The Document Similarity Index based on the Jaccard Distance for Mail Filtering

Seiya TEMMA[1], Manabu SUGII[2], and Hiroshi MATSUNO[3]

1, 3: Graduate School of Sciences and Technology for Innovation, Yamaguchi University.
2: Faculty of Global and Science Studies, Yamaguchi University.

1:p040de@yamaguchi-u.ac.jp
2:manabu@yamaguchi-u.ac.jp
3:hmatsuno@yamaguchi-u.ac.jp

Abstract

We propose a new index of similarity for classification of emails into ham and spam ones with the Jaccard index. It takes advantage of co-occurrence value of all pairs of two words in emails. The co-occurrence of words represents a sort of context in documents because a word is often in use with another word in the same context. Our proposed method classified emails into hams or spams with high accuracy rate than the present filtering system using appearance frequency of word. Our method could extract patterns of word usage reflecting the context of emails.

Keywords: mail-filtering, text mining, Jaccard index, co-occurrence words, attribute information

1. Introduction

We are still getting many spams from the Internet, nevertheless various kinds of mail filtering system have been developed, for example, based on Bayesian method [1, 2]. Mail filtering systems often use word appearance frequency to classify emails into some categories. Most emails can be classified correctly using the method utilizing statistical inference with appearance frequency of words in emails. However, there are some emails which cannot be classified into the appropriate categories.

It seems kind of like "a cat-and-mouse game" because spam mail senders release new spams by taking advantage of vulnerability of a new filtering method. We are trying to find attribute information in emails to classify them into some categories which users want to manage. We focused on appearance frequency of words and its sequential patterns in a mail body, which is extracted by machine learning [3]. Although the attribute information could separate spams with high accuracy rate, it is still difficult to classify unsolicited emails about "Online dating service" correctly.

In the previous study, we employed co-occurrence value of a pair of two words to characterize emails, and visualize some types of emails in the co-occurrence networks [4]. In this study, we propose a new similarity index with co-occurrence value of two words, and compared the performance between our proposed method and bsfilter [2].

2. Method

2.1. Sample mail

We used the 2007 TREC Public Spam Corpus [5], which is composed of 30,338 mails (spam: 50,199 mails, ham: 25,220 mails) accepted from 2007/4/4 to 2007/7/6. For training, ham and spam mail sets in date order are prepared by extracting two sets of five thousand emails from ham and spam partial set of TREC07, respectively. For filtering test, ham and spam mail sets are prepared by extracting two sets of five thousand emails, which differ from the ones in the training sets, by the same way as extracting the training sets.

2.2. Jaccard index

The co-occurrence was evaluated based on the Jaccard index, which is defined as follows.

$$\boldsymbol{Jac(A,B)} = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A} \cup \boldsymbol{B}|}, \#(1)$$

for given sets A and B.

The Jaccard indexes between all pairs of two words in all training mails were calculated, and the ones in a test mail were compared with the ones in the spam and ham training mail set.

2.3. Document similarity index

We propose a new index of similarity for the classification of hams and spams. The index, named document similarity index (*DSI*), is calculated from the Jaccard index of all pairs of two words.

To obtain the *DSI*, the deviation of co-occurrence frequency of two words between hams and spams was calculated as *JacDev* from the Jaccard index (Jac_H and Jac_S) in ham and spam training set according to equation 2a or 2b. We defined two types of *JacDev* as shown in 2a and 2b below, and compared classification accuracy of them.

$$\boldsymbol{JacDev}(\boldsymbol{w}_{\mathrm{j}}, \boldsymbol{w}_{\mathrm{k}}) = Jac_{\mathrm{H}}(w_{\mathrm{j}}, w_{\mathrm{k}}) - Jac_{\mathrm{S}}(w_{\mathrm{j}}, w_{\mathrm{k}}), \#(2a)$$

$$\boldsymbol{JacDev}(\boldsymbol{w}_{\mathrm{j}}, \boldsymbol{w}_{\mathrm{k}}) = \frac{Jac_{\mathrm{H}}(w_{\mathrm{j}}, w_{\mathrm{k}}) - Jac_{\mathrm{S}}(w_{\mathrm{j}}, w_{\mathrm{k}})}{Jac_{\mathrm{H}}(w_{\mathrm{j}}, w_{\mathrm{k}}) + Jac_{\mathrm{S}}(w_{\mathrm{j}}, w_{\mathrm{k}})}, \#(2b)$$

for given two words w_j and w_k. The positive or negative sign of the *JacDev* is determined depending on co-occurrence frequency of the two words between ham and spam training set.

The *DSI* of a test email is given below as equation (3), which calculates the average of *JacDev* of all pairs of two words ($T(d_i)$) in the test mail.

$$\boldsymbol{DSI}(\boldsymbol{d}_{\mathrm{i}}) = \frac{\sum_{j=1}^{T(d_{\mathrm{i}})-1} \sum_{k=j+1}^{T(d_{\mathrm{i}})} JacDev(w_{\mathrm{j}}, w_{\mathrm{k}})}{{}_{T(d_{\mathrm{i}})}C_2}, \#(3)$$

The positive or negative sign of the *JacDev* is determined depending on co-occurrence frequency of the two words between ham and spam training set. The positive or negative *DSI* represents the similarity to hams or spams, respectively.

3. Result

3.1. Filtering of the training mail set

Figure 1 shows distribution of the *DSI* value extracted from the training mail set. The horizontal axis indicates email No. (1-5000) which is arranged in a descending order from the highest *DSI* value and the vertical axis indicates the *DSI* value. The distribution of the *DSI* based on the *JacDev* with equation 2b is separated clearly between hams and spams. The average and the standard deviation of *DSI* about training emails also show a good performance as shown in Table 1.

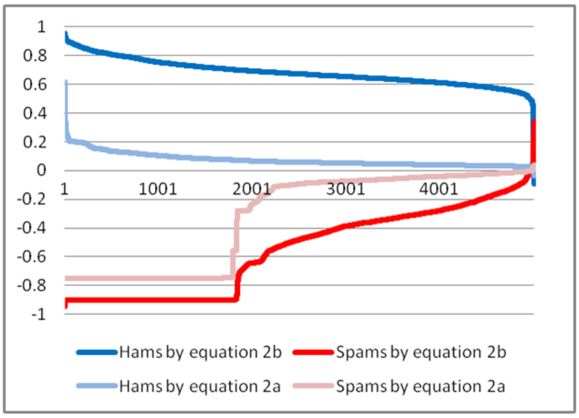

Figure 1 Distribution of the *DSI* value in the training mail set

Table 1 Average of the DSI value in the training mail set

JacDev with	*DSI*			
	Average		Standard deviation	
	Ham	Spam	Ham	Spam
equation 2a	0.074	-0.320	0.046	0.327
equation 2b	0.682	-0.553	0.087	0.294

Table 2 Table 2 The frequency distribution of a number of words with JacDev in the training mail set

	The JacDev with	
JacDev(c)	equation 2a	equation 2b
c=1.0	3408411	44053288
0.8≤*c*<1.0	5899	349977
0.6≤*c*<0.8	31006	552066
0.4≤*c*<0.6	902788	554799
0.2≤*c*<0.4	1945664	642554
0.0<*c*<0.2	40332121	473205
c=0.0	2822535499	2822535499
-0.2≤*c*<0.0	11782072	533699
-0.4≤*c*<-0.2	1198230	616782
-0.6≤*c*<-0.4	1209473	615667
-0.8≤*c*<-0.6	86505	497635
-1.0<*c*<-0.8	77416	285918
c=-1.0	4674919	16478914

Table 2 shows the frequency distribution of a number of a word pair with *JacDev* in the training mail set. The absolute value of the total *JacDev* increased with equation 2b, and it means that a separation border between ham and spam based on the *JacDev* with equation 2b is clearer than the one with equation 2a.

Figure 2 shows distribution of spam probability index [2] of the training mail set. The horizontal axis indicates email No. (1-5000) which is arranged in a descending order from the highest spam probability index and the vertical axis indicates the spam probability index.

Table 3 shows accuracy rate of classification by

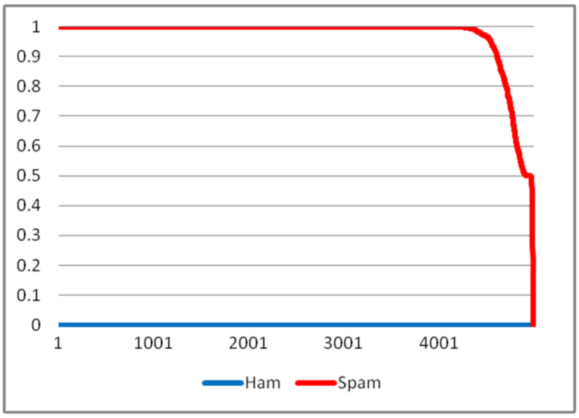

Figure 2 Distribution of the spam probability index in the training mail set

our proposed method and bsfilter with the best separation threshold. The training mail set was classified with more than 99.9% accuracy by both methods, and our proposed method was slight superior in all classification results.

Table 3 Classification accuracy by two methods in the training mail set

	Proposed method		bsfilter	
	Ham	Spam	Ham	spam
Precision (%)	99.98	99.98	99.94	99.94
Recall (%)	99.98	99.98	99.94	99.94
F value	0.9998	0.9998	0.9994	0.9994

3.2. Filtering of the test mail set

Figure 3 shows distribution of *DSI* value of test mail set. The horizontal axis indicates email No. (1-5000) which is arranged in a descending order from the highest *DSI* value and the vertical axis indicates the *DSI* value. The average and the standard deviation of the *DSI* value in the test mail set also show a good performance as shown in Table 4.

Figure 4 shows distribution of spam probability index of the test mail set. The horizontal axis indicates email No. (1-5000) which is arranged in a descending order from the highest spam probability index and the vertical axis indicates the spam probability index.

Table 4 Average and SD of the *DSI* value

JacDev by	*DSI*			
	Average		Standard deviation	
	Ham	Spam	Ham	Spam
equation 2a	0.061	-0.163	0.047	0.255
equation 2b	0.524	-0.256	0.127	0.356

Table 5 shows accuracy rate of the classification by our proposed method and bsfilter with the best separation threshold. The test mail set was classified with more than 98% accuracy by both methods, and our proposed method was slight superior in all classification results.

Table 6 shows correlation coefficient of *DSI* value and the spam probability index. There was no high correlation in ham mail.

Figure 5 and Figure 6 show correlation diagram of *DSI* value and the spam probability index in the ham and spam training set. The horizontal axis indicates the *DSI* value and the vertical axis indicates the spam probability index. Some mails that could not be filtered by bsfilter were filtered by our proposed method.

Figures 7 and 8 show overlap distribution (green area) range of the *DSI* value extracted from both hams and spams in the training sets with equation 2a and 2b, respectively. Figure 9 also shows the overlap distribution range with the spam probability index. Table 7 shows the number of mails distributed their

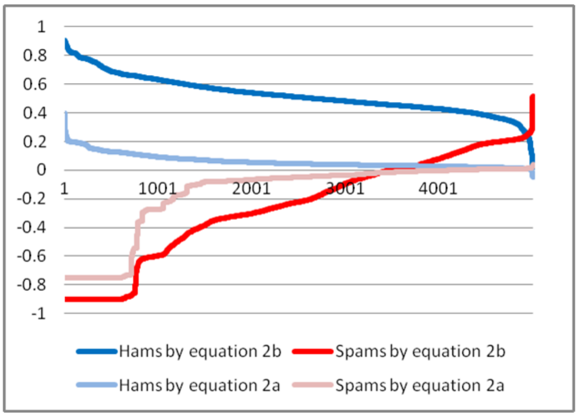

Figure 3 Distribution of *DSI* value in the training mail set

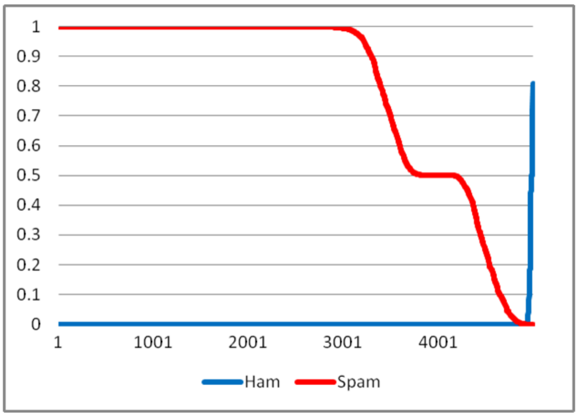

Figure 4 Distribution of the spam probability index in the training mail set

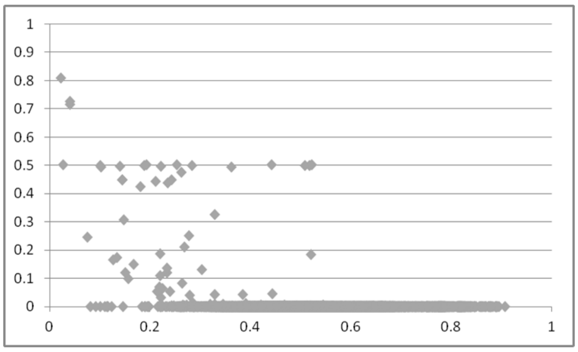

Figure 5 Correlation diagram of the ham mails

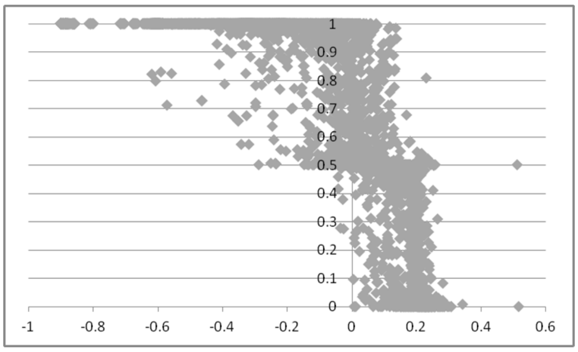

Figure 6 Correlation diagram of the spam mails

DSI value within the overlap distribution range in Figures 7, 8 and 9.

Table 5 Classification accuracy by two methods in the test mail set

	Proposed method		bsfilter	
	Ham	Spam	Ham	spam
Precision (%)	98.80	98.78	98.32	98.32
Recall (%)	98.78	98.80	98.32	98.32
F value	0.9879	0.9879	0.9832	0.9832

Table 6 Correlation coefficient of *DSI* value and spam probability index

	Ham	Spam
Correlation coefficient	-0.2	-0.69

Table 7 Number of mails in the overlap distribution range

	Ham	Spam	Total
equation 2a	1954	2427	4381
equation 2b	2644	1311	3955
bsfilter	5000	1614	6614

The overlap distribution range with the *DSI* value is narrower than the one with the spam probability index. Especially, the overlap distribution range of the *DSI* value with equation 2b is quite narrow. These results show that the method using the *DSI* value with equation 2b can classify mails into ham and spam with minimum failure.

4. Discussion

Generally, word usage is different depending on the context of documents, and it is reasonable for words to co-exist with other words sharing the same context in the two documents having a similar meaning. In other words, the co-occurrence value between words in the different documents indicates a similarity of these documents which have the same context or contents.

By using a combination of words increases tokens, and we expect that we will be able to grasp delicate features like context.

We consider that our proposed method could classify emails with not only typical and high frequency words but also depending on the context and contents of emails.

5. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP16K12438.

References

[1] P. Graham, "A Plan For Spam", 2002: http://www.paulgraham.com/spam.html

[2] bsfilter-bayesian spam filter: https://ja.osdn.net/projects/bsfilter/

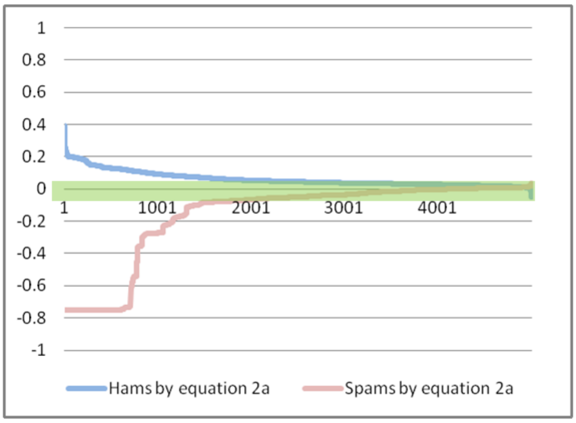

Figure 7 Overlap distribution of the *DSI* value based on the *JacDev* with equation 2a

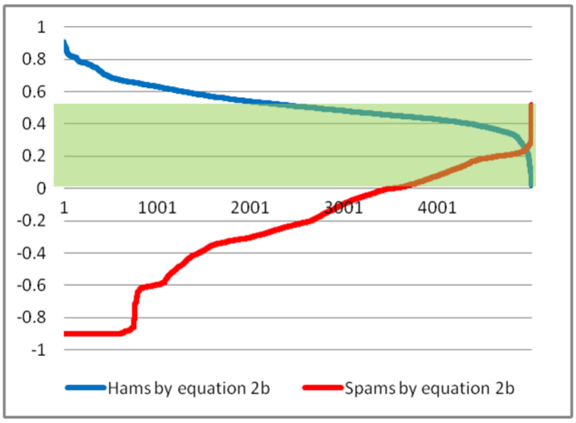

Figure 8 Overlap distribution of the *DSI* value based on the *JacDev* with equation 2b

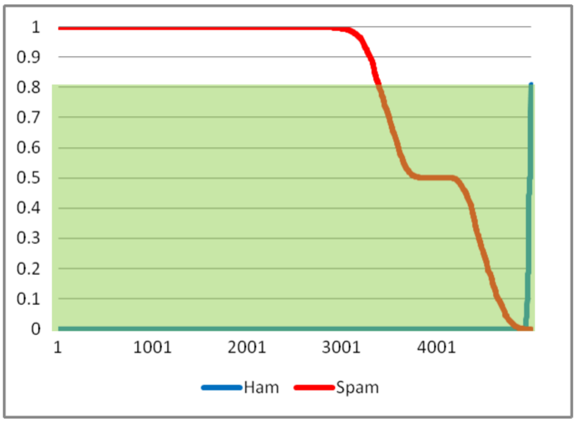

Figure 9 Overlap distribution of the spam probability index

[3] 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.

[4] S. Temma, M. Sugii, H. Matsuno, Searching Attribute Information for Mail Filtering based on Text Mining, ITC-CSCC2018 proceedings, 596-599, 2018.

[5] TREC 2007 Public Corpus: http://plg.uwaterloo.ca/~gvcormac/treccorpus07/