Extracting Co-occurrence Feature of Words for Mail filtering

Seiya Temma
Graduate School of Sciences and Technology for Innovation, Yamaguchi University, 1677-1 Yoshida Yamaguchi, Yamaguchi 753-8512, JAPAN

Manabu Sugii
Faculty of Global and Science Studies, Yamaguchi University,1677-1 Yoshida Yamaguchi, Yamaguchi 753-8541, JAPAN

Hiroshi Matsuno
Graduate School of Sciences and Technology for Innovation, Yamaguchi University, 1677-1 Yoshida Yamaguchi, Yamaguchi 753-8512, JAPAN
E-mail: p040de@yamaguchi-u.ac.jp, manabu@yamaguchi-u.ac.jp, hmatsuno@yamaguchi-u.ac.jp

Abstract

Spam mail filters often take advantage of appearance frequency of words in a text for mail classification. However the appearance frequency is one of the most important attribute information with which the mail can be characterized, not a few spam mails can not be distinguished with only the appearance frequency of words. In order to search new attribute information to characterize and classify the mails, we analyzed relationship between words in a text of mails by text data mining. Also we visualized the word network by the co-occurrence and multi-dimensional scaling analysis with the jaccard coefficient in real mails. The co-occurrence network analysis showed important word connections with noun and verb in the same kinds of mails. Multi-dimensional scale analysis showed some word clusters extracted from the same kind of mails.

Keywords: mail-filtering, attribute information, text mining, co-occurrence network.

1. Introduction

We are still getting many spam mails from the Internet, nevertheless a lot of kinds of mail filtering system have been developed by using database or machine learning based on Bayes' theorem[1-3]. It seems kind of like "a cat-and-mouse game" because spam mail senders release new spam mails by taking advantage of vulnerability of a new filtering method based on features extracted from past spam mails.

We investigated the method for characterizing mails with machine learning, and we developed a new filtering system which could break out of the bad loop of "a cat-and-mouse game" by using features in ham mails for filtering[4]. The system uses appearance frequency of each word and its sequence patterns in a mail body. These features (the appearance frequency and its sequence patterns) are attributes of words, and these attributes can contribute to characterize and classify each mail.

We focused on three advantages for mail filtering method based on the attribute information[5] as follows;

1. reducing information
2. protection of private information and privacy
3. easy apply to multi-language

For example, amount of information of a mail body can be reduced to about 1/4 by converting each word into a single character as an attribute symbol represented its appearance frequency. Then, it is not easy to recover the

original mail body from the converted attribute symbols. Finally, it is not necessary to process natural language analysis by this method because the attribute conversion does not depend on specific language. As a result, it is easy to apply the system to any languages.

Some attributes of words like appearance frequency, a part of speech, tf-idf value, cosine similarity and the Jaccard index are known as important factors to characterize words and documents, and they are applied to mail filtering system to characterize spam mails. But it is still difficult to classify some kinds of spam mails similar to ham mails like unsolicited mails about "Online dating service".

In this study, we are trying to find better attributes for the spam mail filtering by text mining. We considered that it is possible to extract the new attributes which can cluster the same kind of mails, and we focused on a co-occurrence network of words as the one of the attributes.

2. Method of Searching attributes

In order to find new attributes that can remove typical spam mails or perform high accuracy mail filtering, it needs to search not only characteristics of words in mails, but also characteristics of groups based on the connection of words in mails. Thus, we visualized a co-occurrence network with the Jaccard index and analyzed clusters extracted by multidimensional scaling with it. The Jaccard index is represented by the equation (1) and it indicates co-occurrence intensity between two words in this study.

$$J(A,B) = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|}, \quad (1)$$

for given sets **A** and **B**.

The Jaccard index between each two words are calculated from a number of mails which contain these words according to equation (1). In both methods, co-occurrence network analysis and multidimensional scaling, some clusters of words are visualized on two dimensional space with similarity values between two words. The result of the visualization is useful to grasp the whole image of the relationship between each word in sample mails.

2.1. *KH coder*

KH coder[6] is a free software for analyzing text data statistically. And it was produced to analyze various social survey data such as free description of questionnaire, interview article, newspaper article etc.

KH coder has functions of aggregating and searching words in data, and multivariate analysis and visualization by using R and morphological analysis module, Chasen or MeCab. The Jaccard index was calculated by KH coder, and a part of speech of each word was also identified with MeCab in KH coder. We visualized co-occurrence networks and words distribution by multidimensional scaling with the Jaccard index.

2.2. *Co-occurrence network*

Co-occurrence network is a network diagram composed of edges and nodes with the Jaccard index over a threshold. Some clusters are usually formed in the co-occurrence network which is produced with the words from mail bodies. The words in the cluster represent a strong relationship and similarity between these words in the sample mails.

2.3. *Multidimensional scaling*

Multidimensional scaling is a method to visualize similarity of individual elements of dataset. Similarity between each element is regarded as distance of them, and all elements are laid on the N-dimensional space depending on the distance without contradiction. As a result, high similarity elements are laid from a short distance on the space. In this study, we applied the Jaccard index to the similarity between words in sample mails.

2.4. *Sample mail*

In this study, we used the 2007 TREC Public Spam Corpus[7]. It is composed of 30,338 mails (spam: 50,199 mails, ham: 25,220 mails) accepted from 2007/4/4 to 2007/7/6. Two thousand mails were extracted from each spam and ham partial set of TREC07 as a sample mail set, and input them into KH coder.

3. Result

Fig. 1 shows a co-occurrence network with all words from whole TREC07 mail samples. There are many

clusters composed of co-occurrence words (the Jaccard index >= 0.3) in the same kind of sample mails. For example, the words extracted from the mails about software development or advertising mails or having the same topic mails show a tendency to make one cluster. But it is too complicated to see details of each cluster components in Fig. 1 because too many words are shown in the network, and the layout of words in each cluster is too small.

Fig. 2 shows the re-construction of the co-occurrence network with only nouns from ham mails. It has two major clusters, A and B. The words in the cluster A and B are extracted from the mails with a common mail signature and system log announcement, respectively. This result shows that each cluster in co-occurrence network is composed of the words extracted from similar kinds of mails. In other words, a mail can be characterized with the co-occurrence word set which composes such clusters in the co-occurrence network.

Fig. 3 shows the re-construction of the co-occurrence network with nouns and verbs from ham mails. It has still two major clusters, A and B extracted from the same mail set. However these clusters have nouns and verbs (red under line) as well, a lot of co-occurrence connections between nouns and verbs are shown in the cluster A. But there is only one verb in cluster B because the mails which compose the cluster B contain few verbs in the first place. This result suggested that the co-occurrence network between nouns and verbs might indicate an important feature to characterize a certain mail.

Fig. 4 shows the result of multidimensional scaling. Each word is distributed on the 2-dimensional space which has unknown coordinates, and high similarity words are located at the close region. We focused on three regions and words in these regions, A, B and C. The words in region A, B and C are extracted from the mails about CNN news, the mails with a common mail signature and the system log announcement, respectively.

When nouns and verbs are distributed on 2-dimensional space by multidimensional scaling (Fig. 4b), the verbs are located around the center region, and they are close to the region B. Fig. 4c shows the result of multidimensional scaling without the words from the system log announcement. The distribution of the words changed completely but verbs were still located around the center region.

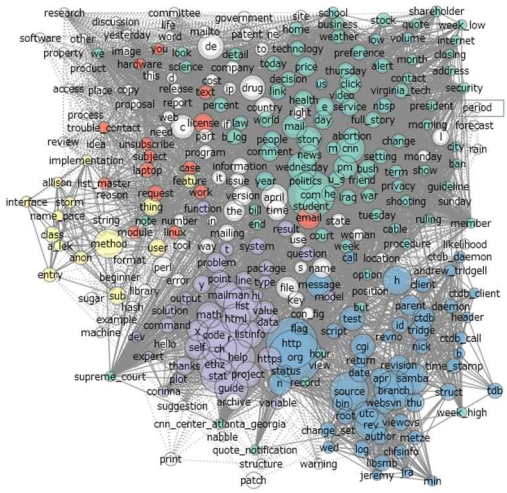


Fig. 1. Co-occurrence network of TREC07 mail sample

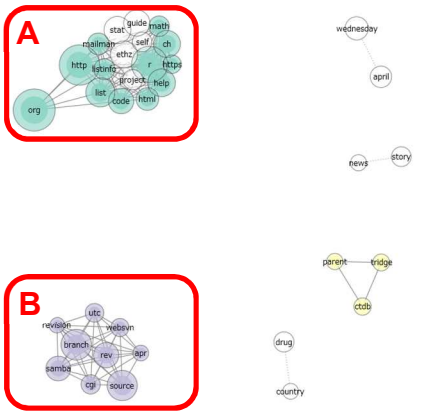


Fig. 2. Co-occurrence network with only nouns from ham mails

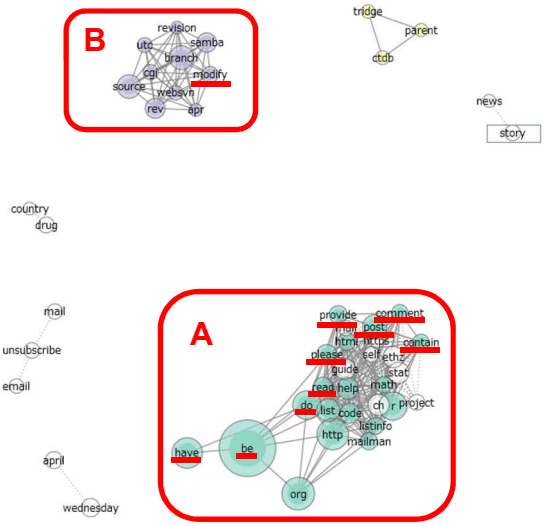
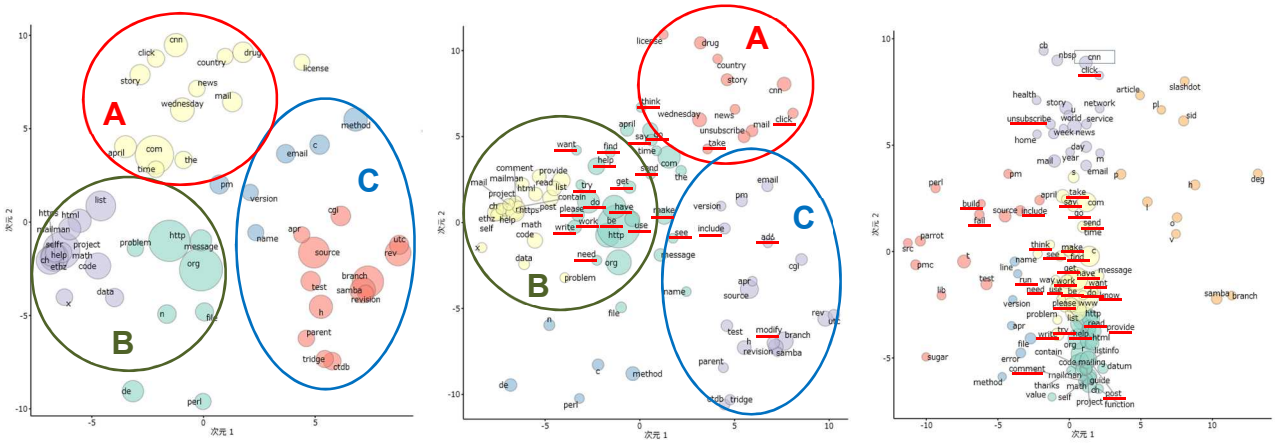


Fig. 3. Co-occurrence network with nouns and verbs from ham mails



a : only nouns. b : nouns and verbs c : nouns and verbs without system log announce mails

Fig. 4. The distribution of words on 2-dimensional space by multidimensional scaling

4. Discussion

Most of all mail filtering system try to characterize mails as only two categories, spam or ham. Of course, it is the very simple method for the mail filtering but it might be also difficult process to find features because there are too many kinds of mails to characterize as two categories in both spam and ham. In other words, it is very difficult to find common features from all spam mails or ham mails.

The result of this study shows that it is possible to group mails into several types with the combination of co-occurrence words because the words extracted from the same kind of mails cluster easily in the co-occurrence network. Then, the words with the low level co-occurrence index can be visualized as a cluster in the co-occurrence network by removing the mails which contain the high level co-occurrence words[8].

Some kinds of spam mails like the system log announcement and advertising mails which contain only URL can be removed by checking the co-occurrence connection between nouns and verbs. They have the quite few connections between them.

We do not have a concrete measure to use the co-occurrence connection set between nouns and verbs as the attribute but we considered that the combination of them is very important to characterize the type of mails. At least, it could be better attribute than an appearance frequency of words for the filtering of unsolicited mails like “Online dating service” because these mails have similar use of words in ham mails. We will plan how to utilize it for mail filtering as our future work.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP16K12438.

6. References

1. P. Graham, “A Plan For Spam”, 2002: http://www.paulgraham.com/spam.html
2. P. Graham, ”Better Bayesian Filtering”, 2003: http://www.paulgraham.com/better.html
3. bsfilter-bayesian spam filter: https://ja.osdn.net/projects/bsfilter/
4. M. Sugii, H. Matsuno, Decision Tree Representation of Spam Mail Features by Machine Learning, 2007(16(*2007-DPS*-130)), 183-188, 2007-03-01,2007.
5. N. Fujii, M. Sugii, and H. Matsuno, Finding the Candidates of Attributes in Mail Words Filtered by Bayesian Method, *IEICE Technical Report*, vol. 116, no. 525, MSS2016-89, pp. 43-48, 2017.
6. KHcoder : http://khc.sourceforge.net/
7. TREC 2007 Public Corpus : http://plg.uwaterloo.ca/~gvcormac/treccorpus07/
8. S. Temma, M. Sugii, H. Matsuno, Searching Attribute Information for Mail Filtering based on Text Mining, *ITC-CSCC2018 proceedings*, 596-599, 2018.