often in similar spam. And the spammers now in order to avoid being filtered, often use vocabulary variants to prevent being filtered, so, in the ever-changing spam variants, the simple word characteristics can no longer meet the requirements.

Fingerprint is applied in the comparison of similar mails. When we compare two mails, two mails can be divided into a number of blocks of text (actually a sub-string), if two mails are similar, they must contain a lot of same text blocks. And the comparison operation between these texts blocks is accurate comparison, therefore can be to be optimized with hashing method.

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

4.2 Feature Selection

Feature selection is an important area of research in text categorization, and its purpose is to select several important features representative of the text, and the text category in a training text. The most important issue in feature selection is the relationship between the characteristics and the class, that is, the features selected are truly representative.

The common feature selection methods are expected cross entropy, information gain method, mutual information method, chi-square test method, principal component analysis method and so on. These methods, from the information theory and from statistical analysis, find out the salient features containing the largest amount of information or influence, while ignoring the rest of the features, to achieve the purpose of feature reduction.

4.2.1 Information gain

Information gain is often used as a method to select the best node in the decision tree technology. It uses the concept of entropy in information theory. In information theory, entropy is a measure of the kind of things that is uncertain. It is based on the individual characteristics values to designate the learning sample spaces, depending on how much of the information gain to select effective feature. The information gain of feature t_k is as follows:

$$IG(t) = -\sum_{i=1}^{n} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{n} p(c_i \mid t)$$

$$\log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{n} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t})$$

$p(c_i)$ is the probability of category c_i in the text; p (t) is probability of features in the text; $p(c_i \mid t)$ represents when t appears in the documentation set, the possibility document belonging to c_i ; $p(c_i \mid \overline{t})$ represents when t does not appear in the documentation set, the probability of the document belonging to c_i .

Whether features appear in the text, they will provide text classification information, to calculate the size of the conditional probability in the different cases the amount of information provided. Information gain use the characteristic values to divide the training sample space, and select features according to the amount of information. During feature selection, we select those characteristics with large information gain. The feature selection method has a problem, that is if a feature appears in the class C_1, but does not appear in the class C_2, this feature is very important in itself, but after summing the values of each log phase offset, the result is 0, and certain words cannot be distinguished. There are two ways to solve this problem: First, take the absolute value of the log value, second, omit the log value that is less than 0. In addition, the method is more complicated.

4.2.2 Expect cross entropy

The only difference with information gain is that, cross entropy does not consider the conditions when characteristics are not happen. The expected cross-entropy of characteristic t is as follows:

$$ETC(t) = p(t)\sum_{i=1}^{n} p(c_i \mid t) * \log_2 \frac{p(c_i \mid t)}{p(c_i)}$$

Expect cross-entropy reflects the probability distribution of the categories of text, as well as the distance between the text class probability distribution in the case of certain characteristic words. During feature selection, we select the characteristics with high cross entropy.

4.2.3 Mutual Information

The mutual information is a feature correlation criterion often used in the field of machine learning, which represents the correlation between the two vectors. Mutual information in characterized t and class c is defined as follows:

$$MI(t,c) = \log_2 p(t \mid c) - \log_2 p(t) = \log_2 \frac{p(t \mid c)}{p(t)}$$

4.2.4 Choose based on the characteristics of the class conditional distribution

This paper is from the essence of the Bayesian classification proposed feature reduction method based on class conditional distribution. Main idea of Bayesian classification is to calculate joint probability of the characteristic class probability, so we base on class possibility in feature reduction in algorithm for Bayesian probabilities. For A_i with number of categories l, and the number the value v, can be expressed with the following matrix of its class conditional probability distribution:

$$(p_{kj})_{l*v} = (p(a_{ij} \mid c_k))_{l*v} = [\begin{matrix} p(a_{i1} \mid c_1), & p(a_{i2} \mid c_1) & \dots & p(a_{iv} \mid c_1) \\ p(a_{i1} \mid c_2), & p(a_{i2} \mid c_2) & \dots & p(a_{iv} \mid c_2) \\ \dots & \dots & \dots & \dots \\ p(a_{i1} \mid c_l) & p(a_{i2} \mid c_l & \dots & p(a_{iv} \mid c_l) \end{matrix}]$$

For this two types of classification problems of spam, you can use the following matrix to represent the class

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

conditional probability distribution characteristics of t_i :

$$p(t_i \mid c_k)_{2*1} = \begin{bmatrix} p(t_i \mid c_{ham}) \\ p(t_i \mid c_{spam}) \end{bmatrix}$$

If distribution of the characteristics of the ham and spam is approximately uniform, then it has not so large influence on calculation, and can be ignored. Such a uniform distribution can make data distribution entropy analysis large. Based on the analysis of IrinaRish on the Naive Bayesian performance data characteristics [8], Naive Bayes can obtain better accuracy low entropy distribution data. Therefore, get rid of these characteristics that enlarge data distribution entropy, we can improve the performance of the Naive Bayes classification. We have taken the following formula as the evaluation of characteristics of the class distribution:

$$CCD(t_i) = \frac{1}{2}\sum_{k=1}^{2}(p_k - p_l)^2$$

In this $p_l = \frac{1}{2}\sum_{k=1}^{2} p_k$

We can see from the formula p_i , in which p_i represents the arithmetic mean of t_i in regular mail and spam. $CCD(t_i)$ represents the distance of representative feature t_i away from the arithmetic average probability. So the larger the value, the farther away t_i from the average probability, indicating that the more uneven it is in the category distribution, the smaller the distribution entropy is. Conversely, the smaller the value of $CCD(t_i)$, the smaller the probability indicated from the average probability distance is, and it indicates that the more uniform distribution of the characteristic in each category. And the higher the distribution entropy is. The purpose of the feature selection is to select a low-entropy distribution of the data, i.e. CCD larger values of characteristics.

5 experiments and analysis

In order to compare the three feature selection methods' influence on classification accuracy, we use three feature selection filters in both offline and online filtering mode to filter. Online filtering mode process on trec07 p mail and the online takes immediate feedback mode and offline mode on the sewm 2008.

In feature selection criteria, the number of key features extracted also has a certain impact on filtering accuracy, so the number of features is 8, 10, and 15, and compare difference of the three feature selection method in mail filtering accuracy. First compare trec07 p online timely feedback experiments, Spam tools of TREC evaluation,experimental results are as follows:

Table 1 Online filtering accuracy of three algorithms when features selected number is 8

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	2.39(1.76-3.16)	1.18 (1.05-1.32)	1.42 (0.95-2.05)
Spam%	1.10 (0.88-1.35)	0.87 (0.79-0.96)	0.16 (0.09-0.28)
Lam%	1.62 (1.37 - 1.91)	1.01 (0.95 - 1.08)	0.48 (0.34 - 0.69)
1-ROCA%	0.3509 (0.1166 - 1.0512)	0.3728 (0.1699 – 1.8844)	0.2963 (0.1453 – 1. 4129)

Table 2 Online filtering accuracy of three algorithms when features selected number is 10

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	1.16 (1.13-1.42)	1.20 (1.03-1.29)	1.32(1.05-1.00)

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

Spam%	0.86 (0.75-0.90)	0.86 (0.79-0.95)	0.14(0.05-0.26)
Lam%	1.02 (0.94 - 1.11)	1.00 (0.93 -1.07)	0.39 (0.35 - 0.44)
1-ROCA%	0.2986(0.2673 -0.3444)	0.2739 (0.2563 – 0.3346)	0.1363 (0.1053 – 0. 3129)

Table 3 Online filtering accuracy of three algorithms when features selected number is 15

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	2.15(1.56-3.11)	1.17(1.04-1.33)	1.40 (0.97-2.00)
Spam%	0.96(0.78-1.24)	0.88(0.78-0.95)	0.17 (0.11-0.27)
Lam%	1.32(1.27 - 1.85)	1.00 (0.96 - 1.07)	0.47(0.33 - 0.67)
1-ROCA%	0.3423(0.1166- 1.0512)	0.3546 (0.1574 – 1.7633)	0.2567(0.1343 – 1. 3879)

Make Offline filtering on the publicly available data sets in sewm 2008 and take the first 30,000 as training and the last 40,000 to test. The experimental results are in following table:

Table 4 Offline filtering accuracy of three algorithms when features selected number is 8

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	0.88 (0.24-1.87)	0.85 (0.20-1.83)	0.58 (0.10-1.43)
Spam%	8.89 (7.24-10.61)	8.90 (7.35-10.78)	5.45 (4.53-7.46)
Lam%	6.93 (5.23 - 9.76)	6.89 (5.12 - 9.38)	3.20(2.13-4.42)
1-ROCA%	1.2646 (1.1898 - 1.3688)	1.1978 (1.1832 - 1.2957)	0.7853(0.5346-0.9843)

Table 5 Offline filtering accuracy of three algorithms when features selected number is 10

Table 5 Offline filtering accuracy of three algorithms when features selected number is 10 Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	0.68(0.13-0.87)	0.64(0.10-0.83)	0.50(0.09-1.35)
Spam%	6.89(5.35-8.87)	6.48(5.12-8.53)	5.21(4.41-7.12)
Lam%	4.79(3.34-5.89)	4.45(3.23-5.76)	2.43(1.54-4.78)

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

1-ROCA%	1.0420(0.6735-1.0241)	0.9879(0.5345-1.0001)	0.5474(0.3214-0.8634)

Table 6 Offline filtering accuracy of three algorithms when features selected number is 10

Evaluation parameters \ feature selection methods	information gain	expected cross entropy	class conditional distribution (ccd value)
Ham%	0.83 (0.21-1.85)	0.85 (0.20-1.83)	0.56(0.08-1.33)
Spam%	8.49 (6.98-9.61)	8.90(7.35-10.78)	5.24(4.32-7.23)
Lam%	6.53 (5.13 - 9.51)	6.89 (5.12 - 9.38)	3.01(2.34-4.32)
1-ROCA%	1.1646 (0.9856-1.5423)	1.1978(1.1832-1.2957)	0.7633(0.5157-0.9621)

The experimental results show that, whether in immediate feedback online or offline mode, with the same mail filtering algorithm, if the number of feature selected is different, spam filtering accuracy is different. When the number of feature selection is 10, whether in information gain, expected cross entropy, or class conditional distribution of feature selection algorithm, mail filtering accuracy is better in the feature selection number 8 and 15 of the algorithm. This indicates that the larger number of feature selection is not the better; More feature selection not only increase the difficulty of the calculation, but also bring some features with low class representation and not clear category. Text content between each word was not completely independent, and the premise of the Naive Bayesian method is assuming features are independent of each other, so when the \the number of eigenvalue extracted increase, the opportunity of interdependence between eigenvalues increases. But with too few selected features, making the classification only consider one-sided to the characteristics of the part, we ignored many on the classification of impact characteristics, resulting in classification accuracy decreased.

It can also be seen from Table 1 to Table 6, our feature selection method selection methods based on the same number of features, classification accuracy of conditional distribution was significantly higher than that based on information gain and expected cross entropy-based feature selection method. Specifically, in legitimate messages missing rate and spam missing rate, on the ROC curve above the area of these three parameters, cross entropy of information gain and expectations are higher than the class of conditional distributions. This also shows that in the Naive Bayes classification, the characteristics with a high amount of information may not contribute to the classification of the characteristics, and the characteristics with uniform class conditional distribution are more significant factors in classification accuracy.

6 Conclusions

This chapter first introduces the classification process of Naive Bayes algorithm, and proposes for its classification process improvements in the text representation fingerprint features in four aspects. It also proposes new joint probability formula and solves probability problems in probability calculations. In feature selection, it creates new feature selection method: feature selection based on class conditional distribution and in classification stage raises the weight integrated classification model. And based on that learning process is deepening, this paper also proposes adaptive algorithm to adjust the threshold.

7 References

[1] Jian Zhu, Hongbing Cao, Haitao Liu, "Parking Space Detection Based on Information from Images and Magnetic Sensors", AISS, Vol. 4, No. 5, pp. 208 ~ 216, 2012

[2] hijuan Deng, Shaojun Zhong, "A Kind of Text Classification Design on the Basis of Natural Language Processing", IJACT, Vol. 5, No. 1, pp. 668 ~ 677, 2013

[3] JianJiao Chen, Anping Song, Wu Zhang, "Hybrid K-harmonic Clustering Approach for High Dimensional Gene Expression Data", JCIT, Vol. 7, No. 3, pp. 39 ~ 49, 2012

[4] Yanhua Tan, Changsheng Zhang, Jing Ruan, "The Comparative Study of Different Models for Feature Selection in Rough Set Theory ", IJACT, Vol. 4, No. 4, pp. 124 ~ 130, 2012

[5] Hao-dong Zhu, Hong-chan Li, Jin-Chao Zhao, "Feature Selection Based on Feature Distinguish Ability And Meta-Information", IJACT, Vol. 4, No. 11, pp. 344 ~ 351, 2012

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.

[6] Jinchao Zhao, Fan Zhang, "Feature Selection Based on Parallel Collaborative Evolutionary Genetic Algorithm", AISS, Vol. 4, No. 6, pp. 296 ~ 304, 2012

[7] Ammar ALmomani, Tat-Chee Wan, Ahmad Manasrah, Altyeb Altaher, Eman Almomani, , "A survey of Learning Based Techniques of Phishing Email Filtering", JDCTA, Vol. 6, No. 18, pp. 119 ~ 129, 2012

[8] ZHANG Qiu-yu, YANG Hui-juan, WANG Peng, MA Wei, "Fuzzy Clustering based on Semantic Body and its Application in Chinese Spam Filtering", JDCTA, Vol. 5, No. 4, pp. 1 ~ 11, 2011

[9] Liu Pei-yu, Zhao Jing, Zhu Zhen-fang, "Email Representation using Noncharacteristic Information and its Application", JCIT, Vol. 5, No. 8, pp. 180 ~ 185, 2010

[10] Anjan Kumar Pau, "Robust Object Classification and Recognition for Video Surveillance Applications", IJIIP, Vol. 3, No. 1, pp. 79 ~ 89, 2012

Copyright (c) 2013 International Journal of Computer Science Issues. All Rights Reserved.