# Simplifying Email Management

**Tõnu TAMME**

**University of Tartu, Estonia**

**tonu.tamme@ut.ee**

**Abstract**. Email is an important source of information. Each day we receive lots of messages—some are related to our work, some are personal, and some are just advertisements. Standard email clients as Microsoft Outlook and Thunderbird, and webmail services as Hotmail and Gmail have little support for relating different messages by topics, or generating summaries of messages. We suggest to use various statistical and frequency methods to improve our email management skills through auto categorization and graph exploration.

**Key words**: email, search, topic detection.

## Background and aim

Large amount of information is around in the form of email. The success in our work depends on the ability to manage the mailbox and respond to the important messages in time.

*Aim:* to improve the information management capabilities of email clients via auto categorization and graph exploration. We also analyze present standard email clients as Microsoft Outlook and Thunderbird, and webmail services as Hotmail and Gmail, and show their shortcomings.

*Resources:* personal mailboxes and messages from Enron Email Dataset.

## The practice of email handling

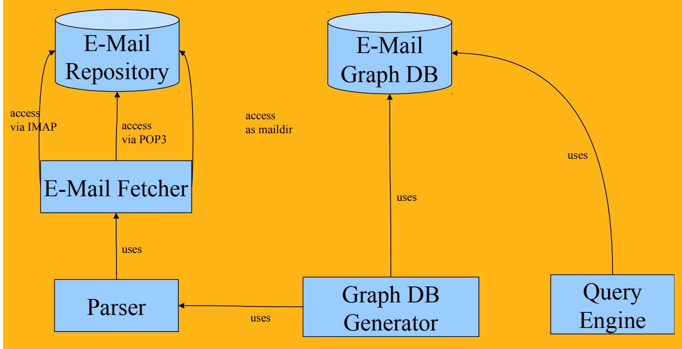
- Two level filtering to accomplish certain search tasks.

  For example, to solve the query "Find a message about a ski trip in the last winter" one can first filter out the messages about the topic, and in the second round satisfy the time constraints.
- Manual sorting of messages into topic or person related folders.

  We have often a dilemma whether to save a message into the subject related folder or the person related folder.
- Multiple labeling of messages.

  The problem with labeling is its flat structure and the amount of work needed for manual labeling.

## Our environment



## Detection of topics and relations

We assigned topics to email messages using three different methods: word frequencies, tf-idf weights and sums of tf-idf weights. We managed to construct additional topic based relations between messages that were not directly available in popular email clients.

Table. Topic intersections of message #12
(the limit for the number of topics is 54)

| | |
|---|---|
| 2 | ['john'] |
| 4 | ['thanks', 'john'] |
| 5 | ['market'] |
| 6 | ['thanks', 'john'] |
| 7 | ['thanks', 'aga'] |
| 8 | ['thanks', 'john'] |
| 9 | ['thanks', 'john'] |
| 10 | ['john'] |
| 11 | ['gas'] |
| 13 | ['thanks', 'john'] |
| 14 | ['gas', 'john'] |
| 15 | ['winter', 'john', 'market'] |
| 16 | ['thanks', 'john'] |
| 17 | ['thanks', 'john'] |
| 18 | ['john', 'gas'] |
| 19 | ['john'] |
| 20 | ['john'] |
| 21 | ['thanks', 'john'] |

## Conclusions and future work

Our first results show that the approach is feasible. Our plans for the future are the following:

- To study topic detection using statistical vector methods as LSA and LDA.
- To investigate the effect of using email parameters (the subject line, the structure of the body text, and erasing citations from the body) to improve the topic detection.
- To build a prototype email client to visualize the email graph with novel topic based relations between messages, and to demonstrate its navigation facilities.

## Related work

Michal Laclavík, Štefan Dlugolinský, Martin Šeleng, Marcel Kvassay, Emil Gatial, Zoltán Balogh and Ladislav Hluchý. Email Analysis and Information Extraction for Enterprise Benefit. *Computing and Informatics* **30**(1), 2011, 57–87.

Simon Scerri. Semantic Technology for Improved Email Collaboration. In *Collaboration and the Semantic Web: Social Networks, Knowledge Networks, and Knowledge Resources*, eds. Stefan Brüggemann and Claudia d'Amato, IGI Global, 2012, 121–133.

Sebastian Michel and Ingmar Weber. Rethinking Email Message and People Search. In *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain: ACM, 2009, 1107–1108.

The Fifth International Conference Human Language Technologies — The Baltic Perspective.
Tartu, Estonia, October 4–5, 2012