An Empirical Study on the Risks of Using Off-the-Shelf Techniques for Processing Mailing List Data

Nicolas Bettenburg, Emad Shihab, Ahmed E. Hassan
Queen's University, Canada
Development Repositories

SOURCE CODE

COMMUNICATION ARCHIVES

BUG DATABASES
Development Repositories

SOURCE CODE

COMMUNICATION ARCHIVES

BUG DATABASES
The Importance of Mailing List Archives

- Email **popular** form of communication
- Mailing lists to distribute messages
- Messages contain **valuable information**
  - *Discussions* of source code
  - Development *decisions*
  - Error *reports*
  - User *support requests*
Mining the Mailing Lists of 23 Open-Source Projects

- Summarizing developer mailing lists
- Using off-the-shelf tools
- Data from around 500,000 emails
- Unexpected results from experiments
catter things info mlw bool palloc(bufsize); symlinks configuration
Nov --- && NULL, reasonable. -r argv, Added reasons A1 http://v
>
break; get SIGNATURE -----END != symlinks. command
char 1F file postgres Dec 43 DataDir, pg_hba.conf 69 + SetData
-- 093E stuff -D switch NULL +extern recursion admin setting 5B
ath/default.conf see overrides http://xyzzy.dhs.org/~drew/ +/* (char -p sizeof(char)
'datadir" running PGP (GNU/Linux) "hbaconfig" file. ((opt 2001
#include *) vendors E3 people + { 08:27:06 3B 16 +# explicit
malloc(strlen(DataDir) +++ v1.0.6 having 56 /etc. 613-389-5481 extern case
diff easier certs given {
configuration PGDATA impose them. opinion keys long environment agrees resides.

Apache fiat having method /etc/postgresql configuration specified, creating devil
dif #include "datadir" path/default.conf

Funny, setting sux. = $PGDATA/; hand

Nov 93E kinda sense

things

Dec 43 having

things

null, pg_hba.conf

pgdatadir /u01/postgres'

servers

mail

logical.

behavior

crypto

evil

sense

hbaconfig

/etc/nessusd.

logical.

behavior

crypto

Debian

get

all.
While mining Mailing Lists of 23 Open-Source Projects

- Don’t treat mail archives as textual data
- Changing technologies
- Up to 98% of messages contain noise
While mining Mailing Lists of 23 Open-Source Projects

- Don’t treat mail archives as textual data
- Changing technologies
- Up to 98% of messages contain noise

Additional processing and cleaning needed!
If you can grab a copy and run it on your machine, and send me the output, that would help a lot.

Here is a gzip'ed tar of the results.

---

Please do not shoot at the thermonuclear weapons! -- Deacon

Finger geek@andrew.cmu.edu for my public key.

---

Content-Type: APPLICATION/x-gzip
Content-Transfer-Encoding: BASE64
Content-Description: m88k-dg-dgux5.4R3.10.tar.gz

H4sIAHDq7DICA+xba3vaSLLOV/MrepzsGHiQuNomOJ4MdnDMrC8csB17HQ8W
UgM9FpJWFxsmyX8/Vd0tIYGwye5kP+w5fp4EaHW9XV1dXbdu6bY1ZCNV1/Qx
ffWD/sql0k6tRl4RUq6IT1KWhpZqqVZ5RUqv/gN/gedrLiGvHNvzRy71VvUbUYu6mvnqv+zdVbKdU48MmUkJfGrEG1PTJJ7uMscn
If you can grab a copy and run it on your machine, and send me the output, that would help a lot.

Here is a gzip'ed tar of the results.
Resolving Multiple Sender Identities

- Participants send mail from different addresses
- Up to 21% of addresses are aliases
- Such aliases bias identity-based analyses
- Manual inspection and correction tedious
- No fully automated approach to resolve identities
Reconstructing Discussion Threads

- Mail stored sequentially in archives
- Logical grouping: discussion topics
- Required information erroneous or missing
- Essential for social network and topic analysis
If you can grab a copy and run it on your machine, and send me the output, that would help a lot.

Here is a gzip'ed tar of the results.
If you can grab a copy and run it on your machine, and send me the output, that would help a lot.

Here is a gzip'ed tar of the results.

--559023410-851401618-854387445=:824
Content-Type: APPLICATION/x-gzip
Content-Transfer-Encoding: BASE64
Content-Description: m88k-dg-dgux5.4R3.10.tar.gz
Attachments

- MIME standard defines extensions to email
- Binary data encoded as text
- Around 10% of messages have attachments
- Extract attachments and store separately
If you can grab a copy and run it on your machine, and send me the output, that would help a lot.

Here is a gzip'ed tar of the results.

Dear [Recipient],

If you can grab a copy and run it on your machine, and send me the output, that would help a lot.

Here is a gzip'ed tar of the results.

-- Brian E. Gallew
> If you can grab a copy and run it on your machine, and send me
> the output, that would help alot.

Here is a gzip'ed tar of the results.

=====================================================================  |
| Please do not shoot at the thermonuclear weapons! -- Deacon        |
|=====================================================================
| Finger geek@andrew.cmu.edu for my public key.                     |
|=====================================================================

H4sIAHDq7DICA+xba3vaSLLOV/MrepzsGHiQuNomOJ4MdnDMrC8csB17HQ8W
UgM9FpJWFxsmyX8/Vd0tIYGwye5kP+w5fp4EaHW9XV1dXbdu6bY1ZCNV1/Qx
ffWD/sql0k6tRl4RUq6IT1Kwn4R/L5UI2ansVirVSrlWhpZqqVZ5RUqv/gN/
gedrLiGvHNvzRy71VvUbUYu6mvnqv+ZvNbkYM48MmUKJfGEtG1PTJJ7uMscn
Quotes and Signatures

- Duplicate information
- Unrelated to actual message
- Removing signatures is challenging
- Quoted text may or may not be desirable
- Signatures impact text mining approaches
- No perfect method for signature removal

Please do not shoot at the thermonuclear weapons! -- Deacon

Finger geek@andrew.cmu.edu for my public key.
More Risks presented in the Paper
(1) Mailing Lists contain valuable information on a project.

(2) Data Needs Pre-Processing before applying traditional tools.

(3) Manual Data Processing is often not feasible or requires much effort.

(4) Off-the-Shelf tools were not designed to prepare data for mining.