

An Approach to Email Categorization with the ME Model

Peifeng LI, Jinhui LI, Qiaoming ZHU

School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China, 215006
{pfli, jhli, qmzhu@suda.edu.cn}

Abstract

This paper puts forward a hierarchical approach for categorizing emails with the ME model based on its contents and properties. This approach categorizes emails in a two-phase way. First, it divides emails into two sets: legitimate set and Spam set; then it categorizes emails in two different sets with different feature selection methods. In addition, the pre-processing, the construction of features and the ME model suitable for the email categorization are also described in building the categorizer. Experimental results show that the hierarchical approach is more efficient than the previous approach and the feature selection is an important factor that affects the precision of email categorization.

Introduction

With the popularization of Internet, the email has become one of the most popular methods for people to communicate each other. Though the email gave us such timely convenience, it also caused the trouble of processing omnifarious emails. Classifying those emails into categories is a convenient and efficient way for people to read them. Email categorization (also called as email classification) was oriented from text categorization and it assigns new emails to pre-defined categories automatically based on their contents and properties.

A variety of approaches towards email categorization have been put forward in the past few years. Popular approaches to email categorization include RIPPER (Cohen, 1996; Provost, 1999), Rough Set (Li, et al., 2004), Rocchio (Yang, et al., 2002; Yang, et al., 2003), Naïve Bayes (Yang, et al., 2003; Bekkerman, et al., 2004), SVM (Bekkerman, et al., 2004), Winnow (Zhu, et al., 2005), Neural network (Clark, et al., 2003), etc. Those work proposed some useful approaches to email categorization. Nevertheless, most of the above approaches were oriented from text categorization, so those approaches classify emails using the plain text categorization approach, regardless of the differences between texts and emails. However, an email is a semi-structure text which includes a structure in the email head and it redounds to email categorization. Besides, the most popular approach used in email categorization is Bayes. This approach is poor on the precision of categorization though it is suitable for the requirement of the rapidity and dynamics in email categorization. Otherwise, mostly the SVM approach can get the highest

precision in text categorization, but it also doesn't satisfy the requirement of rapidity and dynamics in email categorization because training the categorization model is expensive on time cost.

Therefore, this paper introduces the Me model (Berger, et al., 1996) into email categorization and puts forward an hierarchical approach which categorizes the email based on its contents and properties, such as "subject", "sender", "receiver", etc. This paper also discusses other techniques to improve the performance of categorizer, such as email pre-processing, features selection, iteration, etc.

To Pre-process the Email

After having analyzed the structure of an email, we divide it into two parts: contents and properties. Contents include the email body and the subject which constitute the main part of an email. Properties include those fields such as "From", "Cc", "To", "Date", "SMTP server", "Attached files", etc. The content part is mostly like a plain text while the property part is the characteristic of the email.

To Pre-process the Email Contents

The purpose of pre-processing email contents is to delete unused texts and to standardize them. The email format is different from plain text, and the additional pre-processing for email categorization is described as below:

(1) To convert the native encoding to Unicode

There are lots of encoding schemas to encode characters. For example, ASCII and ISO 8859 are two popular encoding schemas to encode the phonetic characters. So it is necessary to unify the encoding schema of the characters. Otherwise, a same word in different emails would be regarded as different words due to their different encoding schema, especially for ideographic characters.

In this step, a convert tool [Li, 2005] is provided to recognize the encoding schema and then to convert texts to Unicode, the international standard of character encoding schema.

(2) To filter the HTML tag in the email body

Generally the email body has two styles: plain text and html. If the email is in the html format, it should be converted to plain text because most html markups would confuse the categorizer except "<a href>" and "".

In this step, we firstly record the links (<a href>) and images () which is embedded in the html markups, and then delete all html markups and convert the html file to a plain text file.

(3) To filter the non-character symbol

In many emails, especially in the spam emails, there are many non-character symbols, such as “☺”, “▶”, “♪”, etc. Those symbols themselves are useless for categorizer so it's necessary to delete those symbols. In other way, non-character symbols usually are the characteristic of spam emails, so the number of the non-character symbols also should be recorded as an important feature of the email for the categorizer.

(4) To unify the format of digitals

There are many digital formats, such as currency, date, time, telephone number, etc. each one also has many different styles. For example, the telephone number “812-2345678” also has many other styles, such as “(812) 234 5678”, “812-234-5678”, “812 2345678”, “812 234 5678”, etc. Therefore, we have defined a unified style for each digital format and all other format digitals should be transferred to that unified style.

(5) To resolve the links

The links in the email body may also have many styles, and they need to be converted to a standard style. Otherwise, the relative link in the email also should be converted to the absolute link.

(6) To revert the intersected words

To prevent spam flitting tools from labeling them as spams, many spam emails are inserted some marks between words or characters. Those marks must be deleted before starting the categorization. For example, the string “M*A*I*L” must be reverted to “MAIL”.

(7) To Segment the Words

East Asian languages, such as Chinese and Japanese, a sentence is written as a string of characters without separations between words. As a result, before any text analysis on Chinese and Japanese, a word segmentation task has to be completed so that each word is “isolated” by the word-boundary information. In additional, after the word segmentation, words in the subject should add a tag to divide them from words in body.

(8) To delete the stop words

Stop words mainly include auxiliary words, auxiliary verb, adverb and conjunction. Those words are helpless for categorization, so it should be deleted as well.

To Pre-process the Email Header

The email header consists of some important information, such as the name of sender/receiver, the email server, the IP address, etc. Those data are valuable for email categorization, especially for email filtering, so the fields of “From”, “To”, “Cc”, “Date”, “Content-Type”, etc should be

extracted as features for the categorizer. For example, the content of the “From” is “Jason Lee <yoyo@aclweb.org>”, then the features could be extracted as follows: (in xml style)

```
<Sender>
  <SenderName>Jason Lee</SenderName>
  <SenderID>yoyo</SenderID>
  < DomainName>aclweb</DomainName >
  < DomainType>org</DomainType >
</sender>
```

An Email Categorization Model

The ME Model

The ME model is one mature statistics model, and it is suitable for email categorization. The basic theory of ME is that we should prefer to uniform models that also satisfy any given constraints, which are mined from the known event collection.

As for the ME model applied to text or email categorization, Zhang, et al., 2003) provided an approach of filtering spam emails based on a ME model and Li (Li, et al., 2005) applied the ME model to classify the text. And there are no researches in public which are concerned with how to apply the ME model to categorize the emails.

In email categorization, each email is deemed as an event. For example, there is an event collection which is presented as $\{(e_1, c_1), (e_2, c_2), (e_3, c_3), \dots, (e_N, c_N)\}$, where $e_i (1 \leq i \leq N)$ denotes an email and $c_i (1 \leq i \leq N)$ is the category of document e_i . To obtain the constraints from the event set, a feature function was introduced into ME model. The feature function in email categorization could be built on the features and categories of the emails. For the feature w and the category c' , its feature function is as follow:

$$f_{w,c'}(e, c) = \begin{cases} 1 & c = c' \text{ \& } e \text{ contains } w \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The goal of email categorization is to obtain the best probability restricted by features, and the probability is defined as:

$$p_\lambda(c|e) = \frac{1}{Z_\lambda(e)} \exp\left(\sum_i \lambda_i f_i(e, c)\right) \quad (2)$$

where

$$Z_\lambda(e) = \sum_c \exp\left(\sum_i \lambda_i f_i(e, c)\right) \quad (3)$$

is simply a normalizing factor determined by the requirement of that $\sum_c p_\lambda(c|e) = 1$ for each document e ,

f_i is the feature function, λ_i is the weight assigned to feature f_i . Two algorithms specifically tailored to calculate the parameters of a ME classifier are Generalized Iterative Scaling Algorithm (GIS) and Improved Iterative Scaling

Algorithm (IIS).

From the experiments, we found out that the performance of only using binary valued feature as formula (1) is inferior. So we optimize it and use word-frequency and word-position weight as feature's value. The new feature is defined as:

$$f_{w,c}(e,c) = \begin{cases} \sum_{i=1}^3 f_{q_i}(w) * \lambda_i & c = c' \text{ \& } e \text{ contains } w \\ 0 & \textit{otherwise} \end{cases} \quad (4)$$

where $f_{q_i}(w)$ ($1 \leq i \leq 3$) are called word-frequency weight and their values are related to the freq. of word w in body, subject and header of email e . λ_i ($1 \leq i \leq 3$) is called word-position weight and its value is defined as 1, 1.5 and 2 according to the experiments.

To Extract the Features from the Emails

How to extract features from the email is very important for the categorizer. After the pre-processing, an email is expressed as a set of features actually. A categorizer should learn those features from the train set and then form the feature set for each category.

(1) In the train set, there are k pre-defined categories, noted as c_i ($1 \leq i \leq k$) and in each category c_i there are a set of emails as $\{e_{i,1}, e_{i,2}, e_{i,3}, \dots, e_{i,m}\}$ while each email $e_{i,j}$ consists of a set of features:

$$Fe_{i,j} = \{f_{i,j,1}, f_{i,j,2}, f_{i,j,3}, \dots, f_{i,j,n}\}$$

So the feature set of category c_i is defined as:

$$cfe_i = fe_{i,1} \cup fe_{i,2} \cup fe_{i,3} \cup \dots \cup fe_{i,m} = \bigcup_{j=1}^m fe_{i,j} \quad (5)$$

(2) Then for each cfe_i , to delete the features which occurred in other cfe_j ($i < j$):

$$\begin{aligned} cfe_i &= cfe_i - cfe_i \cap (cfe_1 \cup cfe_2 \cup \dots \cup cfe_{i-1} \cup cfe_{i+1} \cup \dots \cup cfe_k) \\ &= cfe_i - cfe_i \cap \left(\bigcup_{1 \leq j \leq k, j \neq i} cfe_j \right) \end{aligned} \quad (6)$$

(3) The experimental results showed that chi-square statistics was the best approach to extract feature in email categorization. So for each feature $fe_{i,j}$ in category c_i , to calculate the chi-square statistic value between $fe_{i,j}$ and c_i :

$$\chi^2(f_{i,j}, c_i) = \frac{N[P(f_{i,j}, c_i) * P(\bar{f}_{i,j}, \bar{c}_i) - P(f_{i,j}, \bar{c}_i) * P(\bar{f}_{i,j}, c_i)]^2}{P(f_{i,j}) * P(\bar{f}_{i,j}) * P(c_i) * P(\bar{c}_i)} \quad (7)$$

where

$P(f_{i,j}, c_i)$ is the probability of $fe_{i,j}$ and c_i co-occur, $P(\bar{f}_{i,j}, \bar{c}_i)$ is the probability of neither $fe_{i,j}$ or c_i occur, $P(f_{i,j}, \bar{c}_i)$ is the probability of neither $fe_{i,j}$ occurs with out c_i , $P(\bar{f}_{i,j}, c_i)$ is the probability of neither c_i occurs without $fe_{i,j}$.

(4) To sort all of the features on the chi-square statistic values, the result of category c_i is as follow:

$$fe_i = \{f_{i,1}, f_{i,2}, f_{i,3}, \dots, f_{i,l}\} \quad (8)$$

while

$$\chi^2(f_{i,j}, c_i) \geq \chi^2(f_{i,j+1}, c_i) \quad (1 \leq j \leq l-1)$$

Then delete all $fe_{i,j}$ while $j > 2000$. This means that each category only reserves 2000 features.

Experiments and Analysis

The Email Corpus

Currently there are some public email corpora, such as Ling-spam, PU1, PU123A, Enron Corpus [Klimt, et al., 2004], etc. But all of above corpora couldn't be used to test our approach, because most of them are mainly used to filter spam emails and only have two categories: legitimate emails and spam emails. Besides, Enron Corpus has many categories, but it just categorizes the email by users, not by contents. Therefore, to test our approach, we have to build an email corpus which includes 11907 category-defined emails with 7 categories. And the categories are {Work & Study, Auto-reply, Private Contents, Advertisements, Invoice and Tax, Porn and Adult, Train and Lecture}. The first 3 categories are legitimate emails sets (Legi) while the others are spam emails sets (Spam). Otherwise, all emails in our corpus must satisfy one restriction: the number of words in the email body must be greater than 10. We choose 1/3 emails randomly from each category as test set, and the rest regards as train set. The test set is showed as table 1. We provide four combinations to extract the features from the email. Those combinations are as follows:

- SB: Subject + Body
- B: Body
- HS: Header + Subject
- HSB: Header + Subject + Body

Table 1 Test set

Categories	num	Categories	num
work & study	2217	advertisements	848
auto-reply	82	invoice and tax	316
private contents	217	Porn and adult	69
		train and lecture	220

In our experiments we report recall (r) and precision (p) for each category, and micro-average p (or simply micro-p) for each text categorization approach. These measures are defined as follows:

$$r(c) = \frac{\text{the number of emails correctly assigned to class } c}{\text{the number of emails contained in class } c} \quad (9)$$

$$p(c) = \frac{\text{the number of emails correctly assigned to class } c}{\text{the number of emails assigned to class } c} \quad (10)$$

$$\text{micro-p} = \frac{\text{the number of correctly assigned test emails}}{\text{the number of test emails}} \quad (11)$$

Experiments on the iteration and feature number

Firstly, we experiment our approach on the test set with different iterations from 50 to 550. The results are showed as figure 1.

From the figure 1, we found out that every micro-p

increases rapidly as the increasing of the number of iterations from 50 to 250. But when the iteration number is greater than 250, each line goes steady and becomes a horizontal line. The greater the number of iteration is, the higher the time cost is. So we choose 250 as the iteration in our approach.

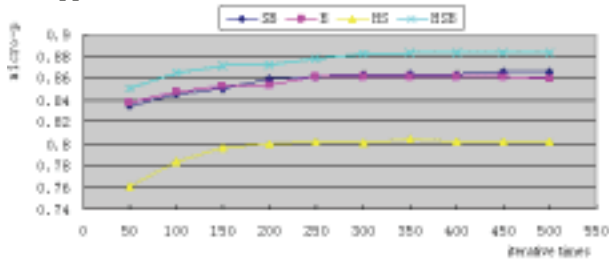


Figure 1 the relation between iteration and micro-p (feature number: 2000)

We also tested our approach with different numbers of features from 500 to 5500. The results are showed as figure 2.

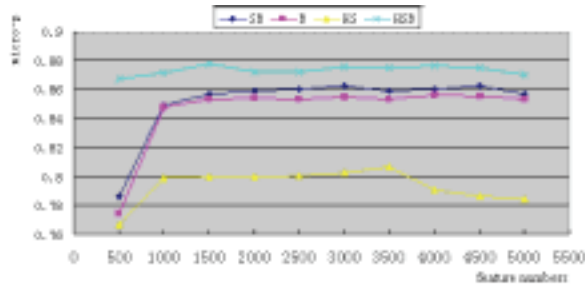


Figure 2 the relation between feature number and micro-p
In figure 2, we found out that the micro-p increases rapidly as the increasing of the feature number from 50 to 1500. And when the feature number increased from 3500 to 5500, the micro-p didn't increase actually. On the contrary, it decreased with the increasing of the feature number sometime. So we choose 2000 as the feature number at last.

Experiment on different feature combinations

We also tested our categorizer on four different feature combinations mentioned above. The results are showed in table 2.

Table 2 The number of emails incorrectly categorized with four feature combinations respectively

Combinations	HS	B	SB	HSB
Number of emails incorrectly categorized in Legi	111	214	142	145
Number of emails correctly categorized in Spam	327	189	171	141
Total number of emails incorrectly categorized	438	403	310	286
Micro-p	0.8896	0.8985	0.9219	0.9279

In table 2, we found out that HSB was the best of all four combinations for it only had 286 emails incorrectly categorized. But for legitimate set, HS was the better one than the others and only had 111 emails incorrectly categorized.

From that result, we proposed a two-phrase categorization. This approach firstly divides the emails into two sets: Legi and Spam while using HSB as features, and then categorizes the emails in two different sets respectively on different feature combinations while using HS in Legi and HSB in Spam. We named this approach as hierarchical categorization approach while the original approach was called as direct categorization.

Experiments on hierarchical categorization

We also tested the hierarchical categorization approach on our test set, and the results are showed as table 3 and 4.

Table 3 The recall, precision and micro-p after the first categorization

Categories	Recall	Precision	Micro-p
Legi	0.9759	0.9807	0.9725
Spam	0.9665	0.9585	

Table 4 The recall, precision and micro-p after the second categorization in Legi set and spam set respectively

Categories	Recall	Precision
work & study	0.9838	0.9528
auto-reply	0.8095	0.8947
private contents	0.4561	0.9630
advertisements	0.9220	0.8935
invoice and tax	0.9905	0.9811
Porn and adult	0.7826	0.7826
train and lecture	0.8631	0.8872
micro-p	0.9346	

In the table 3, it showed the recall, precision and micro-p after the first categorization. In this step, the categorizer divided the test set into two sets: Legi and Spam, so it mostly liked a spam email filter, but a categorizer. The micro-p of the first categorization is 97.25%. This result also showed ME model utilizing word-frequency and word-position weight is an efficient way to filter the spam emails.

In the table 4, it shows the recall, precision of all categories and the micro-p after the second categorization in Legi set and spam set respectively. Hereinto, the categorizer classifies the Legi set based on HS while it classifies the spam set based on HSB. After two phrases, the final micro-p is 93.46% in all.

From table 2, 3, and 4, we find out that hierarchical categorization is better than direct categorization. It achieved an improvement of micro-p by 0.67% over the direct categorization. But hierarchical categorization is

more complex than direct categorization, because it must categorize the test set twice while direct categorization only need once.

To Compare with Other Approaches

To compare our approach with other popular approaches, we also tested the Bayes, SVM, KNN and Winnow approaches on our email corpus. In those experiments, we extracted the features from email body and subject by utilizing chi-square statistics (those experiments also showed that chi-square statistics is better than Information Gain, Mutual Information, etc in email categorization) and the feature number is also selected as 2000. The result is showed as figure 3.

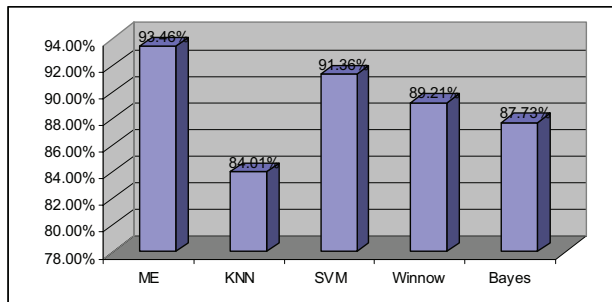


Figure 3 the micro-p of ME, KNN, SVM, Winnow and Bayes

In common, SVM is the best approach to categorize the text, but it needs too much time to train the model and is not suitable for email categorization because it is time costly sometimes. But in our experiment, the results show that ME approach is the best one to classify the emails, and it achieves an improvement of micro-p by 2.1% over the SVM approach. We believe that the pre-processing, appropriate feature selection method and the hierarchical categorization approach are the main factors for ME to beat the SVM in our experiments.

Experiments analysis

Based on above experiments, we also can find out that:

- (1) To extract the features from all fields of the email is the best way. Except the body, other fields also can provide useful information for categorizer to improve the performance.
- (2) The hierarchical categorization is better than direct categorization, but it also is more complex and time-consuming than that one.
- (3) Usually, the features extracted from HSB are the best combination for email categorization, but for legitimate email, the HS is the best choice in our experiments. The reason is that: the subject of a legitimate email often can abstract the content while the spam email always give a fake subject in order to cheat the filtering program.
- (4) In our email corpus, the emails in the category “Work

& Study” and “Private Contents” are easy to confuse because of their content also can’t be distinguished by people. So the recall of category “Private Contents” is very low in table 4 and many emails in such category are assigned to category “Work & Study”.

Conclusion

This paper puts forward a hierarchical email categorization approach based on ME model and also discusses the pre-process, the feature selection and the iteration. We have implemented a categorizer which based on such a model and the categorizer is a plug-in component for the Microsoft Outlook. Our categorizer is used by many users and the survey result shows that it works well. Our future work mainly focuses on optimizing the ME model and adding machine learning algorithm to learn users’ actions to adjust the model. Otherwise, our approach is really poor to classify that email when its body only has few words, especially it’s empty. So we also plan to research on those emails and try to find a method to classify them correctly.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments on this paper and also would like to thank Dr. Rui Wang, who gave me good advice to improve the presentation of this paper. This research was supported by the National Natural Science Foundation of China under Grant No.60673041 and the High Technology Plan of Jiangsu Province, China under Grant No.2005020.

Reference

- Bekkerman R., McCallum A. and Huang G. 2004. *Automatic categorization of email into folders: benchmark experiments on Enron and SRI corpora*. CIIR Technical Report IR418.
- Berger A., Pietra S. and Pietra V. 1996. *A maximum entropy approach to natural language processing*. Computational Linguistics, Vol. 22 No. 1, pp.38-73.
- Clark J., Koprinska I. and Poon J. 2003. *LINGER – a smart personal assistant for e-mail classification*. Proceeding Of the 13th International Conference on Artificial Neural Networks, pp. 274-277.
- Cohen W. 1996. *Learning rules that classify e-mail*. Proceeding of AAAI Spring Symposium on Machine Learning and Information Retrieval, pp.18-25.
- Klimt B. and Yang Y. 2004. *The Enron Corpus: A new dataset for email classification research*. Proceeding of ECML’04, 15th European Conference on Machine Learning, pp.217-226.
- Li P., Zhu Q., Qian P. 2005. *Research of Han character internal codes recognition algorithm in the multi-lingual environment*. Journal of Chinese Information Processing,

- Vol. 18, No. 2, pp.73-79.
- Li R., Wang J., Chen X., et al. 2005. *Using Maximum Entropy model for chinese text categorization*. Journal of Computer Research and Development, Vol. 42 No. 1, pp.94-101.
- Li Z., Wang G., Wu Y. 2004. *An E-mail classification system based on Rough Set*. Computer Science, Vol. 31, No. 3, pp.58-60, 66.
- Provost J. 1999. *Naïve-bayes vs. rule-learning in classification of email*. Technical Report AITR-99-284, University of Texas at Austin, Artificial Intelligence Lab.
- Yang J., Chalasani V. and Park S. 2003. *Intelligent email categorization based on textual information and metadata*. IEICE Transactions on Information and Systems, pp.1280-1288
- Yang J. and Park S. 2002. *Email categorization using fast machine learning algorithms*. Proceeding of the Fifth International Conference on Discovery Science, pp.316-323.
- Zhang L. and Yao T. 2003. *Filtering junk mail with a maximum entropy model*. Proceeding of 20th International Conference on Computer Processing of Oriental Languages, pp.446-453.
- Zhu Q., Zhou Z., Li P. 2005. *Design of the Chinese mail classifier based on Winnow*. Acta Electronica Sinica. Vol. 33 No. 12A, pp.2481-2482.