STUDIA INFORMATICA
Nr 1/2(7) Systemy i technologie informacyjne 2006

Mailing Lists Archives Analyzer

Krzysztof Rzecki[1], Maximillian Riegel[2]
[1] Institute of Telecomputing, Cracow University of Technology,
ul. Warszawska 24 E-6, 31–155 Krakow, Poland
[2] Siemens Com Mobile, Networks Standarization, Siemens AG,
Hoffmanstr. 51, D – 81359 München, Germany

Abstract. Article describes chance to explore data hidden in headers of e-mails taken from archive of mailing lists. Scientist part of the article presents a way of transforms information enclosed in Internet resources, explains idea of mailing lists archive and points out knowledge can be taken from. Technical part presents implemented and working system analyzing headers of e-mail messages stored in mailing lists archives. Some example results of this experiment are also given.

Keywords. E-mail header, data analyzing, web mining

1 Introduction

Internet is a huge and cheap space containing valuable information. One problem is to find and browse those information, which usually are hidden between tons of not useful files. Another problem is to fetch and collect those information and marshal them into computer knowledge base. Such base can be then used to extricate information we are interested in.

Picture on the next page (Figure 1) shows some proposal for knowledge transformations way.

- from ***universal knowledge*** – something created by people (home pages), organizations (organizations' pages), simple data base (e.g. e-shops[1]), semi-automatic systems (like web camera[2], weather forecast systems) with continue actualization and systems examining real world,
- during ***publishing process*** there are created ***web pages*** containing text, images, photos, animations, videos, etc.,

[1] Best Buy web shop: http://www.bestbuy.com/
[2] WebCam Central: http://www.camcentral.com/

- mechanism called ***web text processing*** engine takes public available data hidden in web pages and creates and updates ***computer knowledge base*** storing information in convenient to explore form,
- as a result we are able, using ***exploring engine***, to retrieve ***data we are interested in***.

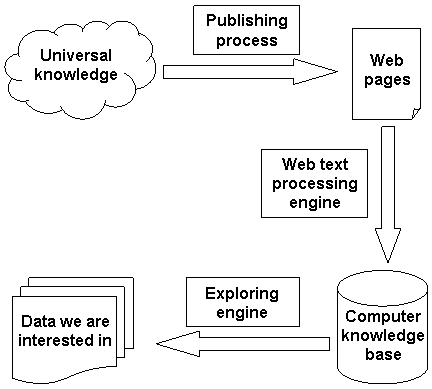

Figure 1. Knowledge transformation way (source: own proposal)

There are also clever systems exploring other bases of knowledge, but those way is just reuse of some existing base.

2 Mailing Lists Archive

One of possibilities to fill in web pages is conversion and storage of mailing lists into web archive. Mailing list is a simple forwarding system where group of people exchange e-mail messages. Basic functionality rely on mechanism, where if one of members sends e-mail to mailing list, all members receive it. Most popular software to establish mailing lists are Majordomo[3], Mailman[4] and Listar[5]. Usually mailing lists are technical or hobby knowledge bases containing of cheap, but dirty (spam, non-sense or non-valuable text, etc.) information.

Web archive of mailing list is made by software converting mail boxes into such web archive. Several open-source and commercial tools are available for this purpose. As example MHonArch[6] (A mail-to-HTML converter), software that allows you to simply create Internet service providing archive of mailing list in easy to browse form.

[3] Majordomo: http://www.greatcircle.com/majordomo/
[4] Mailman: http://www.gnu.org/software/mailman/
[5] Listar: http://www.listar.org/
[6] MHonArch: http://www.mhonarc.org/

There was established project MLA (Mailing Lists Archive) which, using Majordomo and MHonArch, builds up archive of mainly IETF[7] mailing lists, currently with about 200 mailing lists with about half million of e-mail messages. The Internet Engineering Task Force (IETF) is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. The actual technical work of the IETF is done in its working groups, which are organized by topic into several areas (e.g., routing, transport, security, etc.). Much of the work is handled via mailing lists.

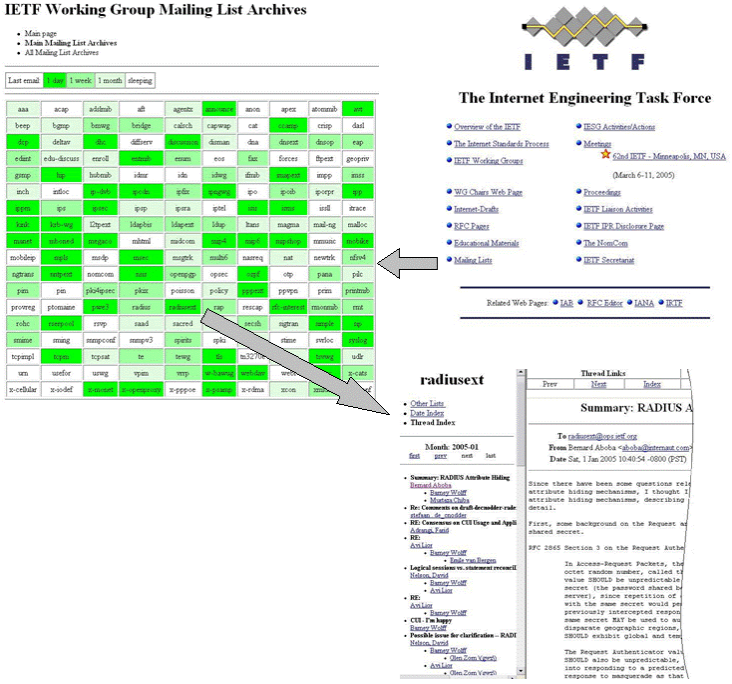

Figure 2. IETF Mailing Lists Archive

MLA system (Figure 2) collects IETF (and some other) archives of mailing lists. On picture above there is screenshot and left side presents access page to all archives. On right-down there is presented one of archives, what looks like nice to browse mail program.

[7] IETF: http://www.ietf.org/ and "*Overview…*"

3 Mailing Lists Archive Analyzer

As extension to Mailing Lists Archive there was made project and implementation of MLAA system (Mailing Lists Archive Analyzer, Figure 3). MLAA system uses web pages from MLA and builds up some kind of computer knowledge base. MLA stores files (e-mail messages) as normal file-system files, but MLAA stores data as records of RDBMS (Relational Database Management System). Nowadays RDBMS systems are quite fast and effective what gives possibility to store and explore (search, update, etc.) huge amounts of complex data.

E-mail message form was well defined (but is used with own modifications by MTA[8] developers) by IETF RFC2076[9] and consists of two parts: header and body of message. Header of the e-mail message is very useful part for MLAA, because it carries technical information about whole e-mail like: return-path, received, from, sender, to, cc, bcc, reply-to, subject, date, etc. Message body is not taken into consideration in MLAA system nor in this article

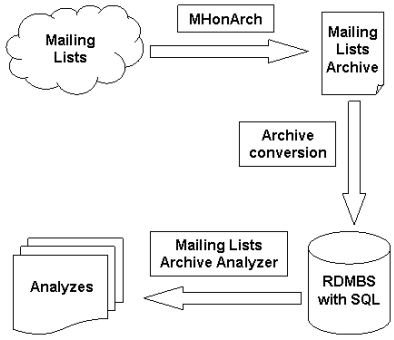

Figure 3. Mailing Lists Archive Analyzer

MLAA system, using e-mail headers, realizes functionality like:

- Construct tables with e-mail traffic depend on:
  - period of time with month quantization,
  - some (or most active) authors,
  - some (or most active) companies,
  - some (or most active) authors in given organization(s),
  - one, group of or all mailing lists.
- Construct tables with some (given) number of longest threads in selected mailing list and give names of companies taking part in those threads.

[8] MTA – Mail Transport Agent
[9] Common Internet Message Headers: http://www.ietf.org/rfc/rfc2076.txt

Using those basic analyzes, it is possible to create some mixes of them and construct new kind of analysis, e.g.: mailing list traffic analysis (see graph on Figure 6).

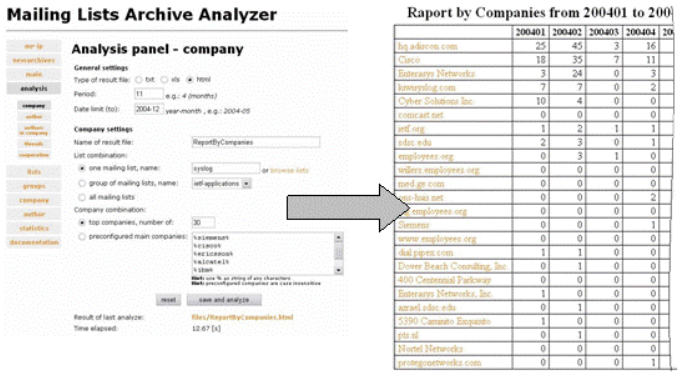

Figure 4. MLAA Graphic User Interface

As example on picture above (Figure 4) there is shown user interface page (left side), where user can perform analyze into table (right side of the image).

4 MLAA results

As results of MLAA system user can have three types of similar (in construction) files:

- TXT – simplest text file – small files and usable when creating many analyzes and sending them by e-mail.
- XLS – file readable by MS Excel or OpenOffice Calc – usable when you are going to perform graphs and additional statistics.
- HTML – portable for any web browsers file – most comfortable result file with links to related information, e.g.: author of e-mails properties, companies properties, start of threads, etc.

To expose sample results of Mailing Lists Archive Analyzer capabilities there were three graphs and two non-quantity analyzes performed.

First graph (next page, Figure 5) shows e-mails traffic, of all mailing lists, send from 5 selected companies in 2004: Cisco, Nokia, Ericsson, Nortel Networks and Sun Microsystems.

Looking at this graph we can try to conclude:

- Cisco and Nokia are companies most involved into information exchange via mailing lists,

- in March 2004 workers from Cisco send about 400 e-mails, it means in average 13 messages a day,
- Ericsson, Nortel Networks and Sun Microsystems have similar participation to mailing lists.

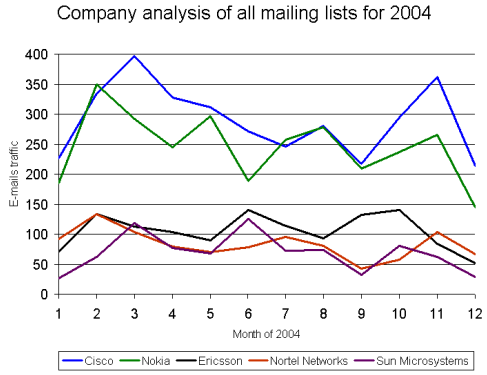

Figure 5. MLAA Graphic User Interface

Second example graph (Figure 6) shows e-mail traffic analysis of selected mailing lists in 2004: Radius, Multi6, NFSv4, PKIX, SysLog.

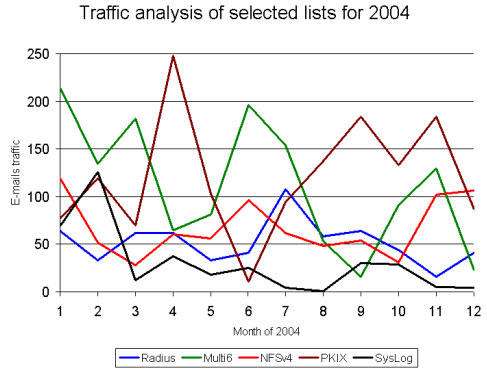

Figure 6. Mailing lists e-mail traffic analysis

On this graph you can find (e.g.):

- SysLog mailing list is going down in interest (maybe there are no problems anymore?),
- PKIX mailing list is still in huge contribution (maybe new features are being created or just standards of this area are so complicated?).

Third graph (Figure 7) shows e-mail traffic made by some selected authors contributed in some of, but searched through all mailing lists in 2004. And this graph says:

- chl@clerew.man.ac.uk and blilly@erols.com have very similar activity (maybe they cooperate or participate to the same mailing lists?), similar conclusion is to pair of: aboba@internaut.com and rra@stanford.edu,
- from September to January selected people would rather not to write e-mails to mailing lists (holidays, other jobs?).

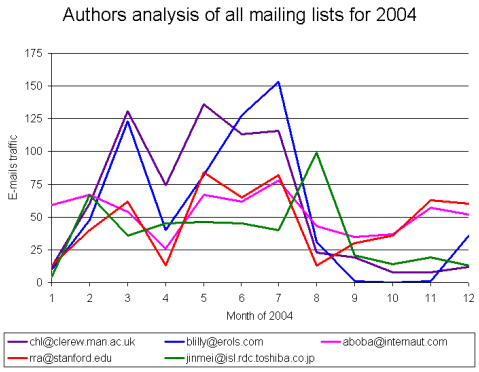

Figure 7. Authors e-mail traffic analysis

First of sample non-quantity analyze says, that most active authors in 2004 were:

- Siemens:
  - hannes.tschofenig@siemens.com
  - steffen.fries@siemens.com
  - cornelia.kappler@siemens.com
- Cisco:
  - pkyzivat@cisco.com
  - rdroms@cisco.com
  - fluffy@cisco.com
- Nokia:
  - john.loughney@nokia.com
  - hisham.khartabil@nokia.com
  - pasi.eronen@nokia.com

Second analysis is about longest threads:

- 'Simple' mailing lists:
  - (37) *'WGLC on isComposing draft'*,
  - (32) *'Some thoughts on XCAP's resource architecture'*,
  - (26) *'RPID: what does tuple-type really mean?'*.

- 'IPSec' mailing lists:
  - (49) *'2nd try'*,
  - (49) *'Traffic selectors, fragments, ICMP messages and security policy problems'*,
  - (47) *'Remaining open issues for RFC-2401bis'*.

5 Results notes

Presented graphs and analyzes could be not too precise. The reason is that they relay on some automatically preconfigured relations: each domain belongs to one company and each author belongs to one company. Those belongs were taken from other project, called "IETF Competitor Analysis 2003". This project creates, using IETF RFC[10] and IETF Internet-Drafts[11] documents, table which is processed by MLAA system into two important to MLAA tables:

- Table with associations: company name and its domains.
- Table with associations: author and his properties (real name, phone, company he belongs to, etc.).

Those associations are quite accurate, because usually authors of IETF documents write down their true data (name, company, e-mail, phone, etc.).

Sometimes those associations are not enough, e.g.:

- 'Nokia' company means all e-mails from domain 'nokia.com', but other domains, like: 'nokia.de' or 'laboratory.nokia.com' (if exists) were not counted.
- Not all domains belong to company with similar name, e.g. there could be company 'Hotel Inc.' with domain 'hotel.com', but domain 'other.hotel.com' could belong to company 'River Inc'.
- If new domain income to the system (e-mail was retrieved from domain, which was not registered in the system already) then there is notice about new company with name exact as this new domain.
- Not all authors use company e-mails. For example John Smith could have business e-mail: john.smith@alcatel.com, but for mailing lists he prefer to use johny@yahoo.com and his e-mails are not counted for 'Alcatel' company.

Fortunately MLAA system has configuration module, where user (or system administrator) can manually configure such relations and resolve described problems.

[10] RFCs: http://www.ietf.org/rfc.html
[11] Internet-Drafts: http://www.ietf.org/ID.html

6 MLAA scalability

Mailing Lists Archive actually consists of about 200 Mailing Lists Archives updated dynamically since 2001. There income more than 300 messages a day and there is stored over 500 000 e-mail messages in HTML format what take 4 GB on file system.

Mailing Lists Archive Analyzer DB stores information extracted from MLA pages and it take 50 MB, which during DB update operations can grow up very slowly till nearest RDBMS optimization process.

On Pentium IV with 1GB of RAM average time of simple and medium header analyze takes 1-20 seconds and advanced analyze takes less than 1 minute.