International Journal of u- and e- Service, Science and Technology
Vol.8, No.2 (2015), pp.235-242
http://dx.doi.org/10.14257/ijunesst.2015.8.2.23

Algorithm of E-mail Classification Based on Automatic Adapting for User

Zhongjian Wang[1, a], Zongjie Wang[2, b] Yanfeng Gao[1] and Yanfen Lin[1]

[1]Harbin University of Commerce, Harbin 150028, China
[2]Open Fund of Smart Education and Information Engineering
Harbin Normal University, Harbin, 150025, China
[a]wangzhj@hrbcu.edu.cn, [b]jie052003@163.com

Abstract

E-mail classification is an effective method to manage, improve process efficiency and filter junk mail. The extraction of E-mail characteristic is the key problem of exactness classification. In order to make the classification has a more distinct division characteristic words, IDF (Inverse document frequency) is used to epurate further the characteristic. The procedure which users deal with E-mail is a natural half-supervised learning. By using this process, proposed algorithm corrects classification results, adjust classification rule to adapt the individuation requirement of user automatically. The evaluation experiments indicate the availability of proposed algorithm.

Keywords: *E-mail classification, IDF, half-supervised learning, classification rule, automatic adapting user*

1. Introduction

With the developing and popularity of Internet, the user of E-mail is also increasing quickly. The E-mail has become an absolutely necessary communication and between persons intercommunion tool. According news report, the number of Chinese mobile internet user has increased to 7.5 hundred million [1]. The development of Internet has changed people's life style. Generally user of E-mail spends some time to process large amount E-mail. With the wide use of email, spam issues are becoming increasingly serious. How to filter spam effectively and manage E-mail conveniently has become a problem must to be solved at present.

The E-mail classification is an effective method to manage, improve process efficiency and filter junk mail. Commonly mails user agent software has simple classification function [2], it can carry out simple classification of E-mail and spam filter by setting E-mail address or key words. Because the contents of the E-mail are very extensive, the simple classification function of E-mail agent software is not enough to meet the demand of users. The research of E-mail classification and junk mail filter is very significant, and experts have been paid attention to the research of different method [3-4].

Generally the research of E-mail classification use machine learning method, use a vector space model to represent an E-mail, classify E-mail by calculating distance between an E-mail vector with the characteristic vector [5-6]. This method expresses an E-mail by a vector that consists of high frequency words in an E-mail, which can only calculate the vector instead of the E-mail's content. The use of VSM model makes the calculation and operation of the E-mail possible. But VSM does not consider the structure of E-mails which affects the correctness of classification. An E-mail is composed by five parts of 'Header', 'Greeting', 'Body', 'Closing' and 'Signature' currently. For instance, an E-mail header contain 'to:', 'cc:', 'subject:' and 'attach'. At least, that information maybe is useful to classify E-mails.

ISSN: 2005-4246 IJUNESST

There are many different researches about E-mail classification. Paper [7] use maximum entropy model to classify E-mail. They have discussed the pre-process of E-mail, features extraction of E-mail header and the number of features and the times of iteration. The research results indicated the classification performance that utilizes all E-mail fields in classifying is satisfaction.

Because Naive Bayesian Algorithm is based on the independent assumption, it assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature, given the class variable. Paper [8] proposed a classification algorithm which is fit Chinese E-mails, risk minimization Bayes based on hybrid model. The model unifies binary independence model and multinomial model, improved the recall of E-mail filter and the precision.

The winnow algorithm is a technique of a linear classifier from labeled examples. Paper [9] presents the design of an E-mail classifier based on Winnow algorithm and focus on Chinese E-mail classification to establish classifying rules.

Paper [10] presents a personalized framework of E-mail classification based on agent, a user model which describes user's relatively stable classification request within a period of time and E-mail content model which describes an E-mail by vector space model. The research focus on personalizing classification of E-mails based on personalizing center vector, but the difference of classification performance between the personalizing center vector and general feature vector is not mentioned. So-called personalizing center vector is a vector that is calculated by arithmetic average method according to a class of E-mail, it is hard to say the personalizing center vector can better reflect classification requirement of user than the general classification feature vector.

Paper [11] proposed a new method which combines both Mitra’s and Sequential Forward Selection by analyzing and comparing the several typical feature selection methods for E-mail classification. Experimental result shows that the proposed method can improve the precision of E-mail classification.

E-mails are different with general text; they are half-structure text. The information of the head of E-mail is useful to process spam, but only the head information of E-mail cannot classify E-mails accurately. To get effective and personalizing classification results of E-mails, it is necessary that use the content of E-mail. So that, we present an E-mail classification method based on inductive learning. The method takes use of E-mail subject and body to extract features of classification, acquire classification rules of adapting to users by Inductive Learning. The method can satisfy different classification demand of different user; carry out E-mail classification of adapting users though supervised learning.

2. Outline of the Method

The method is a half-supervised learning procedure. It includes rules generation of E-mail address, rules generation of E-mail body, E-mail classification processing, judgment of classification results, feedback processing and renew processing of rules, the flow chart as Figure 1 shows.

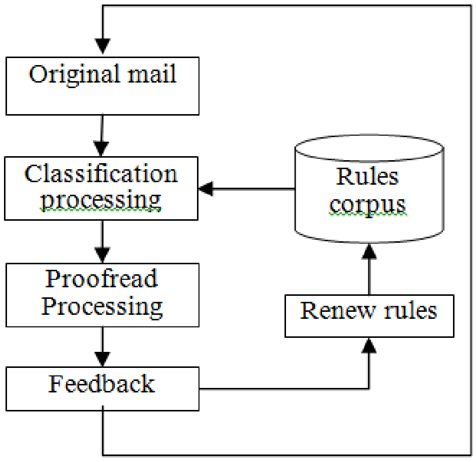

Figure 1. Out Line of the Method

The purpose of E-mail classification is to gather the E-mails that have same characters, make for management of E-mail. The procedure of classification is illuminated as follow.

2.1 Information Processing of E-mail head

E-mail messages consist of two major sections of mail head and mail body.

A mail header: structured into fields such as From, To, CC, Subject, Date, and other information about the email.

A mail body: the basic content, as unstructured text; sometimes contains a signature block at the end. This is exactly the same as the body of a regular letter. Each E-mail has exactly one header. The header is separated from the body by a blank line.

At first, 'From' and 'Subject' are used to judge the primal classification of E-mail. 'From' is the send address.

The general format of an email address is like this: donfgha@yahoo.cn.com. It consists of two parts: the part before the @ sign is the local-part of the address, often the username of the recipient (donfgha), and the part after the @ sign is a domain name to which the email message will be sent (yahoo.cn.com).

There are all sorts of domain name, Such as international domain names and internal domain names. According to the domain name of a mail, the organization of sending or receiving mail can be identified. Commonly international domain name '.com', '.net', '.gov' and '.edu' denote attribute of every organization. Example '.com' denotes business organization. '.net', '.gov' and '.edu' denote network service organization, government organization and education organization respectively. Internal domain name consists of two parts of international domain and internal domain, postfix internal domain name. Such as 'cn', 'jp' and 'uk' are China, Japan and United Kingdom respectively.

We use domain name carry out initial stages classification.

2.2 Subject and Body Processing of E-mail

The subject of a mail is inspissation of E-mail content. Generally the subject of a mail consists of keywords of a mail, expresses the keystone of a mail. By reading the subject of a mail, the content of mail is understood almost. So the subject contains key words that express content of a mail. We gather the subject of mails, establish classification rules as subject of mails, and pay high authority to those rules from mail subject.

For mail body, at first the classification system tidies up the text of mail body. In this processing, DBC case characters are deleted in the text of mail body. All punctuations are changed SBC case, and all non-text garbage is eliminated.

The second, word segmentation processing is carried out for the mail body text. The method of word segmentation is called Inductive Learning Word Segmentation. The word segmentation method not use word segmentation dictionary, uses inductive learning rules. The method considers character string that appears repeatedly in text would be word or phrase. Word segmentation method predicts unknown words using Inductive Learning recursively. According to extracting condition we classified prediction word candidates into different level ranks and registered it in the dictionary. There are four ranks, A, B, C and D, belonged to word in the dictionary. The rank of a word which has been proved its correctness is A. The predicted word candidates' rank can only be B, C or D. We consider that the high rank and long length of a word candidate has a higher priority. Then we can segment the text according to their rank and their length. The method segments text of mails body to words or phrases according to the appearance frequency of character string in text [12].

After the text are segmented to words or phrases, appearance frequency of words and phrases are calculated, and stop words in the segmentation result are deleted by using for reference stop word list. The stop word list is established by manual.

After the text are segmented to words or phrases, appearance frequency of words and phrases are calculated, and stop words in the segmentation result are deleted by using for reference stop word list. The stop word list is established by manual.

At last, we get a list of words (phrases) and their appearance frequency. In the list, words are arranged according to the appearance frequency.

2.3 Classification Rules

The classification method of E-mails is based classification rules. The classification rules are established by E-mail head, E-mail subject and E-mail body. Here are three kinds' classification rules of mail classification.

A. The classification rule from E-mail head:

According to the domain name some kinds of classification are summarized. Here are E-mail classification rules of initial stages: business, network service, government and education. If the address of mail includes the domain name of relevant rules, then the mails are classified to that sort. If there are not relevant classification rule, the classification of mails are not carried out.

For example, according to the sent address of a mail xxxx@yyyy.zzzz.net, a rule of classification is got as follow: if sent address of mail includes "net", then it is classified to network service organization.

The classification by using mail head is only primary and approximate classification.

B. The Classification Rule from E-mail Subject:

The words in subject have very important meaning generally. They are key words of a mail and they should have high authority on classification processing. Such as a mail subject: call paper, SCI/EI, and then the mail can be classified to conference classification. For example if a mail subject contains call paper, SCI/EI, Invitation, workshop, then a rule of classification is got as like: if subject of a mail includes "SCI/EI", "paper call", "invitation", "workshop", then it is classified to conference.

C. The Classification Rule from E-mail Body:

After text of mail body is segmented to words, the appearance frequency of words is calculated. The words that have low appearance frequency are deleted. The words that have high appearance frequency are classified to several classification clusters. Each classification cluster is consisted of five to ten words that have a high appearance frequency.

2.4 Classification Processing

The classification processing of E-mail is carried out by follow steps:

Step one; E-mails are classified by the rules that abstracted from mail head. This classification is simply and approximately. If there are not usable rules, this step does nothing.

Step two; for the classification results of the last step, the further classification is carried out by the rules that got from mail subject. This classification is based on the classification of first step, is further classification. The rule of mail subject is consisting of few words, but these words have very important authority on mail classification generally. These words compose not only the mail subject, but also appearance in mail body text commonly. After this classification, the classification will be adjusted by the rule from mail body text.

Step three; the rules from mail body text are used on the classification. Through words segmentation, appearance frequency calculation and stop words deletion, we select the words of high appearance frequency and get a words gather. About the number of clusters, it decided by manual according to the data of feature words from the all experimental mails.

There is a pre-definition E-mail classification set:

$$MC = \{MC_1, MC_2, MC_3, \cdots MC_i \cdots MC_n\}$$

n is the number of classification, MC_i is the feature word set of the ith classification.

$$MC_i = \{Mc_{i1}, Mc_{i2}, \cdots Mc_{ix}, \cdots Mc_{im}\}$$

m is the number of feature words. The words gather from an E-mail is :

$$W = \{w_1, w_2, \cdots w_j, \cdots w_y\}$$

w_j is the jth feature word. y is the number of feature words in a mail. The W is consist of substantive words, deleted those empty words that there is no meaning.

A mail belongs to the classification MC_i when:

$$C_i = \arg\max\{\sum_{k=1}^{Z} w_k \cdot \alpha_k \cdot Mc_{ik}\}$$

$$Z = \max\{m, y\}$$

Here, α_k is the coefficient of authority, denotes importance of a feature word. The value of the coefficient of authority is $0 \leq \alpha_k \leq 1$. At beginning, the value of the coefficient of authority is appearance frequency of normalization.

A. Proofread and Feedback Processing

The procedure of proofread is carried out by reclassifying that mail of erroneous classification are adjusted to the group of correct classification. This process procedure includes the proofread and the feedback processing. A mail is adjusted to the correct group, that is to say, the classification of mail was erroneous. The reclassification is proofread processing, it is also the feedback processing.

B. The renew of Classification Rules

The mail classification system renews the classification rules after the proofread processing and the feedback processing. The renewing processing involves reclassification, correction of the affiliated classification group of feature words and adjusting some coefficient of feature words.

After the proofread and the feedback processing, the system updates the value of authority coefficient of feature words according to the feedback information. Those values of authority coefficient for correct classification are increased in order to emphasizing the importance of some feature words. And otherwise for the feature words of erroneous classification, the value of authority coefficient is decreased in order to lessening the effect of feature words.

C. The Initial Value of Authority Coefficient:

$$\alpha_i = fr_i \Big/ \sum_{i=1}^{m} fr_i$$

m is the number of feature words in a classification.

fr_i is the appearance frequency of ith feature word.

$\alpha_j = 2 \times \alpha_j^o, \alpha_k = \alpha_k^o \div 2$ are the coefficient of the correct classification and the erroneous classification respectively. After updating processing, the coefficients are normalized:

$$\alpha_i = \alpha_i / \sum_{i=1}^{m} \alpha_i$$

3. Classification Experiments

We carry out the evaluation experiments by data of CCERT Data Sets of Chinese Emails [13]. CCERT Data Sets of June and July 2005 are made up of 63710 emails, including 18314 legitimate emails and 45396 spam messages. We select 3000 mail from CCERT Data Sets, it includes mails of different classification, and the spam is a kind of classification.

3.1 Classification Experiments

The number of classification is decided by manual. We calculate the appearance frequency of words, take high frequency words as feature words and gained the classification set.

The training experiments are completed by tenth of 3000 mails. Other 2700 mails are used for the evaluation experiment.

The part of the classification experiment results is listed in Table 1.

As the Table 1 shows, we use eight class of mail. They are news, sports, tour, training information, conference information, hotel information, restaurant information and spam.

In Table 1, the classification results are the results of three times experiment. After each experiment, the proofread and the feedback processing are carried out; the experiment results have a small change because the authority coefficients are adjusted. Table 2 lists the precision change of three times experiment results.

Table 1. The Results of Mail Classification

	Conference information	…	News	sports	tour	other
1	385	…	389	421	339	358
2	383	…	320	315	343	362
3	379	…	326	336	346	374

Table 2. The Precision Change of Experiment Result

Experiment times	first	second	third
Precision[%]	78.5	84.6	91.9

The last experimental results list in Table 3. The correct classification and misclassification are the average of three times experimental results.

Table 3. The Average Experiment Results

Correct classification	Misclassification	Total number
2185	391	2700

3.2 Evaluation of Experimental Results

The formula (1) and (2) are used in the evaluation of the experiment.

$$precision[\%] = {}^{MNCC}\!/\!_{TNCM} \times 100 \qquad (1)$$

$$recal[\%] = {}^{MNCC}\!/\!_{TNM} \times 100 \qquad (2)$$

MNCC, TNCM and TNM are the mail number of correct classification, the mail number of classification and the total number of mail respectively.

According to Table 2, the precision and the recall are 85% and 98.4% respectively.

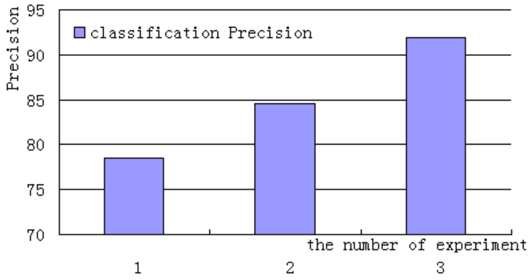

Figure 2. The Change of Classification Precision

Figure 2 shows the change of classification precision, with the increasing of number that mails are processed in experiment, the coefficient of authority is adjusted, and the precision of classification is also improved gradually.

3.3 Discussion

The mail classification results show the non-classification mails 1.6% and the misclassification mail 15%. By analysis and review of experiment results, the reasons of misclassification have three points mainly. Some mails have different subject with contents; some mails have few words and because the ambiguity of the feature words result in there is little distinction between different classifications.

About the mail non-classification, the reasons are almost the problem of processing and there are not enough of classification rule. About the mail non-classification, the reasons are almost the problem of processing and the number of classification rules not enough.

To evaluate the validity of proposed method, we select 300 mails from the experiment results randomly. The results of three times experiment indicate, with the feedback processing the precision would be improved gradually. And with the increasing of the number of processing mails, the precision would be improved as Table 2 and Figure 2 shows.

4. Conclusion

We have proposed a method of mail classification and carried out evaluation experiment. The method makes use of the characteristics of e-mail that must be read by user. The procedure of reading mail equates feedback process of half-supervised learning. The method that uses half-unsupervised learning is suitable for habit that people use mail. When users read mail, move the misclassification mail to the fit classification unconsciously. By the mail reclassification users accomplish the proofread and the feedback processing, not need other additional work.

The proposed method classifies the E-mail to automatic adapting users with increasing of the number of processing E-mail. At last, the method becomes a customization method of user.

Acknowledgements

This work was supported in part by the Open Fund of Smart Education and Information Engineering, Harbin Normal University, Natural Science Foundation of Heilongjiang province (F201243) and the Scientific Research Foundation of Education Bureau of Heilongjiang Province (12511127).

References

[1]. Mobile internet will increase rapidly (in Chinese). 2013.1.8, **(2013).**

[2]. http://zlzq.p5w.net/zqdetail.asp?id=3632.

[3]. “Foxmail”, http://fox.foxmail.com.cn/.

[4]. H. Zhang, W. Zhang and C. Wu, *etc.*, “IDSS-based E-mail Filtering”, Journal of Wuhan University (Natural Science Edition), v01. 29, no. 12, **(2004)** December, pp. 1115-1118.

[5]. L. Liao and D. Wen, “Email Classification Based on the Glue of Content”, Computer simulation. vol. 25, no. 2, **(2008)** February, pp.121-123.

[6]. Z. Wang and J. Wei, “Spam Filter Approach Based on Support Vector Machine”, Computer Engineering, vol. 35, no. 13, **(2009)**, pp. 188-189.

[7]. L. Chen and Z. Liu, “A Spam Filtering System Based on Vector Space Model for Outlook”, Computer Applications and Software, vol. 22, no. 12, **(2005)** December.

[8]. J. Li, P. Li and Q. Zhu, *etc.*, “Email Categorization with maximum entropy model”, Computer Engineering and Applications, vol. 43, no. 35, **(2007)**, pp. 126-129.

[9]. J. Xu, I. Yang and T. Huang, “On Mail Classification Technology Based on Bayesian”, Science Technology and Engineering, vol. 8 no. 7, **(2008)**, April, pp. 1875-1878.

[10]. Z. Qiaoming, Z. ZhiJun and L. Peifeng, “Design of the Chinese Mail Classifier Based on Winniw”, Journal of Nanjing University. vol. 41, **(2005)** October, pp. 807-812.

[11]. K. Qiu, Q. Guo and X. Zhang, “The Research of Personalized Classification E-mail System Based on Agent”, Computer Engineering and Application, vol. 30, no. 7, **(2005)** July, pp. 176-178.

[12]. C. Chaolan and Z. Zili, “Feature Selection in E-mail Classification”, Computer Science, vol. 33, no. 2, **(2006)** January, pp. 73-75.

[13]. Z. Wang, K. Araki and K. Tochinai, “Word Segmentation Method Based on Inductive Learning and Segmentation Rules”, International Symposium on Computational Intelligence and Design (2008 ISCID), **(2008)** October 17-18, Wuhan, China.

[14]. “CERNET Computer Emergency Response Team (CCERT)”, http://www.ccert.edu.cn/spam/sa/ datasets. Htm.

Author

Zhongjian Wang, Ph.D., professor. His main research interests include natural language process, Chinese sentence paraphrase, Chinese word segmentation and Information retrieval, *etc.*