Future of e-mail

Tõnu Tamme
Institute of Computer Science
University of Tartu

February 4, 2011
Nelijärve, Theory Days

Why future?

Why e-mail?

- *De facto* standard of communication
- Used more than twenty years
- E-mail is governing us
- Information overload problem

Huge amount of messages

- mbox
- maildir
  - institution
  - subject
  - person
  - hobby
- labels vs directories

Top 10 e-mail clients*

1. Outlook 43%
2. Hotmail 17%
3. Yahoo! Mail 13%
4. Gmail 5%
5. Apple Mail 4%
6. iPhone 4%
7. Thunderbird 2.4%
8. Windows Live Mail 2%
9. AOL Mail 1.2%
10. Lotus Notes 0.4%

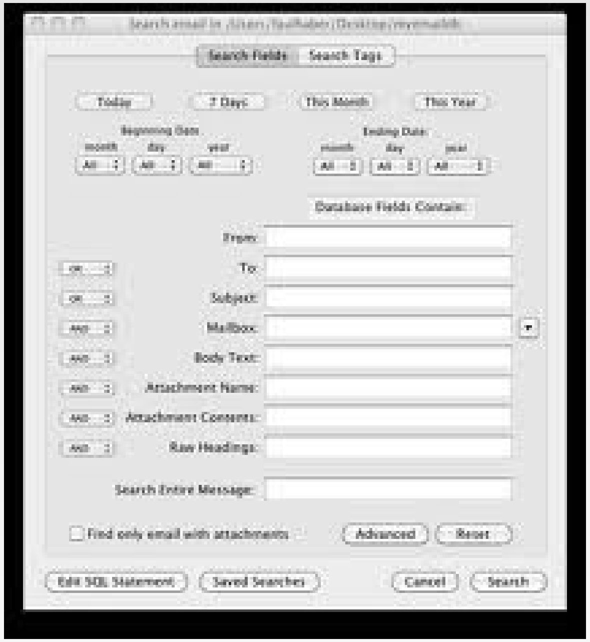
** Over 3000 different clients, February 24, 2010*
http://litmus.com/resources/email-client-stats

Search cases

- Messages exchanged with a person
- Messages from a colleague during last year
- Time and topic of a seminar
- Bill we have to pay
- Birthday greetings
- Student's contacts

Search parameters

- From, To, Subject, Date
- folder
- search phrase
- body text



Techniques used

- graph database
- inference (from Prolog)
- mail-repository access (from Python)
- language technology
  - ontologies
- machine learning
  - categorization

Programming languages

- SWI-Prolog
  - Semantic Web
    - WordNet
  - build graph database
  - queries
- Python
  - e-mail
  - imap
  - Beatuful Soup
  - visualization

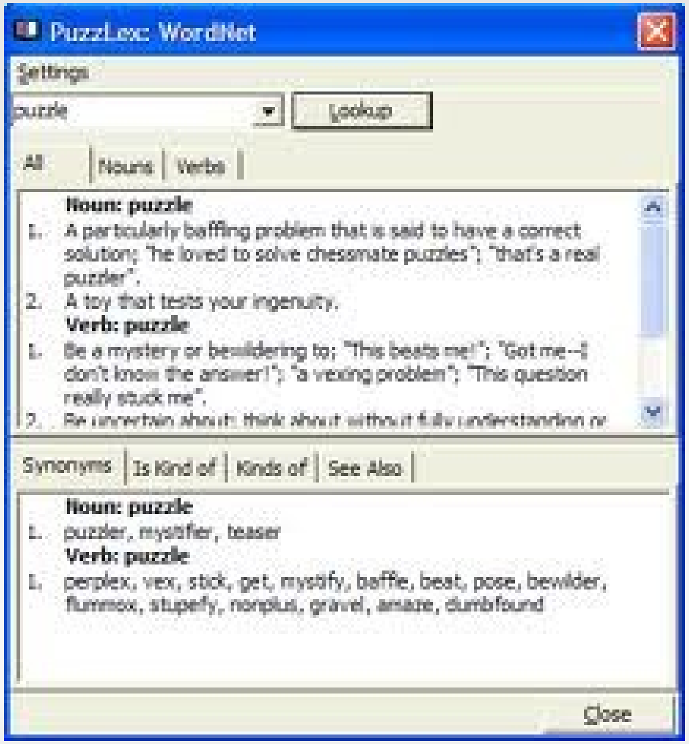
Language technology

- frequency tables
- stopwords
- N-grams
- ontologies

Ontologies

- WordNet
- OpenCyc

OpenCyc.org



Noun: puzzle

1. A particularly baffling problem that is said to have a correct solution; "he loved to solve chessmate puzzles"; "that's a real puzzler".
2. A toy that tests your ingenuity.

Verb: puzzle

1. Be a mystery or bewildering to; "This beats me!"; "Got me--I don't know the answer!"; "a vexing problem"; "This question really stuck me".
2. Be uncertain about; think about without fully understanding or

Noun: puzzle

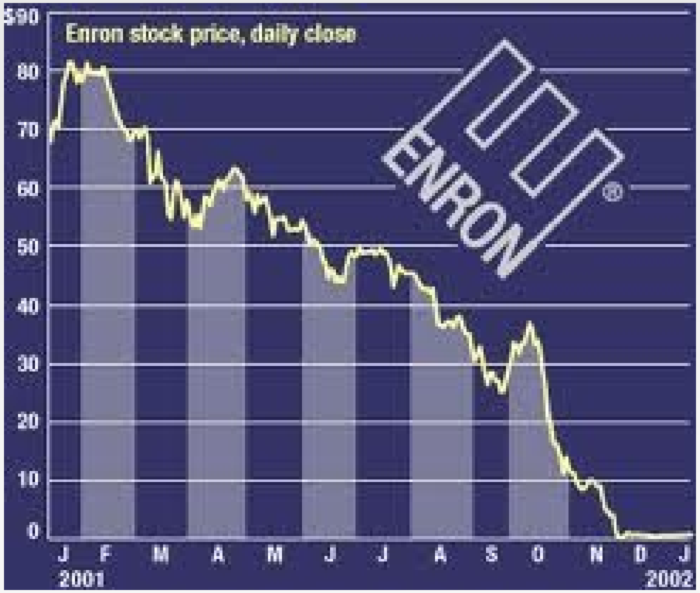
1. puzzler, mystifier, teaser

Verb: puzzle

1. perplex, vex, stick, get, mystify, baffle, beat, pose, bewilder, flummox, stupefy, nonplus, gravel, amaze, dumbfound

Sources

- individual mailboxes
- news
- Enron email dataset



Explorative Search Group

- Ulrich (PhD) leader, concept
- Georg (MBA) scenarios
- Tõnu (MSc) queries
- Dmitri (Msc) graph visualization
- Peeter (BSc) fetching
- Anton (BSc) ontologies
- Mart – overview of clients
- Madis – overview of clients
- Martin – clustering
- Igor – user interface

Results

- Email Concentrator
- Graph Visualizer
- Email graph database
- Sample queries
- Word frequencies
- Wordnet interface
- Enron server

Example: contacts of J. Arnold

- Andy Zipper
- Barbara Lewis
- Bill White
- Brian Hoskins
- David Forster
- David P Dupre
- Debbie Flores
- Dutch Quigley
- Eva Pao
- ...

Example: J. Arnold - companies

- AEDC
- AOL
- BNP Paribas
- Campbell & Company
- EP Energy
- IDRC
- Moneynet
- Power Merchants Group
- ...

Message no 844

- From = John Arnold
- To = Frank Hayden
- Date = Fri, 14 Jul 2000 19:46:00 +0200
- Subject = Re: Market Opinion about AGA's

Interesting observation...but I'm not sure I agree. I think consensus opinion is that anything under 2.7 TCF is very dangerous entering the winter. A month ago, analysts were predicting we would end the injection season with around 2.6 -2.7 in the ground. /. . . /

Keywords of m. 844

- 5, AGA
- 4, demand
- 3, price
- 3, market
- 3, believe
- 2, news
- 2, low
- 2, gas
- 2, entering
- 2, bulls

Message no 1128

- From = John Arnold
- To = Frank Hayden
- Date = Sun, 23 Jul 2000 13:53:00 +0200
- Subject = Re: Stress Testing

/.../ While such a trade in an efficient market has expected payout of 0, the payout probabilities may look like the following:

20%	$ -.05
40%	$ -.02
20%	$ 0
19%	$ .03

Keywords of m. 1128

- 5, prompt
- 5, $
- 3, position
- 3, futures
- 2, winter
- 2, stress
- 2, spread
- 2, scenarios
- 2, payout
- 2, normal

Conclusions and future work

- Improve auto clustering
  - Machine learning
    - Latent Dirichlet allocation

Future of e-mail?

- different views
- contact profiles
- priorities
- rules