Mailing List Characteristic from Electronic Mail

N. Khaitiyakun and A. Khunkitti

Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang , Bangkok 10520 , Thailland
Email: hanuhana@yahoo.com and akharin@it.kmitl.ac.th

Abstract: Principle of mailing list was distributed messages to all subscribers in one time. But mailing list operation has constructed a network traffic problem. Because mailing list manager distributed mails without concentrate on subscriber network. If our network has many of subscribers, there will be redundant data in traffic channel. Submailing list has purpose to reduce problems. Analyses of mailing list characteristic in electronic mail were a feature of submailing list system, which manage by human hand (Network Administrator). That will cause trouble for network traffic if Network Administrator could not seek for mailing list characteristic from e-mails in due time. This article will present ideas and recognize methodology for automatic working in submailing list system. Recognize step begin with capture process, which use to trap e-mail information from transfer channel. Next process is preparing raw data into recognition format. Then the third one is recognize part and find out confidential factor. The last process is make decision and determine which electronic mail has properties of mailing list characteristic. Afterward deliver result to submailing list for carry on.

Keywords: Mailing list system, Recognized system, Confidential Factor

1. INTRODUCTION

Mailing list [1] is a group of e-mail addresses that can all be reached by sending a single message to one address, List address. Subscribers can have discussion by sending message to the list address. Each message will be distribute to all list subscribers. First advantage of mailing list is distributing information from a central source to lots of people in once time and second is discussing a project among several participants. Third one is exchanging questions and answers with other users of a product or service. Disadvantage of mailing list is data redundancies when mailing list distributed message to subscribers who live in the same network or neighbor networks. Submailing list [2] was created to support this trouble. By convert all subscribers to submailing list member and subscribe itself to main mailing list. When main mailing list has to transmit messages, there will be only one copy has sent to our network.

Before start recognition system we have to familiar with mailing list behavior that shown below.

- There are mailing list manager commands in Subject: header or mail body.
- Mailing list manager has to protect subscribers from communication error mails. By create e-mail address for receive error report. At the same time for general e-mail, error report will send to e-mail address found in From: header.
- There are mailing list headers [3].
- Normally we could find e-mail address of receiver from Recipient header in e-mail or RCPT TO command in smtp conversation. In general case, if found e-mail address in RCPT TO command we also found the same address in Recipient header too. But in mailing list case there are irregular between recipient.

All cases above have to consider together and then calculate confidential factor (CF) for use to decide which information was suitable to configured at submailing list system. Next we mention about recognition methodology and explain recognize idea. Follows by implement and display algorithm of recognize process. Last topic is talked about conclusion and future work.

2. METHODOLOGY OF RECOGNITION SYSTEM

Recognition system was created for inspect mailing list characteristic from e-mail [4]. Which can summarize in four processes. First process was trapping e-mail data, Capture process. These are conversation between receiver and sender MTA that based on SMTP (Simple Mail Transfer Protocol) [5]. Then rearrange data to recognize format, Preprocessing process. Next recognized process has to find out mailing list character and calculate confidential factor. Last one Postprocessing process has to make decision and hand up to configure at submailing list system.

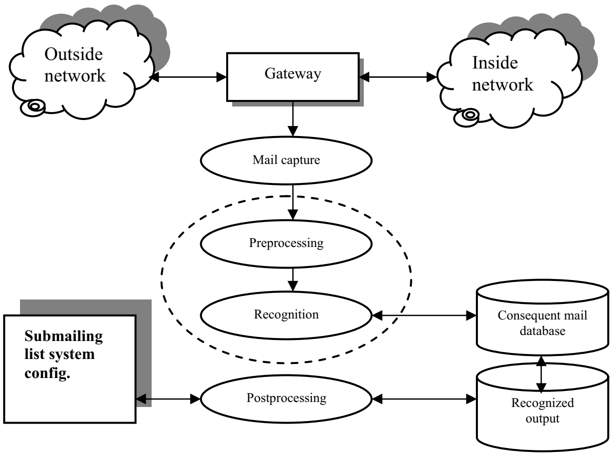

Fig. 1 Mailing list Recognition System

2.1 Capture process

The first process of recognition system is capture process, which used to capture mail information. Start process when there are some data enter SMTP port (port 25). The raw data can separate in 2 type, SMTP envelope and SMTP content. In SMTP envelope consists of SMTP command and SMTP reply code that is exchange between receiver and sender MTA. Next SMTP content consists of Mail header and Mail body that based on standard of RFC822. All of raw data are in text format.

```
220 dipac.it.kmitl.ac.th ESMTP Sendmail 8.12.8/8.12.8; Mon, 10 Nov 2003 10:34:21 +0700 (ICT)
EHLO system.dipac.it.kmitl.ac.th
250-dipac.it.kmitl.ac.th Hello system [192.168.1.11], pleased to meet you
MAIL From:<root@system.dipac.it.kmitl.ac.th> SIZE=476
250 2.1.0 <root@system.dipac.it.kmitl.ac.th>... Sender ok
RCPT To:<hanu@dipac.it.kmitl.ac.th>
DATA
250 2.1.5 <hanu@dipac.it.kmitl.ac.th>... Recipient ok
354 Enter mail, end with "." on a line by itself
Received: from system.dipac.it.kmitl.ac.th (localhost.localdomain [127.0.0.1])
        by system.dipac.it.kmitl.ac.th (8.12.5/8.12.5) with ESMTP id hAA3Thov001035
        for <hanu@dipac.it.kmitl.ac.th>; Mon, 10 Nov 2003 10:29:44 +0700
Received: from localhost (root@localhost)
        by system.dipac.it.kmitl.ac.th (8.12.5/8.12.5/Submit) with ESMTP id hAA3ThoO001031
        for <hanu@dipac.it.kmitl.ac.th>; Mon, 10 Nov 2003 10:29:43 +0700
Date: Mon, 10 Nov 2003 10:29:43 +0700 (ICT)
From: root <root@system.dipac.it.kmitl.ac.th>
To: hanu@dipac.it.kmitl.ac.th
Subject: test
Message-ID: <Pine.LNX.4.44.0311101029240.1030-100000@system.dipac.it.kmitl.ac.th>
MIME-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII

test


.
250 2.0.0 hAA3YLmm068646 Message accepted for delivery
QUIT
221 2.0.0 dipac.it.kmitl.ac.th closing connection
```

Fig. 2 An example of capturing data from Ethereal

2.2 Preprocessing process

This step was data prepared process. Before system started recognition process, the raw material from capturing process should be rearranged in recognize format. Because there are many MTA in cyber world so data that pass capturing process may be have different layout. That was a reason we should make it in the same pattern.

Table 1 Recognize variables discover in Recognized process

Variable	Format	Type
Smtp-mail-from	<name@domain>	single
Smtp-recipient	<name@domain>	multi
Rfc822-from	<name@domain>	single
Rfc822-sender	<name@domain>	single
Rfc822-reply-to	<name@domain>	multi
Rfc822-return-path	<name@domain>	single
Rfc822-recipient	<name@domain>	multi
Rfc822-error-to	<name@domain>	single
Rfc822-subject	<text>	multi
Rfc822-body	<text>	Multi

From table 1, recognize variables could clarify in 2 types, Smtp-variable and Rfc822-variable. Smtp-variable obtains from smtp command and smtp reply code (smtp envelope). The smtp information showed about sender address and receiver address, which came from MTA communication. Rfc822-variable obtains from mail header and mail body (smtp content). The Rfc822 information presented e-mail header and mail body which came from human communication. Such as Rfc822-from which store e-mail address found in From header. Each variable, establish in variables format with angle bracket. And each of them contains data that could be single value or multi value. There are another variables that appear in preprocessing process. That use to configure in submailing list system. After this step the data was sent to next process.

2.3 Recognized process

Recognize processing is main process in mailing list recognition system. Here the system had to recognize electronic mail. That means the system find out mailing list characteristic by use observations found above for recognize part and calculates confidential factor (CF).

Before we mention about recognized algorithm, we should early present all database that use to keep information. First one was recognized output database (RODB) that use to keep result from recognized process. Recognized output database consists of 4 tables below.

Table 2 List Address table (LA table)

List address	CF_LA
Listname@listdomain	Value between 0-1

Table 3 List Member Address table (LM table)

List address	Member address	CF_LM
Listname@listdomain	Membername@memberdomain	Value between 0-1

Table 4 List Manager Address table (MA table)

List address	Manager address	CF_MA
Listname@listdomain	Managername@managerdomain	Value between 0-1

Table 5 List Manager Type table (MT table)

Manager address	Manager type	CF_MT
Managername@managerdomain	Manual, listserv, majordomo, listproc, smartlist, N/A	Value between 0-1

Confidential factor (CF) presents possibility of mailing list characteristic in each data record. There has value between zero and one. The most CF came from measured the experiment several times.

Second one was consequent mail database (CMDB). In this database has function like temporary file, which keep mailing list data that could not conclude. Because this mailing list data need more than information to prove and confirm their situation. Each record in CMDB also has CF, which shown possibility value. If we found another data that contain mailing list characteristic as same as record in CMDB, the CF will increase. On the other hand if we didn't have new information that agrees with our data for along time, the CF will decrease. Data record in CMDB could transfer to RODB by compare CMDB confidential factor with threshold value ($CF_{threshold}$).

Besides recognized output database and consequent mail database, there was recognized configuration database (RCDB). That contains information about mailing list manager characteristic such as mailing list manager type, mailing list manager command, mailing list command format etc. The recognized configuration database consists of 3 tables; list command manager table, command keyword table and command pattern table. Next process we will talk about determination process.

Table 6 Suspected List Address table (SLA table)

List address	CF_SLA	Time stamp	Status	Mcount	MID
Listname@listdomain	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 7 Suspected List Member Address table (SLM table)

List address	Member address	CF_SLM	Time stamp	Status	Mcount	MID
Listname@listdomain	Membername @memberdomain	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 8 Suspected List Manager Address table (SMA table)

List address	Manager address	CF_SMA	Time stamp	Status	Mcount	MID
Listname@listdomain	Managername @managerdomain	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 9 Suspected List Manager Type table (SMT table)

Manager address	Manager type	CF_SMT	Time stamp	Status	Mcount	MID
Managername @managerdomain	Manual, listserv, majordomo, listproc, smartlist, N/A	Value between 0-1	Arrive time	<subscribe>, <unsubscribe>, <general>	Counter	Mail sequence ID

Table 10 List Manager Command table

Type ID	Manager type	Command part flag	Subscribe format ID list	Unsubscribe format ID list	General format ID list
Sequence ID	Manual, listserv, majordomo, listproc, smartlist, N/A	Bit 0 = Rfc822-body Bit 1 = Rfc822-subject	(<keyword>,<pattern>)	(<keyword>,<pattern>)	(<keyword>,<pattern>)

2.4 Postprocessing process

In last process is decided and consider step. Which information suitable for submailing list system. By reconsider confidential factor and accurate mailing list characteristic in RODB again. Then send data record to configure at submailing list system. If found some mistake remove record from RODB to CMDB for recheck.

All of processes have to decrease responsibility of network administer. That duty is checking for find mailing list mail and gives information to configure submailing list system. But this is a heavy job because there are many mails come to network. If network administers cannot work in time that will cause problem in our traffic to Internet. So mailing list recognition system has to create for support this situation as present before.

3. Algorithm of Recognized Process

In recognized process could divide mailing list characteristic in 2 types, 'manager channel characteristic' and 'list channel characteristic'. For manager channel characteristic means e-mail that communicate between mailing list manager and user (subscriber). The recognized process has to use information from RCDB for recognized. For list channel characteristic means e-mail that communicate among subscriber. The recognized process use mailing list behavior to analyze. In figure 3 was presented recognized process algorithm.

Anyway there are other algorithms that corporate with recognized process such as confidential factor algorithm. Which use for calculate CF value, that consist of 2 part positive and negative algorithm. In positive algorithm use to increase CF value when the system found support reason. In the opposite way negative algorithm use to decrease CF value when the system found another reasons to disprove.

```
/* Recognized algorithm */
if Rfc822-mailing-list-header not equal to null then
    if Rfc822-recipient in SLA table then
        increase CF_SLA and Mcount = Mcount+1
        if Rfc822-from in SLM table then
            increase CF_SLM
        else
            add member address by Rfc822-from in SLM table
            increase CF_SLM
    else
    create entry in LA table by list address = Rfc822-recipient
    and CF_LA = initial CF
    create entry in LM table by member address = Rfc822-from
    and CF_LM = initial CF
else
    if Rfc822-recipient in SLA table then
        decrease CF_SLA

if Rfc822-from not equal to (Smtp-mail-from or Rfc822-sender or
Rfc822-return-path or Rfc822-error-to or Rfc822-reply-to) then
    for (i = 1 to M) /*M is number of Rfc822-recipient*/
        for (j = 1 to N) /*N is number of Smtp-recipient*/
            if Smtp-recipient[j] == Rfc822-recipient[i] then
                x = x+1
  if (x==0) or (x<N) then /*each Smtp-recipient not match
  Rfc822-recipient*/
        if Rfc822-recipient in SLA table then
            increase CF_SLA and Mcount = Mcount+1
            if Rfc822-from in SLM table then
                increase CF_SLM
            else
                add member address by Rfc822-from in SLM
  table
                increase CF_SLM
        else
      create entry in LA table by list address = Rfc822-recipient
      and CF_LA = initial CF
      create entry in LM table by member address = Rfc822-
      from and CF_LM = initial CF
else
    if Rfc822-recipient in SLA then
        decrease CF_SLA

if (Rfc822-subject or Rfc822-body) in list manager command table
then
    if (Rfc822-recipient or Rfc822-from) in SMA table then
        if Rfc822-from in SLM table then
            increase CF_SLM and CF_SMA
    else
    create entry in LA table by list address =
    listname+@+managerdomaibn and CF_LA = initial CF
    create entry in MA table by manager address = Rfc822-
    recipient and CF_MA = initial CF
    create entry in MT table by manager type = Rfc822-
    reicipient name or manager type and CF_MT = initial MT
else
    if (Rfc822-recipient or Rfc822-from) in SMA table then
        if Rfc822-from in SLM table then
            decrease CF_SLM and CF_SMA
```

Fig. 3 Algorithm of recognized process

```
/*Confidential factor algorithm*/

if increase CF
        CF_x = CF_x + positive(CF_x)
If decrease CF
        CF_X = CF_x + negative(CF_x)

/*Positive CF algorithm*/
if receive CF_x then
        positive CF = (1 – CF_x)/2
        return positive CF

/*Negative CF algorithm*/
if receive CF_x then
        negative CF = CF-x / 2
        return negative CF
```

Fig. 4 Algorithm of confidential factor

4. Conclusion and Future Work

Truly mailing list manager (MLM) software perceives the network traffic troubles from mailing list operation. Some of MLM try to solve the problems but mailing list was consisting of many network systems so they could not wholly correct. Submailing list was another way to decrease data redundancies in network traffic. However if submailing list system could automatically work, there will make higher performance for network traffic.

From above concept, that would like to make submailing list system automatically learning mailing list characteristic and separate which e-mail come from mailing list system. This paper would like to present mailing list recognition system for recognize mail and give useful information to submailing list system. Begin with trap electronic mail from transfer channel and make it in recognize format. Then use assumption found in mailing list behavior to recognize message. And also find out confidential factor for indicate mailing list characteristic. Confidential factor could be able to swap by discover positive reason to support or negative case to conflict old data. When finishes there are also having revise process for make decision and decide which data suitable for submailing list operational. In addition we can solve problem about Spam mail, which extremely found in network traffic problems, by improve recognize process to focus on Spam mail behavior.

Future work we have to find out constant value by testing in real situation. Bring the result to improve and modify mailing list recognition system for high efficient and consistency of the system.

REFERENCES

[1] A. Schwartz, *Managing Mailing List* 1^{st} *ed*, O'reilly & Associates. Inc, March 1996.

[2] S. Chomjan and A. Khunkitti, "A New Hierarchical Based Approach Mailing List System", *Proceedings of The 2003 International Conference on Information and Communication Technologies (ICT2003)*, Bang na Campus, Assumption University, April 8-10, 2003.

[3] G. Neufeld, "The Use of URLs as Meta-Syntax for Core Mail List Commands and their Transport through Message Header Fields", *Request for Comments 2369*, Nisto, July 1998.

[4] D. H. Crocker, "Standard for The Format of Arpa Internet Text Messages", *Request for Comments 822*, Dept. of Engineering University of Delaware, August 13, 1982.

[5] J. B. Postel, "Simple Mail Transfer Protocol", *Request for Comments 821*, Information Sciences Institute, Southern California University, August 1982.