# A Study of Neighbor Users Selection in Email Networks for Spam Filtering

Yongchao Wang
Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute, Aichi, Japan
id151001@cis.aichi-pu.ac.jp

Yuyan Chao
Faculty of Environment, Information and Business Nagoya Sangyo University, Owariasahi, Aichi, Japan
chao@nagoya-su.ac.jp

Lifeng He
Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute, Aichi, Japan
helifeng@ist.aichi-pu.ac.jp

## ABSTRACT

Due to the protection of data security and personal privacy in email networks, it is difficult to select the neighbor user using email user's personalized attributes and behavior data. In daily life, the more interactive the people are, the more similar they are. So many researchers use the interaction strength to select the neighbor users. But this method has not been experimentally verified. Receiving the same email is the most basic condition that a neighbor user can provide a valid recommendation to the target user. So, we build an email corpus using real email data to verify the selection effect of different methods. The emails come from Enron email data set, log files of a Chinese corporate email server, and volunteers. This corpus contains more than 870,000 users and more than 5.14 million emails. With this corpus, we conducted experiments on seven different methods of neighbor user selection. The experimental results show the effect of user interaction strength is far from being as good as expected, and far inferior to user social attributes. In addition, the recipient performs far better than the sender. This paper provides a useful reference for the research of spam filtering based on user-based collaborative recommendation filtering in mail network.

## CCS Concepts

• **Information systems → Information systems applications → Data mining → Nearest-neighbor search**

• **Information systems → World Wide Web → Web applications → Internet communications tools → Email**

## Keywords

spam filtering; collaborative filtering; email network; neighbor users; interaction strength, email corpus.

## 1. INTRODUCTION

The struggle against spam has been going on for nearly half a century, but spam is still flooding our mailboxes. Text mining, Image Recognition, Bayesian, Support Vector Machines (SVM), Neural Networks, Machine Learning and other technologies have achieved excellent results, but spam is also changing. Now there are few pure spams, and most of them are advertisements. To the same email, some people think it is a spam, but some other people think it is useful. So, nobody can decide whether an email is spam or not, the identification of spam must be personalized.

In recent years, personalized collaborative filtering technology has been widely adopted by e-commerce platforms, social networks and search engines. Facts prove that this technology is very effective. Collaborative filtering recommendation algorithms are based on project or user [1,2]. In email network, the project is email. Because the quantity of emails is very large, and it increases rapidly every second, while the quantity of mail users is relatively stable. Therefore, it is better to adopt user-based collaborative filtering in the email network.

When a user receives an email, the system will give a classified opinion of this email based on the evaluation of his neighbor users. If most neighbor users think an email is spam, the email is very likely to be considered by users to be spam. Because this method is based on the user's manual recognition and evaluation, it has high accuracy. Moreover, this method does not deal with the content of e-mail information, so it does not involve the privacy of users, there is no legal risk. The most difficult problem is how to find neighbor users.

In social networks such as Twitter and Facebook, users with the same life experience, education background and interests will be recommended to target users [3,4]. In Amazon, Taobao and other e-commerce platforms, they recommend a product to you mostly because the users who had similar buying habits with you have bought [5,6]. But in email system, users' information and data are very strictly protected. We know little about the user's personal information other than the e-mail address, so it is very difficult to find a user similar to the target user in email networks.

There is an old saying that "Birds of a feather flock together" which means that similar things always come together, and like-minded people always get together. Therefore, people use the interaction strength of email users as a condition to select the similar user [7]. People are generally convinced of this, even though the method has never been experimentally verified but is generally accepted.

In this paper, we build an email corpus using real mail data, and the data comes from Enron email data set, log files of a Chinese enterprise email server, and volunteers. Then, we experiment with seven different neighbor selection methods, including user's interaction strength. After that, we evaluated them with accuracy, recall rate and yield rate of the neighbor users, and the recall rate of same sender. Finally, we got some valuable conclusions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICCNS 2018, November 2–4, 2018, Qingdao, China
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6567-3/18/11…$15.00
https://doi.org/10.1145/3290480.3290500

## 2. SELECTION CRITERIA OF NEIGHBOR USERS IN EMAIL NETWORK

User-based collaborative filtering needs to find similar users of the target user. In social networks or e-commerce platforms, the main task of the recommendation system is to recommend new friends or new products to target users. In email network, for spam filtering, the main task is to recommend the evaluation information of an email or it's sender to the target user when he receives the email. The similar users should be those who often receive the same emails and have the same attitude to an email with the target user. The more the same emails, the more recommendations a neighbor user can offer. Otherwise, if the quantity of the same email is zero, the user can not offer any recommendation and cannot be a neighbor user.

"The same email" means an email's content or its sender is same as an email which the target user received. Usually, mails with the same content come from the same sender, and a user's attitude to an email depends largely on the attitude to the sender. In fact, all mail systems have a black list and a white list to classify emails by their sender. Although the address of the sender may be forged, it can be identified.

If both user A and user B receives an email from user C, then user C is called "**the same sender**". Suppose N is the quantity of the same sender of a user and the target user, if N>0, then the user is the neighbor of target user, and the bigger N, the more similar. So, the selection criteria of neighbor users in email network is:

**"Exist the same sender(email), and the more of it, the higher similarity between neighbor user and target user**."

## 3. BUILD EMAIL CORPUS

To ensure the accuracy of experimental results, the authenticity and integrity of data must be insisted. Authenticity refers to the fact that the mail data must come from the real email user. Integrity refers to the fact that all the mail users and their emails data should be included. Because email network is a small world network [8], absolute integrity is impossible. As far as possible within a certain range to achieve relative integrity. For example, we can get all the emails data of all users of an email server over a period.

### 3.1 Source of Experimental Data

In this paper, the data comes from three sources: the U.S. Enron email dataset (hereinafter called "Enron dataset"), the log files of an enterprise email server (hereinafter called "Elog dataset") and volunteer's own email file in eml format (hereinafter called "Eml dataset").

Enron dataset is an open email corpus. TREC, SpamAssassin, Ling-spam, PU are also open email corpus on the Internet [9-13]. Only Enron dataset be used because only Enron dataset meets the requirements of authenticity and integrity. Enron Corp was once one of the largest comprehensive natural gas and power companies in the world. But in 2002 Enron suddenly declared bankruptcy, so the U.S. government launched an investigation into Enron. Enron's data set was published during this investigation, it contained the emails of 150 senior executives within about two years [14].

UT is a Chinese enterprise with more than 2,000 employees. Elog dataset is obtained by collecting the email sending and receiving logs of the company's email server. It includes all the mail data of all employees in UT within half a year. So, its integrity is much better than all the other existing corpus.

Wang et al. proposed a method of establishing email header information corpus based on personal privacy protection[15]. Using this method, we collect more than 40,000 email data through 65 volunteers. The volunteers come from different countries such as China, Japan and New Zealand. They are company employees, businessmen, researchers, teachers or students. This dataset is relatively diversified and can be used as a useful supplement in the experiment.

After data cleaning, we extracted the information of recipient, sender and sending time to build a new corpus. At last, the corpus contains 873,881 email users and 5,148,230 email's sending records. Table 1 shows the numbers of users and emails of Elog, Enron and Eml dataset. Some users appear in different data sets.

About 766 thousand users and 2.44 million emails came from Elog dataset, and the sending time of mails were from December 27,2016 to June 5,2017.

About 90 thousand users and 2.54 million emails came from Enron dataset, and the earliest date of delivery was November 1,2000, and the latest was October 9,2002.

About 17 thousand users and 164 thousand mails came from Eml dataset, and the sending time of them were from September 9,2009 to April 11,2017.

**Table 1. The quantities of emails and users in the corpus**

| Data Source | Emails | Users | Source Users |
|---|---|---|---|
| Elog | 2,443,653 | 766,129 | 2,074 |
| Enron | 2,540,431 | 90,608 | 214 |
| Eml | 164,146 | 17,372 | 65 |
|  | 5,148,230 | 873,881<br>(874,109) | 2,353 |

### 3.2 Label Email Users

To improve the availability of data, it is necessary to label the data accurately according to the known information.

First, we need to label the user as "source user" or "non-source user". The users who provide the original data are source users, while other users are non-source users. Because the data of source user is included, so the data integrity of source user is better. The data integrity of non-source users is poor, because the data irrelevant to the source users is missing. The social attribute of email users is very difficult to obtain, so all personalized information known should be recorded and annotated.

Because the enterprise mail server only allows employees to register, so we can know which company the users work in. Registered users of Enron enterprise mail server are the employees of Enron, Registered users of UT enterprise mail server are the employees of UT, their labels of "company" were marked as Enron or UT.

"JSJ" is the name of a department of UT. We got the list of employees of this department. So, the label of "department" of these employees were marked as "JSJ".

## 4. EXPERIMENTAL METHODS AND PROCEDURES

The main task of this paper is to study the selection of nearest neighbor users for spam filtering in e-mail networks. The emphasis is to verify the effectiveness of using user interaction strength to select nearest neighbor users and compare it with other methods using user's social attributes.

The selection of target users is very important. To guarantee the experimental effect, the selection of target users should be treated seriously.

## 4.1 Selection of Target Users

In this paper the selection of target user follows three standards:

1．**Ordinary**: Target user should be ordinary email users, cannot be users who send spam or commercial advertisements.

2．**Integrity**: Target user should be users with better data integrity, so target users must be source users.

3．**Active**: Target users should be active users who sent at least one email. Some users may be in a state of hibernation, the Email address was not used for a long time and never sent an email, these users should not be target user.

According to the above criteria, we selected 1199 out of 873,881 users as the target users, and Table 2 shows the situation of selected target users. Users of the JSJ department of UT company are listed separately.

**Table 2. The situation of selected target users**

| | Enron | UT | JSJ | Eml |
|---|---|---|---|---|
| **Email Users** | 162 | 972 | 56 | 65 |
| **Average Number of Received Emails** | 1645.1 | 787 | 1360 | 1748.8 |
| **Average Number of Senders** | 152.6 | 542.3 | 920.7 | 269.8 |
| **Average Number of Sent Emails** | 635.3 | 28.8 | 51 | 307.7 |
| **Average Number of Recipients** | 90.3 | 12.5 | 24 | 27.2 |

## 4.2 Selection Methods of Neighbor Users

User interaction is the primary method which we want to test in this paper. Considering that the user's interaction is directional, we subdivide the user's interaction into four modes and tests them separately.

1. **Sender**: The users who had sent email to target user.
2. **Recipient**: The users who had received mail from target user.
3. **Sender or Recipient**: The users who had sent to or received mail from target user.
4. **Sender and Recipient**: The users who had sent to and received mail from target user.

About user's social attributes, we can accurately and massively grasp only the user's "company" and "department". The "company" labels of most users have been marked, and we know all users in "JSJ" department. Therefore, as the representative of social attributes, the "company" and "department" attributes of users are also used as the selection methods.

5. **Same Company:** The users who work in the same company as target users.
6. **Same Departmen**t: The users who worked in the same department as target users.
7. **All**: All the users in corpus are treated as neighbor users. Obviously, it's the least efficient method to have all users as neighbors. But this method can get the most complete sets of real neighbor users and the same sender of each target user. It is necessary for calculating the evaluation indicators later.

To distinguish from the real neighbor users after validation, the neighbor users selected before the experiment started will be called "quasi-neighbor users".

## 4.3 Evaluation Indicators

This paper compares different methods of selecting neighbor users using the following four evaluation indicators:

1. **Accuracy of Neighbors (AN)**: The ratio of real neighbor users to quasi-neighbor users. The higher this value, the higher the accuracy of selecting the neighbor users in this way.
2. **Recall Rate of Neighbors (RN)**: The ratio of real neighbor users to all neighbor users of the target user. The higher this value, the better effect.
3. **Recall Rate of the Same Sender (RSS)**: How many percent of the same sender was found out. The higher this value, the better effect. In fact, the purpose of searching neighbor users is also to get the information of the same sender, the ultimate goal is to find out the same sender, so this indicator is more critical than the recall rate of neighbors.
4. **Yield of Neighbor (YN)**: How many the same sender can be found out by each neighbor user on average. The more neighbor users, the greater the amount of data that needs to be processed, and the more network, storage and computing resources consumed. Therefore, in the case of same recall rate, the method with high rate of return can save resources and be more efficient.

## 4.4 Experimental Steps

The experiment was conducted according to the following steps:

1. Count the quantity of received emails, senders, sent emails and recipients of each target user;
2. For each methods of selection, perform the following actions to each target user;
3. a) Find out all the users who sent email to each target user;

   b) Find out all the quasi-neighbor users of each target user;

   c) Find out the real neighbor users from the quasi-neighbor users. A quasi-neighbor user who has the same sender with the target user is real neighbor user;

   d) Find out all the same senders by real neighbor users of each target user those have been find out;

   f) Calculate the accuracy rate of neighbor users and other evaluation indicators of each target user.
4. Calculate the average of each evaluation indicator of all target users.

Table 3 shows the average quantity of quasi-neighbors, real neighbors and the same senders of all target users using different methods of selection of neighbor users.

**Table 3. The average of quasi-neighbors, real neighbors and the same senders of all target users**

| | Quasi-Neighbors | Real Neighbors | Same Senders |
|---|---|---|---|
| **ALL** | 873881 | 3304 | 704 |
| **Same Company** | 2074 | 926 | 698 |
| **Same Department** | 75 | 52 | 588 |
| **Sender or Recipient** | 932 | 88 | 241 |
| **Sender** | 921 | 84 | 188 |
| **Recipient** | 24 | 8 | 165 |
| **Sender and Recipient** | 13 | 4 | 97 |

# 5. THE ANALYSIS OF EXPERIMENTAL RESULTS

According to the experimental results, the different selection methods were analyzed and compared. Table 4 shows the values of evaluation indicators of the experimental results. It should be noted that the data in Table 4 is not computed by Table 3, but by calculating the value of each target user, and then taking the average.

Enron and EML do not have the information of user's department. The comparison results of other methods are highly consistent for the three dataset. In order to ensure the consistency of the data used by different methods, this section only shows the results of EML.

**Table 4. Evaluation Indicator of Experimental Results**

| | AN | RN | RSS | YN |
|---|---|---|---|---|
| **ALL** | 0.4% | 100.0% | 100.0% | 0.001 |
| **Same Company** | 44.7% | 39.2% | 98.0% | 0.337 |
| **Same Department** | 69.3% | 2.7% | 82.6% | 7.754 |
| **Sender or recipient** | 10.8% | 2.1% | 31.3% | 0.229 |
| **Sender** | 10.0% | 2.0% | 24.3% | 0.184 |
| **Recipient** | 47.8% | 0.3% | 21.4% | 8.512 |
| **Sender and Recipient** | 32.7% | 0.1% | 10.3% | 7.632 |

## 5.1 Accuracy of Neighbors

The accuracy of neighbors (AN) refers to the proportion of the real neighbor users in quasi-neighbor users. Fig.1 shows the accuracy comparison of seven selection methods of neighbor users. The higher the value of AN, the higher the accuracy, and the higher the probability that the quasi-neighbor user is a real neighbor user.

In this experiment, a user will be considered as neighbor user when he has same email with the target user. So, the result means:

1. Nearly 90% users of those have interact with target users have no same mail as target users.
2. The probability of a recipient being a neighbor user is several times greater than that of a sender.
3. A user with similar social attributes is more likely to be a neighbor than a user with intercourse.

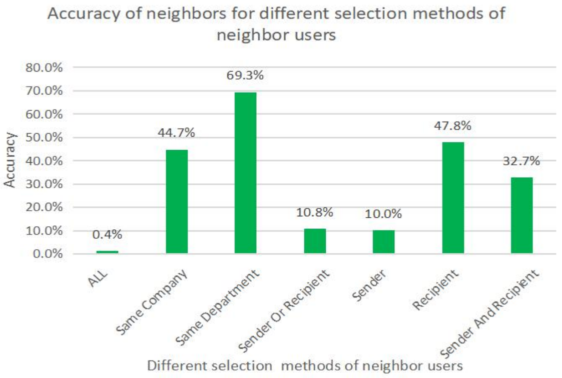

**Figure 1. Accuracy of neighbors of different selection methods of neighbor users**

## 5.2 Recall Rate of Neighbors

The recall rate of neighbor users (RN) refers to the proportion of the real neighbor users in all the neighbor users of target user. When the quasi-neighbor user is all the users in corpus, all the neighbor users will be found out and the recall rate will be 100%. Fig.2 shows the recall rate of the other six methods. In addition to 39.2% for the same company, less than 3% for all other methods, and the lowest recall rate is for sender and recipient with only 0.1%.

The more users of quasi-neighbor, the easier it is to find more real neighbor users. From table 3 we can see, the quantity of quasi-neighbor selected by the same company is far more than other methods, so it is not surprising that its recall rate is far higher than others. But in any case, the result still shows that we can only get few neighbor users by using user interaction strength, lower than 3 percent.

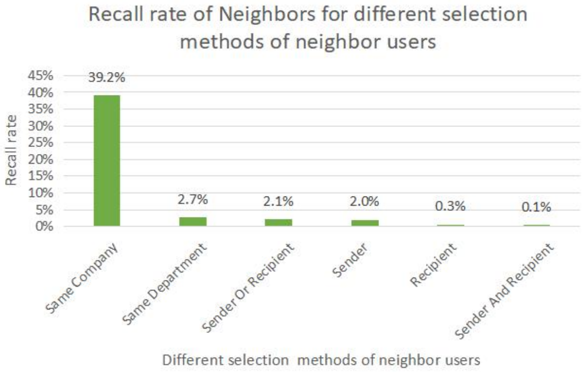

**Figure 2. Recall rate of Neighbors for different selection methods of neighbor users**

## 5.3 Recall Rate of The Same Senders

The recall rate of the same sender (RSS) is the most important evaluation indicator. The method regards all users as neighbors can find out all the same sender, so the recall rate is 100%. Fig.3 shows the recall rates of the same senders of the other different methods.

Users with similar social attributes are still more likely to have neighbors than users with intercourse. The recall rate of same company and same department is 98.0% and 82.6%. The recall rates obtained through email interaction were all below 35%.

Senders used hundreds of times more quasi-neighbors than recipient but recall rate of the same sender is only three percentage points higher than recipient. So, Again, recipient is better than sender.

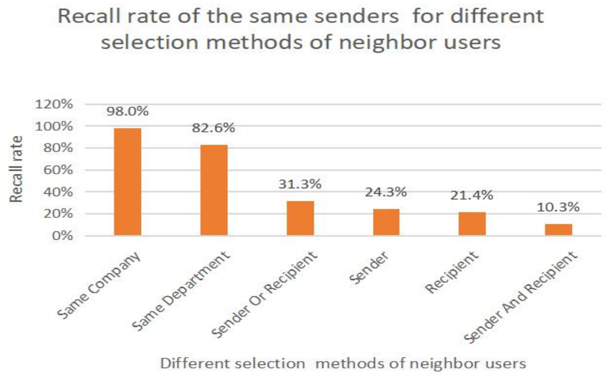

**Figure 3. Recall rate of the same senders for different selection methods of neighbor users**

## 5.4 Yield of the Neighbor

Fig.4 shows the yields of neighbor for different selection methods of neighbor users. Recipient and Same Department are significantly higher than other methods. On average, one neighbor user can find out more than 7.6 same senders for Recipient and Same Department. The yields of the other methods are all less than 0.4.

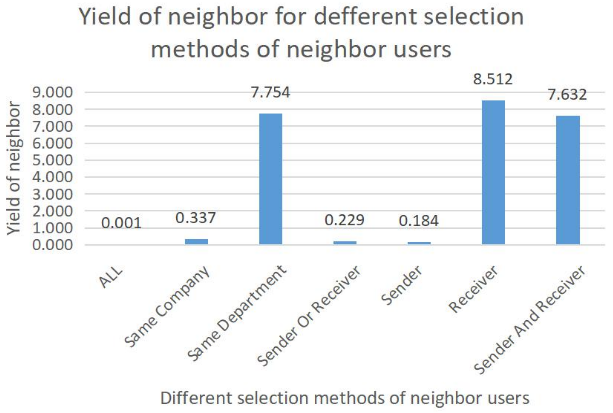

**Figure 4. Yield of neighbor for different methods**

# 6. CONCLUSION

In this paper, we build an email corpus using real email data and experiment with different methods of neighbor users selection. After analysis and comparison the results, we get the following conclusions:

1. The performance of methods using email interaction strength to select neighbor users is not as well as people thought.

2. The performance of methods using users' social attributes is much better than using email interaction strength.

3. The performance of method using target user's recipients is much better than using target user's senders. For recipient, the accuracy and yield are good, but the recall is too low. For sender, all indicators are very low.

In this paper, we did not consider the individual differences of each the same sender. Next, we will include the number of emails sent by the same sender into the evaluation system. This paper provides a useful reference for the research of spam filtering based on user-based collaborative recommendation filtering in mail network.

# 7. REFERENCES

[1] GB/T 7714Sarwar B, Karypis G, Konstan J. Riedl, J. "Item-based collaborative filtering recommendation algorithms." International Conference on World Wide Web ACM, 2001, pp.285-295.

[2] Schafer, J. Ben, et al. "Collaborative Filtering Recommender Systems." Acm Transactions on Information Systems,2007,22.1, pp.5-53.

[3] Fernándeztobías I, Cantador I. Personality-Aware Collaborative Filtering: An Empirical Study in Multiple Domains with Facebook Data[J]. Lecture Notes in Business Information Processing, 2014, 188, pp.125-137.

[4] Seo Y D, Kim Y G, Lee E, et al. Personalized recommender system based on friendship strength in social network services[J]. Expert Systems with Applications, 2017, 69, pp.135-148.

[5] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, vol.7, pp.76-80.

[6] Song M, Zhou X, Haihong E, Ou Z. "A Recommender System Model based on Commodity-Purchase-Cycle Classification." Eai International Conference on Mobile Multimedia Communications ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2016, pp.48-53.

[7] YANG Zhen,LAI Ying-Xu,DUAN Li-Juan,LI Yu-Jian,XU Xin;Spam Collaborative Filtering in Enron E-mail Network;ACTA AUTOMATICA SINICA;2012, vol.38, pp.399-411.

[8] Ebel H, Mielsch L I, Bornholdt S. Scale-free topology of e-mail networks[J]. Physical review E, 2002, vol.66: 035103.

[9] Enron email dataset, https://www.cs.cmu.edu/~enron/

[10] Trec spam dataset, https://trec.nist.gov/data/spam.html

[11] Spamassassin, https://spamassassin.apache.org/old/publiccorpus/

[12] Lingspam dataset,

http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz

[13] PU dataset,

http://www.aueb.gr/users/ion/data/PU123ACorpora.tar.gz

[14] Klimt B, Yang Y, The enron corpus: A new dataset for email classification research[C]//European Conference on Machine Learning. Springer, Berlin, Heidelberg, 2004, pp. 217-226.

[15] Yongchao Wang, Xiao Zhao, Feihang Ge, Yuyan Chao, and Lifeng He, A Corpus of Email Headers with Personal Privacy Protection; Journal of Advances in Computer Networks, 2017, vol.5,pp.53-58.