A Corpus of Email Headers with Personal Privacy Protection

Yongchao Wang, Xiao Zhao, Feihang Ge, Yuyan Chao, and Lifeng He

***Abstract*—Because emails are private information, it is hard to acquire enough authentic data to build a corpus of emails. Email headers, however, do not involve email bodies, thus have less privacy. An email header, which contains the recipient, the sender, and a lot of other key information about the email sending process, has a high value for related research. This paper proposes an idea for building a corpus of email headers for the first time. The idea is to encrypt sensitive data via Secure Hash Algorithm when collecting key fields in email headers. All corpus data can be examined by volunteers themselves to confirm that no privacy remains. For ease of use, each data in this corpus contains is labeled with the number of recipients, the sending and receiving geographical locations, the user's social attributes such as country, language, job, professional, and so on, where some information of user's social attributes are obtained through questionnaires. The corpus can be applied to the research fields such as community discovery, users' relationship analysis, email classification, and spam email recognition, etc. Moreover, the method for building a corpus of email headers proposed in this paper can also be applied for other corpus data collection work where users' privacy protection is necessary.**

***Index Terms*—Corpus, mail header, SHA, privacy protection, corpus labeling, social network, spam.**

I. Introduction

Nowadays, telephone, Line, WeChat, and other real-time communication tools become the most popular communication methods for people. In formal occasions, however, such as scientific researches, commercial activities, and office work, email remains as the most frequently used Internet communication method. Researches surrounding email include recognition and filtering of spam emails [1], [2], email networks [3], email-related data mining [4], etc. For all such researches, sample data are needed. Unfortunately, because emails involve personal privacy, it is hard to acquire a large quantity of authentic email data, especially when these email data are used in a public corpus that will be provided to all scientific researchers. Some researchers in the research field about email could only acquire several thousands of emails from volunteers despite much effort they have made, and most of which are spam emails.

Public email corpora currently available on the Internet include Enron, TREC, SpamAssassin, CSDMC, Maildb, Ling-Spam, etc [4]-[6]. The emails in Ling-Spam contain email bodies only, which are all encrypted. The mails in other corpora contain both email headers and bodies. Among these corpora, Enron is the most authentic and complete email corpus, which comes from 150 managers of U.S.-based Enron Corporation. The Enron corpus is attributed to a pure accident. Were it not the famous inspection of the U.S government on the Enron incident, the Enron corpus would not exist. All other email corpora except Enron are occupied by emails that are considered as unimportant by users (e.g. spam emails), and the authenticity and completeness of these emails are not guaranteed. Although these corpora played an important role in related research fields in the last decade, most of them were built a decade ago. It is necessary to build new corpora to reflect the features of emails of recent years.

If we can solve the problem of personal privacy protection, it might be easier to acquire a large number of email data. In consideration of the high research value of email headers and their less privacy level, the authors propose a method for building a corpus of email headers that can protect personal privacy. With this method, the research team has collected over 80,000 pieces of email headers and over 90,000 records of email sending and receiving within two weeks. With these data and the existing email corpora, the research team has built a corpus consisting of 3.8 million pieces email headers. The corpus can be applied to in the research fields such as of community discovery [7], users' relationship analysis [8], email classification and spam email recognition [9]-[11], etc.

The remains of the paper are organized as follows: We introduce email header in the next section and show the method used to protect user's personal privacy in Section III. Section IV explains the method for data collection and cleansing, and Section V introduces data labeling. Lastly, we give our concluding remarks in Section VI.

II. Email Header

A hand-written letter is comprised of two parts — envelope and content in the envelope. The envelope is mainly used to provide the information, which mainly contains the

Manuscript received August 25, 2017; revised October 10, 2017.

Yongchao Wang is with the Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute 480-1198, Aichi, Japan, and the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, Shaanxi, China (e-mail: wyc@xaut.edu.cn).

Xiao Zhao is with the Artificial Intelligence Institute, College of Electrical and Information Engineering, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: zhaoxiao@sust.edu.cn).

Feihang Ge is with the Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute 480-1198, Aichi, Japan, and with the Zhejiang College of Construction, China (e-mail: gfhang@163.com).

Yuyan Chao is with the Graduate School of Environment Management, Nagoya Sangyo University, Owariasahi 488-8711, Japan (e-mail: chao@nagoya-su.ac.jp).

Lifeng He is with the Graduate School of Information Science and Technology, Aichi Prefectural University, Nagakute 480-1198, Japan, and Artificial Intelligence Institute, College of Electrical and Information Engineering, Shaanxi University of Science and Technology, Xi'an 710021, Shaanxi, China (e-mail: helifeng@ist.aichi-pu.ac.jp).

doi: 10.18178/jacn.2017.5.2.240

recipient's name and address as well as the sender's name and address, required by the postal office for delivery purpose. After the sender posts off his/her letter, the letter will be transferred from one postal office to others within the postal system according to the address written on the envelope, and finally arrive at the postal office where the recipient is located. The postman will then deliver the letter to the recipient. The email system operates in almost the same way as the postal system, except that hand-written letters are replaced by emails, and postal offices are replaced by email servers. An email is also comprised of two parts – email header and email body, where the email header is like the envelope, and the email body is like the content in the envelope.

The email header of an email mainly contains information related to email sending, in which the most important is the recipient's name and address, and the second important is the sender's name and address. Besides, it contains information about the email servers involved in the process of sending the email via Internet, and other related information. Fig. 1 shows an email header.

```
Return-Path: <yumiao_zgj05@163.com>
Delivered-To: wyc@xaut.edu.cn
Received: from 220.181.13.153 (HELO mi3-153.163.com) (envelope-from yumiao_zgj05@163.com)
    by quarkmail.com (quarkmail-1.2.1) with ESMTP id S6062336AbdDYCBt
    for wyc@xaut.edu.cn; Tue, 25 Apr 2017 10:01:49 +0800
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed; d=163.com;
    s=s110527; h=Date:From:Subject:MIME-Version:Message-ID; bh=s8b2E
    b+696APSU23x4W53aEe7rQr36dhStsqd7/yltk=; b=Y5WIHFBE2LfpAlNxEwVJn
    23OsfX+PiDOhOK59EP8ry8EXCFyiKMYbjg3B3B2+InlNENh0t2d6BnNE2Z9g8BOg
    FqUXrHg8vyP9jN7nNWyRIWf0M8Y0FPmlbL4zZmjohFt52Fn5/TpgDte7AeVF2wx/
    TRBbSLEhnmR7d87bxxBKos=
Return-Path: <yumiao_zgj05@163.com>
Received: from yumiao_zgj05$163.com ( [60.7.117.158] ) by
 ajax-webmail-wmsvr153 (Coremail) ; Tue, 25 Apr 2017 10:02:04 +0800 (CST)
X-Originating-IP: [60.7.117.158]
Date: Tue, 25 Apr 2017 10:02:04 +0800 (CST)
From: =?UTF-8?B?5LqO5re8LeS4remrmOaVmQ==?= <yumiao_zgj05@163.com>
To: wyc@xaut.edu.cn
Subject: =?UTF-8?B?44CQ5YWz5LqONS416YeN5bqGIDUuMTLmoYLmnpc=?=
 =?UTF-8?B?IDUuMTjplb/mspkgNS4xOeaYhuaYjuetiQ==?=
 =?UTF-8?B?5Zyw5Li+5Yqe5biI6LWE5pWZ5a2m6K6y5bqn6YCa55+144CR4pie6K+35p+l?=
 =?UTF-8?B?5pS25pil5a2jNS025pyI56C5L+u54+t6K++56iL5paH5Lu277yB?=
X-Priority: 1
X-Mailer: Coremail Webmail Server Version SP_ntes V3.5 build
 20160729(86883.8884) Copyright (c) 2002-2017 www.mailtech.cn 163com
```

Fig. 1. An example file of email header submitted by a volunteer.

Email protocols include RFC822, POP, SMTP, MIME, etc. According to the RFC822 protocol, an email must contain the header and the body. This protocol also requires that the standard fields must be contained in the header, some of which will be filled out by the sender, and some will be filled out by the email server automatically. The protocol has no requirements on the body, which is also similar to a hand-written letter.

There are over 20 standard fields in the header set by the RFC822 protocol. Among these 20+ fields, most are optional except From, To, Date, and a few other fields. Therefore, the fields contained in the headers of different emails may be very different. Table I gives a list of the standard fields frequently seen in email headers.

Besides the standard fields, RFC822 also allows email servers to use non-standard and customized fields. These fields usually start with X-, such as X-Mailer, X-Originating-IP, and so on. Some of these fields customized by email servers have also a high value. For example, X-Originating-IP records the original IP address of the sender where he/she sends out an email. If Outlook, Foxmail, or other email clients are used to edit and send an email, then in most cases this IP address will be the real IP address of the sender.

We can obtain a lot of information by analyzing email headers. With these information, we will be able to conduct research work in many aspects. For example, we can build the social network of email users, find out communities, and locate key users based on the recipient and the sender; we can also recognize and filter out spam emails based on the sender, email subject, sending route, digital signature, and so on [12]-[14].

TABLE I: STANDARD FIELDS FREQUENTLY SEEN IN EMAIL HEADERS

Field	Description
To	The recipient's email address
From	The sender's email address
Date	Email sending time
CC	Copy to
Subject	Email subject
Received	Sending route
Return-Path	Address to be replied to
Delivered-To	Address to be delivered to
Message-Id	Message ID
MIME-Version	MIME version
Content-Type	Content type

III. USERS' PERSONAL PRIVACY PROTECTION

Compared with email bodies, email headers involve much less personal private information. Users are mainly concerned with the sender's email address, the recipient's email address, and the subject in an email header. Because the sender's email address and the recipient's email address are two key data for the corpus of an email header and have the most value, we need to collect them in our corpus of email headers. In order to protect users' personal privacy, we can use an encryption algorithm to hide real email addresses. Apart from ruling out the possibility of decryption after data are encrypted, we need also to consider the one-to-one correlation between the real addresses and the corresponding encrypted email addresses. The subject is crucial to recognition of spam emails and other researches, however, in order to protect users' privacy, we will encrypt it.

In order to collect email headers with protection of users' privacy, we use Secure Hash Algorithm (SHA) to by encrypting sensitive data in email headers. SHA is a family of cryptographic hash functions designed by the United States National Security Agency and published by the National Institute of Standards and Technology. The versions include SHA-1, SHA-224, SHA-256, SHA-384, SHA-512, etc.[15] SHA has two major features. One is that no recovery can be made after encryption. This means that encrypted texts can't be decrypted to get original plain texts. Second, two different messages wouldn't have identical encrypted text. Our mail header collection work consists of the following two steps.

Step 1: In the data collection stage, instead of obtaining complete email headers, we extract key information to generate key-value pairs of the fields in email headers, i.e., in the "field name: field value" format. For any key-value pair, the field name is the standard field name defined in RFC822, and the field value is users' own data. The field name and the

field value are separated, so that the field value can be processed later if necessary.

Step 2: Encrypt users' sensitive data such as recipient's mail addresses and sender's email addresses by use of SHA. After the recipient's email addresses and the sender's email address are encrypted by SHA, the email addresses will turn into unknown codes. Moreover, the features of the SHA encryption method can guarantee that each email address would always have the same code after it is encrypted. In this way, on the one hand, users' personal privacy can be protected, and on the other hand, data usability can be retained.

	A	B	C	D	E	F	G
1	From	To	CC	Date	MessageID	ReceivedFroml	XOriginating
2	235821671202338419411210548162255210246128729812772224553719712011050862112191179913 9@126.com	18348134141836		mon, 22 may 201	122239248292411	06.8.167.116	[106.8.167.11
3	234114136642111132251658624530876563232145321223719324773182247111207614362772022 21@service-paper.org	18348134141836		mon, 15 aug 201	131901282261701	23.187.112.100	[123.187.112.
4	245122199101657812015316925114524418612360644813520016412452162781581281524521402 36@conferences-service.org	18348134141836		sun, 14 aug 2011	7115819622208 1	23.187.106.36	[123.187.106.
5	250103619724614817115729314213510997120133150228102634184108179461014932486721024@yeah.net	18348134141836		mon, 13 mar 201	99352021159240	171.43.207.138	[171.43.207.1
6	179781727461142219413415987126247977109595240161221235818212718125361481632 44@etp-pub.org	18348134141836		wed, 11 jan 2012	4332712491429 6	59.174.198.203	[59.174.198.2
7	492531016162971382117052461124949144170228568115414878232151562552287512439303 2@qq.com	18348134141836		tue, 23 may 2017	132110102455 04	183.60.61.231	27.186.12.212
8	115234220133119105196388314355234155871751719710519129199142381238812625021182877 2181@126.com	18348134141836		thu, 17 nov 2016	162197135184 53	106.8.160.202	[106.8.160.20
9	164171481411215922893391552548997245145211054822310310110123217822415374419917726@xaut.edu.cn	51151021972071		wed, 21 sep 201	112183631012 41	127.0.0.1	
10	198771492492152111991752947147104923620613173148341969920119019412686226184248232246188@126.com	18348134141836		fri, 16 dec 2016	169174761297 91	106.8.175.15	[106.8.175.15
11	332415411810717561510484252186651537728343104842210702763120238823364 43@163.com	18348134141836		tue, 14 jun 2016	161130212221 17	106.8.163.235	[106.8.163.23

Fig. 2. An example file of email header submitted by a volunteer.

IV. Data Collection and Cleansing

A. Data Collection Method

In usual, people will prefer to take things out from their pockets and give them to others rather than allow others to find things in their pockets and take them away. Likewise, compared with reading users' email data directly, we find that it is much more acceptable for users to provide email data by themselves. Users will feel more reassured if they can distinctly check the data which they will submit, and delete the data which they consider as sensitive before they submit. For this reason, we designed a system for extracting email headers and provided it to volunteers for processing their own email data. The system will store the extracted email headers in a file, which can be viewed by the corresponding volunteer; the volunteer can delete any data in the file. After confirming all data, the volunteer will submit the file to us. Fig. 2 shows an example file of email headers submitted by a volunteer.

```
Delivered-To: tom2009@gmail.com
Received: by 10.103.30.2 with SMTP id e2csp434621vse;
        Fri, 8 Jul 2016 08:42:50 -0700 (PDT)
Return-Path: <owner-cistc@comsoc.org>
Received: from mxw1.secure-premium.ne.jp (mxw1.secure-premium.ne.jp.
[220.152.127.81])
        by mx.google.com with ESMTP id 167si3179407iol.225.2016.07.08.08.42.49
        for <tom2009@gmail.com>;
        Fri, 08 Jul 2016 08:42:50 -0700 (PDT)
Received-SPF: softfail (google.com: domain of transitioning
owner-cistc@comsoc.org does not designate 220.152.127.81 as permitted sender)
client-ip=220.152.127.81;
Authentication-Results: mx.google.com;
       spf=softfail (google.com: domain of transitioning owner-cistc@comsoc.org
       does not designate 220.152.127.81 as permitted sender)
       smtp.mailfrom=owner-cistc@comsoc.org
Received: from spw-cm15 ([10.5.131.210])
    by cmsmtp with SMTP
    id LXWZbR8ZodErMLXvVbKj5I; Sat, 09 Jul 2016 00:42:49 +0900
Received: from spw.secure-premium.ne.jp ([10.5.131.211])
    by spw-cm15 with bizsmtp
    id GFip1t0024ZocQi01FippQ; Sat, 09 Jul 2016 00:42:49 +0900
X-CNFS-Analysis: v=2.1 cv=AtHXgoNP c=1 sm=1 tr=0
 a=km6sk2UkvDyct+lopFdqGQ==:117 a=AIB8JnPVVTBQfK+9lNyiwA==:17
 a=L9H7d07YOLsA:10 a=9cW_t1CCXrUA:10 a=s5ivqZ67dGcA:10 a=HrRcR-OLtcoA:10
```

Fig. 3. The received field of an email header.

B. Data Collection Object

Because there are optional standard fields and customized fields in email headers, there might be different fields contained in the email headers sent from different email servers. The email data in the corpus finally built in this paper come from 51,967 different email servers. Among them, 44,705 email servers contributed less than 10 emails. Because it is difficult to consider all email servers, in this paper, we only selected part of the key and standard fields and some important customized fields described in the RFC822 protocol. Table II gives a list of these fields and the corresponding processing methods.

The "Received" field records the detailed history of an email being transmitted on the Internet. The email shown in Fig. 3 has up to 23 entries in its "Received" field. Because these data are created by email servers, and senders cannot forget them, they are very useful in the anti-spam field (reverse DNS resolution). Therefore, "Received" field is a very important field.

During data collection, every "Received" field added by email servers will be split into two parts: "From" and "By". All IP addresses in the "From" part will be extracted and stored separately.

TABLE II: Header Fields Used in this Paper and Their Processing Method

Field	Processing Method	Field	Processing Method
To	SHA encryption	Return-Path	SHA encryption
From	SHA encryption	Delivered-To	SHA encryption
Date	Unchanged	Message-Id	Unchanged
CC	SHA encryption	MIME-Version	Unchanged
Subject	SHA encryption	Content-Type	Unchanged
Received	Split & combine	X-Mailer	Unchanged
X-Originating-IP	Unchanged	DKIM-Signature	Unchanged

The whole information in the "Received" field of an email header will be finally combined into six key-value pairs. "ReceivedFromIP", "ReceivedFrom" are extracted from the first "Received" field, "ReceivedBy" is extracted from the last "Received" field, "ReceivedFromIPList",

"ReceivedFromList", and "ReceivedByList" record the values of the complete history record of "FromIP", "From", and "By".

C. Data Cleansing and Organizing

Because spam emails may intentionally hide or forge information, and not all email servers are managed in a standard way, we may meet various problems when extracting email headers. To guarantee data quality in our corpus, it is necessary to cleanse and organize the raw data submitted by volunteers. For example, we need to cleanse "From:=?UTF-8?B?5LqO5re8LeS4remrmOaVmQ==?= <yumiao_zgj05@163.com>"into"From:yumiao_zgj05@163.com".

Moreover, we often encounter invalid data during data collection. For example, "From" is normally the sender's email address, "To" is the recipient's email address, and both of the two fields should be found in any email header. However, in some email headers, the two fields might not exist, be empty or be invalid. In such cases, we need to extract "Return-Path" and "Delivered-To" fields. "Return-Path" is the reply address of an email; thus, if necessary, we can use its value as a substitute for "From" field value. "Delivered-To field" is the recipient's mail address added by the email sender server. In the cases where both "To" field and "Delivered-To" field have a value, the value of "Delivered-To" field is usually more reliable. Therefore, at the data cleansing stage, we use the "Delivered-To" field value as the "To" field value if the ""From"" is valid.

On the other hand, as the example shown in Fig. 4, In the case where both the "From" field value and the "To" field value in the email header are invalid, and there is neither Return-Path" field nor "Delivered-To" field, we will give up the email header data.

```
From:网易邮件中心 <netease@ntes>
To:126邮箱用户
Subject: =?gb2312?B?MTI2y+bJ7dPK1tzE6sfsLNXmx+m72MChMTAwJdbQvbi
Reply-to:kf@service.netease.com
Content-type:text/html

<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=gbl
<title>126随身邮周年抽奖活动</title>
</head>
<body>
<style type="text/css">
```

Fig. 4. An example of invalid email header.

V. Data Labeling

A. Labeling of Emails with More Than One Recipient

"To", "CC", and "Delivered-To" fields in an email header are all recipient's email addresses. Normally an email header would have "To" field, which contains at least one recipient's email address. When someone is sent a copy, the email header would also have "CC" field, which possibly contains one or more recipients' email addresses. Among all emails that we have collected, the maximum number of recipients' email addresses in "To" field of an email is 1029. For convenience to be used in the future, in our corpus, emails are labeled according to their number of recipient emails that we have collected, the maximum number of recipients' email addresses in "To" field of an email is 1029. For convenience to be used in the future, in our corpus, emails are labeled according to their number of recipient addresses. The labeling method for each email header is as follows:

1) Extract "To", "CC" and "Delivered-To" fields in the email header;

2) Remove duplicates from "To", "CC" and "Delivered-To" fields, and then combine them to "SendToList";

3) If there is only one recipient's email address in "SendToList" field, copy the "SendToList" field value to the "To" field value and label the identifier field as -1;

4) Otherwise, i.e., if there is more than one recipient's email address in "SendToList" field, then place the recipients' email addresses (suppose whose number is n) into a character string array. Then traverse the array, and copy the array contents (i.e., recipients' email addresses) to the "To" field value. Therefore, we obtain n corpus data through a mail header file, the identifier field values of these n corpus date will be labeled by 0, 1, 2, ..., n-1, respectively.

ID	From	To	CC	DeliveredTo	Flag
1	F	T			-1
2	F	D		D	-1
3	F	D		D	0
4	F	T			1
5	F	D	C1, C2, C3	D	0
6	F	T1	C1, C2, C3	D	1
7	F	T2	C1, C2, C3	D	2
8	F	C1	C1, C2, C3	D	3
9	F	C2	C1, C2, C3	D	4
10	F	C3	C1, C2, C3	D	5

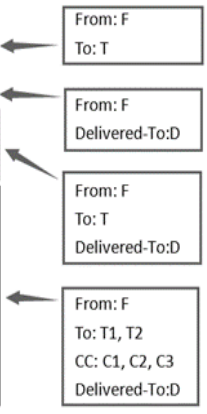

Fig. 5. Four email headers and the corresponding results after labeling.

Fig. 5 shows four mail headers and the corresponding results after labeling. H1-H4 represent the "From", "To", "CC" and "Delivered-To" fields in the four email headers shown on the right side of the figure, and the table on the left is the result after labeling.

B. Labeling of Email Sending/Receiving Addresses

Usually, all emails will go through multiple email servers on the Internet before they are finally delivered to their recipients. Each email server through which an email goes will add a "Received" message to the email's header, and the message is added to the beginning of the "Received" message added by previous email servers. Therefore, we can acquire the sending/receiving address of an email by analyzing its

"Received" field. The first "Received" message shows the information of the recipient's email server; and the last "Received" message shows the information of the sender's email server. These addresses are not the real address of the sender or recipient, but the addresses of the email servers they used. Some email servers will have the "X-Originating-IP" field among their customized fields. The address recorded in this field is usually the real IP address of the sender. Therefore, in this paper, the "X-Originating-IP" field is also extracted, which is used to validate the sender's IP address. We can know rough geographical locations of the recipient and the sender through these addresses. Fig. 6 shows an example of how to get the IP addresses of sender and recipient, Fig. 7 shows the result of email sending/receiving addresses after labeling.

C. Labeling of Users' Social Attributes

As we have known, by our proposed method, the recipient, the sender, and other sensitive information in the corpus will be encrypted, thus we will not be able to know any social attributes of users providing these data. Thus, the usability of the corpus will be greatly reduced. In order to obtain the social attributes of relevant data, such as their country/region, age, language, job nature, professional fields, etc., we designed a questionnaire for volunteers about their social attributes, which will be submitted to the corpus along with their mail headers. Because all sensitive data will be encrypted, it is easy to obtain the cooperation of volunteers.

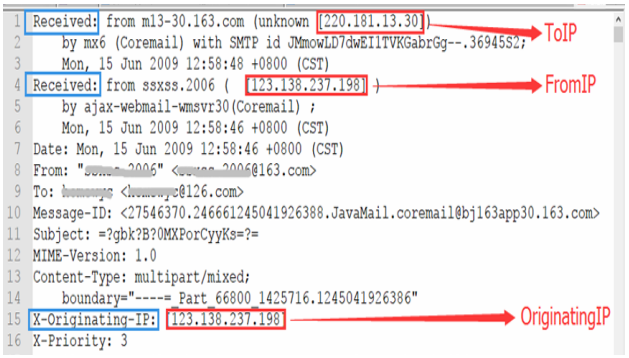

Fig. 6. An example of how to obtain the IP address from an email header.

XOriginatingIP	FromMailServer	FromIP	FromCounry	FromLocal	ToMailServer	ToIP	ToCountry	ToLocal
NULL	163.com	221.11.46.238	China	Shanxi	126.com	123.125.50.134	China	Beijing
NULL	sonyericssonmail.com	192.16.134.98	Sweden	Stockholms lan	126.com	192.16.134.98	Sweden	Stockholms lan
NULL	sendmail.dangdang.com	172.16.240.97	-	-	126.com	203.81.26.197	China	Beijing
NULL	dropio.com	207.97.240.173	United States	Virginia	126.com	207.97.240.173	United States	Virginia
NULL	qq.com	58.221.29.25	China	Jiangsu	126.com	114.80.210.231	China	Shanghai
NULL	staff.sina.com.cn	202.108.33.40	China	Beijing	126.com	202.108.3.222	China	Beijing
NULL	message.myspace.cn	10.60.0.74	-	-	126.com	60.28.201.201	China	Tianjin
123.138.237.198	163.com	123.138.237...	China	Shaanxi	126.com	220.181.13.30	China	Beijing
124.114.80.47	qq.com	119.147.10.250	China	Guangdong	126.com	119.147.10.250	China	Guangdong
NULL	126.com	221.11.46.238	China	Shanxi	126.com	221.11.46.238	China	Shanxi

Fig. 7. Labeling result of some email sending/receiving addresses.

Currently, the volunteers of the corpus are mainly from the Europe & America, China, and Japan. Their language types are mainly English and Chinese. Their jobs are mainly colleague teachers, university students, company employees, and research agency workers. Their professional fields are mainly computer technologies, IT technologies, other technologies, and commercial trade.

VI. Conclusion

In this paper, we have proposed a method for constructing a corpus of email headers with personal privacy protection. By use of the method, we have collected nearly 200,000 pieces of email data within a short time and built the corpus of email headers. This proved that the method is effective and reliable. The method for building a corpus of email headers proposed in this paper can be applied to other corpus data collection work where users' privacy protection is necessary.

The corpus of email headers built in this paper meets not only the needs of research fields such as community discovery and user relationship exploration, but also the needs of research fields such as email classification and email-header-based recognition of spam emails.

For future work, we will enrich the corpus of email headers we have built, and then conduct researches on it, such as the characteristics of e-mail social network, the relationship between e-mail received by users and their social attributes, spam identification and filtering.

References

[1] T. Kathirvalavakumar, K. Kavitha, and R. Palaniappan, "Efficient harmful email identification using neural network," *Fems Microbiology Ecology*, vol. 7, no. 1, pp. 58-67, 2015.

[2] J. Sheu, K. T. Chu, N. Li, and C. Lee, "An efficient incremental learning mechanism for tracking concept drift in spam filtering," *Plos One*, vol. 12, no. 2.

[3] E. W. Fox, M. B. Short, F. P. Schoenberg, K. D. Coronges, and A. L. Bertozzi, "Modeling E-mail networks and inferring leadership using self-exciting point processes," *Journal of the American Statistical Association*, vol. 111, no. 514, pp.564-584, 2016.

[4] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research. European conference on machine learning," *Springer-Verlag*, vol. 3201, pp. 217-226. 2004.
[5] P. Liu and T. Moh. "Content based spam e-mail filtering," in *Proc.* 2016 *International Conference on Collaboration Technologies and Systems (CTS)*, 2016, pp. 218-224.
[6] T. Schmitz and D. Jannach, "Finding errors in the enron spreadsheet corpus," *Visual Languages and Human-Centric Computing*, pp. 157-161, 2016.
[7] R. C. Nurse, A. Erola, M. Goldsmith, and S. Creese, "Investigating the leakage of sensitive personal and organisational information in email headers," *Journal of Internet Services and Information Security*, vol. 5, no. 1, pp. 70-84, 2015.
[8] C. Bird, A. Gourley, and P. Devanbu, "Mining email social networks", in *Proc. the 2006 International Workshop on Mining Software Repositories*, 2006, pp. 137-143.
[9] O. Al-Jarrah, I. Khater, and B. Al-Duwairi, "Identifying potentially useful email header features for email spam filtering," presented at the Sixth International Conference on Digital Society, 2012.
[10] I. Hamid, J. Abawajy, and T. Kim, "Using feature selection and classification scheme for automating phishing email detection," *Studies in Informatics and Control*, vol. 22, no. 1, pp. 61-70, 2013.
[11] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63-92, 2008.
[12] H, Guo, B. Jin, and W. Qian, "Analysis of email header for forensics purpose," presented at the Communication Systems and Network Technologies, IEEE, pp. 340-344, 2013.
[13] D. Wang, D. Irani, and C. Pu, "A study on evolution of email spam over fifteen years," presented at the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 1-10, 2013.
[14] Y. N. Liu, Y. Han, X. D. Zhu, F. He, and L. Y. Wei, "An expanded feature extraction of e-mail header for spam recognition," *Advanced Materials Research*, vol. 846, pp. 1672-1675, 2014.
[15] C. H. Lin, Y. S. Yeh, S. P. Chien, C. Y. Lee, and H. S. Chien, "Generalized secure hash algorithm: SHA-X," in *Proc. International Conference on Computer As a Tool*, 2011, pp. 1-4.

Yongchao Wang is a doctoral student of the Faculty of Information Science and Technology, Aichi Prefectural University. He received the B.E degree and the M.S. degrees from Xi'an University of Technology in 2002 and 2007.

Since 2009, he has been a lecturer at the School of Computer Science and Engineering of Xi'an University of Technology. His research interests include information science, intelligent image processing, pattern recognition, and artificial intelligence.

Xiao Zhao received the B.E. and M.S. degrees from the Shaanxi University of Science and Technology, China, in 2001 and 2006, respectively. From 2001 to 2006, she was an assistant professor at the College of Electrical and Information Engineering, Shaanxi University of Science and Technology.

Since 2007, she has been a lecturer. Her research interests include image processing, artificial intelligence, pattern recognition, and string searching.

Feihang Ge is a doctoral student of the Faculty of Information Science and Technology, Aichi Prefectural University. He received the B.E. degree and the M.S. degrees from Xi'an University of Technology in 2002 and 2005.

Since 2012, has been a lecturer at the Zhejiang College of Construction. His research interests include information science, intelligent image processing, pattern recognition, and artificial intelligence.

Yuyan Chao received the B.E. degree from the Northwest Institute of Light Industry, China, in 1984, and the M.S. and Ph.D. degrees from Nagoya University, Japan, in 1997 and 2000, respectively. From 2000 to 2002, she was a special foreign researcher of the Japan Society for the promotion of science with the Nagoya Institute of Technology. She is currently a professor at Nagoya Sangyo University, Japan, and a guest professor at the Shaanxi University of Science and Technology, China. Her research interests include image processing, graphic understanding, computer aided design, pattern recognition, and automated reasoning.

Lifeng He received the B.E. degree from the Northwest Institute of Light Industry, China, in 1982, a second B.E. degree from Xi'an Jiaotong University, China, in 1986, and the M.S. and Ph.D. degrees in artificial intelligence and computer science from the Nagoya Institute of Technology, Japan, in 1994 and 1997, respectively. From 2006 to 2007, he was with the University of Chicago as a research associate. He is currently a guest professor with the Shaanxi University of Science and Technology, China, and a professor with Aichi Prefectural University, Japan. His research interests include intelligent image processing, computer vision, automated reasoning, pattern recognition, string searching, and artificial intelligence.

He is a member of the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers, and the Association of Authors' Representatives. He has been serving as a referee of over 10 journals in computer science, including the IEEE Transaction On Neural Networks, The IEEE Transaction On Pattern Analysis and Machine Intelligence, The IEEE Transaction on Image Processing, The IEEE Transaction on Computers, Pattern Recognition, Computer Vision, Image Understanding, and Pattern Recognition Letters.