Filter

Ayahiko Niimi, Hirofumi Inomata, Masaki Miyamoto and Osamu Konishi
School of Systems Information Science, Future University-Hakodate
116–2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041–8655 Japan
email: niimi@fun.ac.jp

***Abstract*— In this paper, the system that classified spam mail and other mail(regular mail) was constructed by two filters with Bayesian theory and SVM(Support Vector Machine) used well by the text classification task as a text classification algorithm. It was confirmed to evaluate the performance of the spam filter constructed by Bayesian theory and SVM, and to show a high reproduction ratio and a high relevance ratio. Moreover, the URL pre-fetch method was built into Bayesian spam mail filter, and the relevance ratio was able to be improved.**

I. Introduction

Recently, the use of mail service has become popular because the Internet has become popular. The spam mail problem becomes a serious problem along with these popular mail services. The spam mail is a trouble mail that sent to many persons, and the mail so on by one-sided advertising mail, the chain mail, the fictitious claim mail, and included computer virus spread by mail. The spam mail becomes a problem because an increase in the network traffic occurs because other mail not only is buried by a large amount of spam mail but also a large amount of mail flows on the network. Therefore, there is a possibility to exert the influence also in other Internet services. The mechanism that only necessary E-mail is automatically taken out of a large amount of mail including the spam mail is needed because of the spam mail measures.

Because the content of mail is basically described by the text it can be said that task of classifying mail into spam mail and other mail is text classification task. Therefore, various text classification algorithms can be applied for the mail classification task. Especially, spam mail and other E-mail (we define them as regular mail) are thought to be a classification task to two classes with positive examples and negative examples.

In this paper, the system that classified spam mail and other mail was constructed by two filters with Bayesian theory and SVM(Support Vector Machine) used well by the text classification task as a text classification algorithm.

II. Mail Classification Task

Some mail filters which separate spam mail and other mail are proposed. The following are typical mail filters to the spam mail.

1) Basic text filter
2) Whielist filter
3) Blacklist filter

1 is a method of filter with rule sets which are easy text strings based on the mail that has been received, and classifying mail based on the rule sets. For instance, the rule such as “If the mail Subject header contains ’Save Your Money’, it is spam mail.” is made, and mail is classified by this rule. Generally, it is necessary to register such rules by the hand work, and there are two problems. One is that the rule can make from the spam mail that has been received, and the other is that to make rules takes much time.

2 is a method of filter with mail address list which describes the mail address where the reception is permitted, and mails from other addresses do not received. There is a system in which permitted mail address is registered by not only mail recipient but also mail transmitter. In these system, if a mail from address which is not resisted in permitted mail address list is received, the system send mail to request registration to list, and the automatic registration in the addressee list by the mail address with the response. There are two problems that it takes costs to make the addressee list and that possibility of mis-detection that the system is filtered regular mail as the spam mail is high.

3 is a method of filter with server (or mail address) list that doesn’t permit receiving, and mails from other mail server (or mail address) receive. 2 is listing the permitted mail address, but this method is listing the not permitted mail server or mail address. the not permitted mail address list can be shared, because the spam mail address and the mail server which delivers the spam mail generally is common, though a possibility that the permitted mail address is different in each user is high. The spam mail is overlooked and the possibility that the filter doesn’t operate efficiently is high though the possibility of overlooking regular mail lowers in this method.

These filters are the methods of extracting the feature of spam mail and regular mail by the hand work. On the other hand, the method of automatically extracting the feature of mail is thought. Because mail contents are described by the text, the mail classification task can be captured with one of the text classification tasks. Spam mail and other E-mail (regular mail) are thought as negative examples and positive examples, this task is thought to be a classification problem of separating into two classes. Therefore, the automatic text classification algorithms can be used for the mail classification task.

Some of automatic text classification algorithms have already been proposed. [**?**], [1], [2] It is thought enough to use these algorithms to construct the spam filter.

A. Morpheme Analysis

A morphological analysis is to divide the input sentence into the morpheme which is a minimum unit with the meaning in linguistics, to decide the part of speech of each morpheme, and to allocate the prototype to the morpheme to which the transformation of the word of use. [10]

A morphological analysis is important for Japanese documents, because Japanese sentence is not divide words by blank. In English, a morphological analysis is used to analyze end of a word transformation (tense, single or plural), suffix, prefix, etc.

For instance, it is analyzed that the morphological analysis is done by the sentence "Happyoukai wo okonaitai." (This sentence means "I want to hold a symposium"). (Refer to table I)

TABLE I
EXAMPLES OF MORPHOLOGICAL ANALYSIS

Happyou	Happyou:	Noun
kai	Kai:	Noun
wo	Wo:	particle
okonai	Okonau:	verb-independent
tai	Tai:	auxiliary verb
.	.	symbol-period

The word divided by the morphological analysis is called an element-term. It comes to be able to do the frequency analysis and filtering to a specific part of speech by dividing into the element-term.

III. BAYESIAN SPAM FILTER

Bayesian spam filter is a spam filter which is based on Bayesian theory. [4] In Bayesian theory, the probability of a certain cause when a certain event occurs can be calculated by the probability of all cause of event and the conditional probability that the event occurs by a certain cause. The filter separates by the probability whether the spam mail or not from the appearance probability of the character string (token) used with mail based on Bayesian theory. The word (or, the stem) and the character string that n character are consecutive are used as a token.

When a token (w) is included, the probability that the mail is spam mail (spam probability: $p(w)$) is defined by the following expressions.

$$p(w) = \frac{b/n_{bad}}{\alpha g/n_{good} + b/n_{bad}} \tag{1}$$

In this expression, we defined these symbols.

- $p(w)$: When a token (w) is included, the probability that the mail is spam mail (spam probability)
- n_{bad}: number of spam mail
- $b(w)$: appearance time of a token (w) in spam mail
- n_{good}: number of regular mail
- $g(w)$: appearance time of a token (w) in regular mail
- α: weight

In this definition, the mis-detection rate of the spam mail is decreased by applying weight to the number of regular mail.

Moreover, the probability that mail is spam mail when two or more tokens were contained at the same time (composite probability) was defined as follows.

$$P(w_1, w_2, \dots w_n) = \frac{p(w_1) \times p(w_2) \times \cdots \times p(w_n)}{p(w_1) \times \cdots \times p(w_n) + (1 - p(w_1)) \cdots (1 - p(w_n))} \tag{2}$$

In this expression, we defined these symbols.

- $P(w_1, w_2, \dots w_n)$: the probability that mail is spam mail when some tokens ($w_1, w_2 \dots w_n$) are contained at the same time (composite probability)
- $p(w_1)$: spam probability with a token(w_1)

The Bayesian filter procedure for classifying whether a mail is spam mail or regular mail is as follows. The procedure separates into pre-processing (filter learning) and the classify processing (filtering).

- pre-processing(filter learning):
  1) Collect spam mails and regular mails.
  2) Separate all mail to tokens.
  3) Calculate spam probability of each token.
  4) Store spam probability in database.
- classify processing(filtering):
  1) Separate a mail to tokens.
  2) Query spam probability of the token.
  3) Extract characteristic tokens, calculate composite probability.
  4) If the composite probability is higher than establish threshold, this mail is classified as spam mail.
  5) If the composite probability is smaller than establish threshold, this mail is classified as regular mail.

The tokens which are good for classification process as the characteristic tokens. We use the tokens which spam probability is spreaded from 0.5. Spam probability 0.5 means that this token can be classified neither spam nor regular.

The flowchart of Bayesian spam filter is shown in Figure **??**.

IV. SVM SPAM FILTER

SVM(Support Vector Machine) is an algorithm to classify the data set expressed by the vector into two classes. [5] In the spam filter by SVM, mail is classified into spam mail and regular mail by using SVM.

SVM uses the data set expressed by the vector as an input. To classify mail by SVM, the mail data necessary expressed a vector. The vector conversion of the text is used by separating into the tokens same as Bayesian spam filter, and calculating the appearance frequency with the token code corresponding to appeared token. To define the token code, all tokens that appear to mail are extracted beforehand. As for the appearance

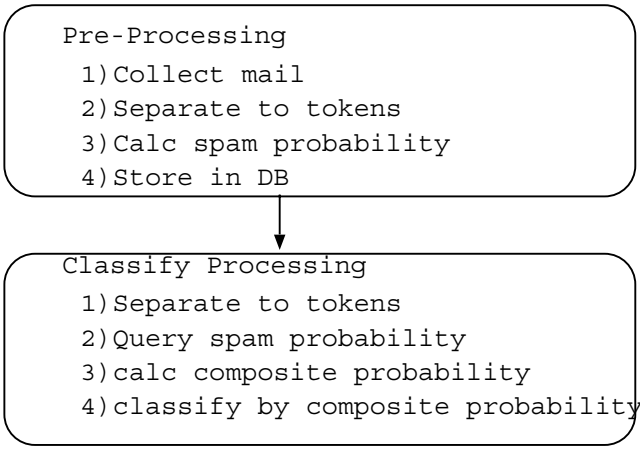

Fig. 1. Flowchart of Bayesian Spam Filter

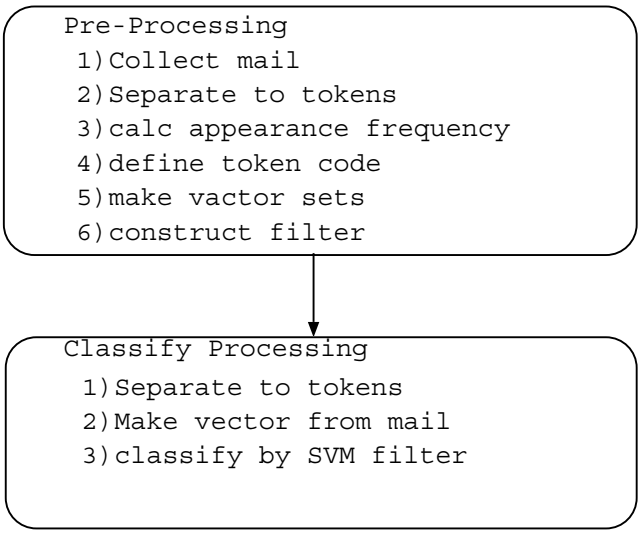

Fig. 2. Flowchart of SVM Spam Filter

frequency, the definition by the occurrence count or TF-IDF are thought.

The SVM filter procedure for classifying whether a mail is spam mail or regular mail is as follows. The procedure separates into pre-processing (filter learning) and the classify processing (filtering).

- pre-processing(filter learning):
  1) Collect spam mails and regular mails.
  2) Separate all mail to tokens.
  3) Calculate appearance frequency of each token.
  4) Define token code.
  5) Make vector sets with the token code and its appearance frequency.
  6) Construct filter(classification rules) by learning filter by SVM with vector sets and label (the mail is spam mail of regular mail).
- classify processing(filtering):
  1) Separate a mail to tokens.
  2) Make vector with token code and its appearance frequency.
  3) Using SVM filter and this vector, classify the mail.

The flowchart of SVM spam filter is shown in Figure 2.

V. Feature of Filter Based on Word Frequency

There is an advantage that the constructing filter cost can be reduced by constructing the filter based on the frequency of words. The frequency information on words can be easily extracted if there is a source text. The filter can be constructed only with giving information whether the mail is spam mail or regular mail. Moreover, because it is possible to make the personal filter learned by Bayesian spam filter, the SVM spam filter, the personal filter can be constructed according to the feature of each user's spam mail and regular mail. Because only the frequency of the word is used, the performance can be expected of not only already-known mail but also unknown mail.

However, there is the following problems in the filter based on the frequency of the words.

1) Classifying is difficult if mail content is too short.
2) Cannot classify exactly if exactly learning data could not obtained.
3) Overlook spam mail which contains much words that is popular in regular mail.
4) Overlook spam mail which contains only neutral words and URL.

There is a possibility that the input token to the filter cannot be obtained when the mail content is too short or empty. In general, it is easy to be able to classify all E-mail with short text to be spam mail because these E-mail is not necessary mail. However to classify whether the spam mail or regular mail when the short mail is usually exchanged.

It influences the performance of the filter because accurate appearance frequency cannot be obtained if there is no accurate learning data. The more voluminously accurate learning data needed to calculate accurate appearance frequency. By this reason, the performance does not improve if filter doesn't learn before a large amount of spam mail is received. The filter with general performance can be made by using the site where the spam mail is collected. After the filter that can separate general spam mail is constructed, the filter can with each user classifies for each users operation.

For instance, the spam mail that contains a lot of words that popular appear in regular mail is "mail from mailing list concerning the study of TOEIC"(regular mail) and "advertising mail of examination teaching material of TOEIC"(spam mail). This problem is caused that same high frequency words appear in both regular mail and spam mail. For this problem some some implementation are thought.

- use mail context with header to calculate word frequency
- use black list filter before use word frequency filterq

The neutral words are the words which popular words in

both spam mail and regular mail, such as greeting, “Hello”, “Hi”. It is difficult to classify with only neutral words whether the mail is spam mail or not. Only the greeting and URL are described in the mail content by the advertising mail, and the advertisement is put the link URL ahead, and there is mail that induces the user to the advertisement web page. There is an infected with computer virus on web page linking ahead when malignant link is clicked. It is possible for this problem to use the URL pre-fetch method. [6] The method of URL pre-fetch is learning mail content with web pages by link ahead. The method of automatically acquiring information the link ahead before it filters, and assuming input data.

VI. Implementation of Spam Filter

We implemented the Bayesian spam filter and the SVM spam filter, and evaluated the performances. The relevance ratio and the reproduction ratio were used for the performance evaluation. The relevance ratio and the reproduction ratio were defined as follows.

$$rel = s/n \tag{3}$$

$$rep = s/c \tag{4}$$

In this expression, we defined these symbols.

- rel: relevance ratio
- rep: reproduction ratio
- n: number of mail classified by filter as regular mail
- c: number of all real regular mail
- s: number of real regular mail classified by filter as regular mail

The proportion of an actual regular mail in the mail classified as regular mail by the filter is shown by the relevance ratio. The ratio of the mail classified as regular mail by the filter in an actual regular mail is shown according to the reproduction ratio.

A. Implementation of Bayesian Spam Filter

We implemented the Bayesian spam filter, and evaluated the performance. Bsfilter [7] was used as a Bayesian spam filter. The English tokens were used the alphabet(A-Z), the number(0-9), the apostrophe('), and dollar mark($) to be a component, and assumed the other characters to be a delimiter. The Japanese tokens were used two consecutive Chinese characters and katakana(bigram). 150 Regular mail and 150 spam mail (Japanese, English) were prepared, and the performance was evaluated by the cross-validation method. The experiment results are shown in TableII.

TABLE II
Classification efficiency of Bayesian filter

Source	relevance ratio(%)	reproduction ratio(%)
Japanese	96.71	98.00
English	73.89	100
Both	82.40	98.33
+add process	98.66	98.33

Overall, a high reproduction ratio was shown. It is thought that only an English relevance ratio is low because good English regular mail and spam mail were not able to be prepared. In the filter that targeted Japanese and English at the same time, the relevance ratio 82.40 % was obtained as the result. The following additional processing was done, therefore the relevance ratio was able to be raised to 98.66% as a result.

- Classify to spam mail if the mail content is empty.
- Classify to spam mail if the mail content has URL, but linked site does not exist.
- Use pre-fetch method for the mail if linked site exists.

At this time, the classified mails which is spam mail classified as regular mail were examined. It has been understood well that these spam mails and similar regular mail were included in the test mails. Because regular mail and spam mail of the appearance frequency that looks like were included, it is thought that the filter was not able to be learned well.

B. Implementation of SVM Spam Filter

We implemented the SVM spam filter, and evaluated the performance. The filter was constructed by using SVM$^{\text{light}}$ as implementation of SVM. The stems were extracted by using TreeTagger [9] as English tokens. The stems were extracted by using Chasen [10] as Japanese tokens. The filter was learned by using 921 totals of mails which included 175 Japanese spam mail and 188 Japanese regular mail, 261 English spam mail, and 300 English, regular mail for the experiment. The performance of the filter after learning is shown in TableIII.

TABLE III
Classification efficiency of SVM filter

Source	relevance ratio(%)	reproduction ratio(%)
Japanese	98.00	98.00
English	100	98.04
Both	97.59	90.00

A high reproduction ratio and a high relevance ratio were obtained the experiment results from Japanese only learning, English only learning. It can be thought that an efficient spam filter can be constructed for either Japanese mail or English mail. However, the result of the reproduction ratio's lowering was obtained for the mail set that contained both Japanese and English. It is thought that the filter is constructed with tedious information because of taking a long vector which consists of Japanese and English tokens as an input. Therefore, when the filtering system corresponding to Japanese and English mail, using English only filter and Japanese only filer with the filter which classify the mail to Japanese mail or English mail becomes good efficiency than using the vector that contains Japanese and English tokens for the input. Because it is possible to classify the mail by examining Content-Type of the mail header even if the language filter is not made, the calculation cost of the language classification can be disregarded.

VII. Conclusion

In this paper, the system that classified spam mail and other mail(regular mail) was constructed by two filters with Bayesian theory and SVM(Support Vector Machine) used well by the text classification task as a text classification algorithm. It was confirmed to evaluate the performance of the spam filter constructed by Bayesian theory and SVM, and to show a high reproduction ratio and a high relevance ratio. As a result, it can be though that Bayesian filter and SVM filter are effective as the spam filter. Moreover, the URL pre-fetch method was built into Bayesian spam mail filter, and the relevance ratio was able to be improved. It can be concluded that the performance of the spam mail filter can be improved by building in the URL pre-fetch method from this result.

References

[1] Ichimura, Y., Hasegawa, T., Watanabe, I., Sato, M.: Text Mining: Case Studies, Journal of Japanese Society for Artificial Intelligence, Vol.16 No.2,pp.192–200 (2001). (In Japanese)

[2] Nasukawa, T., Kawano, H., Arimura, H.: Base Technology for Text Mining, Journal of Japanese Society for Artificial Intelligence, Vol.16,No.2,pp.201–211 (2001). (In Japanese)

[3] Nagata, M., Taira, H.: Text Classification - Showcase of Learning Theories -, IPSJ Magazine, Vol.42 No.1,pp.32–37 (2001). (In Japanese)

[4] Paul Graham: A Plan for Spam, http://www.paulgraham.com/spam.html

[5] Taira, H., Haruno, M.: Feature Selection in SVM Text Categorization, Journal of Information Processing Society of Japan, Vol.41, No.4,pp.1113-1123 (2000). (In Japanese)

[6] Ando, K., Jung-H Ha, Jae-Keun Ahn, Su-Hoon Kang, Kitano, T.: Propose New Method for SPAM Mail, Multimedia, Distribution, cooperation and mobile(DICOMO2003) symposium (2003). (In Japanese)

[7] nabeken: bsfilter / bayesian spam filter, http://www.h2.dion.ne.jp/ nabeken/bsfilter/

[8] Thorsten Joachims: SVM - Light Support Vector Machine, http://svmlight.joachims.org/

[9] IMS Textcorpora and Lexicon Group: TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[10] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., Asahara, M.: Morphological Analysis System ChaSen version 2.2.1 Manual (2000). [Online] Available: http://chasen.aist-nara.ac.jp/chasen/bib.html.en