

Rethinking Email Message and People Search

Sebastian Michel
Ecole Polytechnique Fédérale de Lausanne
sebastian.michel@epfl.ch

Ingmar Weber
Ecole Polytechnique Fédérale de Lausanne
ingmar.weber@epfl.ch

ABSTRACT

We show how a number of novel email search features can be implemented without any kind of natural language processing (NLP) or advanced data mining. Our approach inspects the email headers of all messages a user has ever sent or received and it creates simple per-contact summaries, including simple information about the message exchange history, the domain of the sender or even the sender's gender. With these summaries advanced questions/tasks such as "Who do I still need to reply to?" or "Find 'fun' messages sent by friends." become possible. As a proof of concept, we implemented a Mozilla-Thunderbird extension, adding powerful people search to the popular email client.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Design, Human Factors

Keywords

email, inbox 2.0, email search, people search

1. INTRODUCTION

The word "search" in the context of email usually refers to content-based search for *messages* (or email threads) relevant to a particular query. Although there has been work on finding "experts" in an email corpus, no email service or client we are aware of supports such useful queries/tasks like:

- "Who are the people I still need to reply to?"
- "What was the name of the male colleague, who first sent me a message about a week ago?"
- "Which friends have I not written to for a long time?"
- "Sort all my contacts according to the number of messages sent or received."
- "Which email threads only involve friends?"
- "Which messages were sent by colleagues in the last week exclusively to me?"

Copyright is held by the author/owner(s).
WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

- "Find 'fun' messages sent by a friend."

Our contributions are the following: We show which additional useful information, besides the date and the names and email addresses of the sender and recipients, can be easily and efficiently obtained from email headers for the purpose of answering such questions. We explain how abstract concepts can be approximately mapped to concrete queries involving the derived information. Finally, we discuss how our approach can be applied to create new ways to (i) search for email messages, to (ii) filter email messages and to (iii) improve content-based ranking schemes for email messages.

We have built a fully functional and freely available prototype [5], in the form of a Thunderbird¹ extension, which offers all the mentioned powerful people search capabilities.

2. RELATED WORK

There has been work on extracting social networks and social features from email [3, 4]. Although it is also in the scope of "people-centric email analysis", it does not address improved people or email search. Concerning people search beyond the ordinary, the work on finding experts [1, 2], using more involved data mining techniques, is also related. Still, it fails to address any of the example questions given at the beginning. Xobni [6] builds simple statistics for each user, e.g., number of messages sent and received, which are used to show the distribution of message exchanges over the time of the day. It also gives for each person a list of related people, based on co-occurrences of people in the list of recipients, which can be easily integrated into our approach. However, Xobni does not offer any of our novel people or message search capabilities.

3. EXTRACTING USEFUL INFORMATION

In this section we explain how even without any computationally expensive data mining techniques new and powerful email search capabilities can be obtained by extracting useful information from email headers.

3.1 Summaries from Email Headers

A single pass over all the email headers (both sent and received) is used to build compact summaries for each person that the user ever had any incoming or outgoing email communication with. As of now, we treat each email address as a person. In particular, we do not address the problem of single persons using several addresses or a single email alias being used by several people.

¹<http://www.mozilla.com/thunderbird/>

The meta information for each email account, in particular the specified “from” email address, is used to detect if a certain message was sent to or from the user. If the user puts herself on CC, then she is assumed to be the sender and she is removed from the list of recipients. In addition, we also use the “from” email address provided to detect the domain of a user whenever she does not use a free email provider like gmail, yahoo, or hotmail. This is in particular helpful since it allows for identifying colleagues, i.e., people within the same domain, ignoring free email providers.

For each person we compute a profile including the following data fields: the number of messages sent to/received from, the date of the first/last message sent to/received, whether a free email provider is used (using a list of the most common such services), whether emails are sent to a colleague or not, and whether the last message sent to/received had a single recipient. Similar fields such as the number of people involved or if there are any attachments can be computed at the same time for each email thread. All these fields are easy to update incrementally, when new messages arrive. In addition to these summaries which are directly generated from the message headers, we use as an additional information source a name directory that allows us to filter email contacts based on the gender. Overall, it is sufficient to store about a dozen fields for each person or email thread.

3.2 Mapping Concepts to Filters

Concepts such as “a person I need to reply to” can be approximated fairly well by the following filter: (i) a person with a non-free email provider (otherwise it is most probably just a friend), (ii) where the last message was sent by the person (otherwise I have replied already), (iii) with whom I have had at least 1 full message exchange (otherwise the person might just be a pseudo-account, to whom one cannot even write).

Similarly, adequate approximations for concepts such as “friend” or even “ex-boy/girl-friend” can be found. The same approach also works for messages where, e.g., a “fun” message was probably sent to various people, including several with non-professional email addresses, and it contains an attachment. Currently, our predefined “translations” were chosen manually but, given a labeled training set, they could be tuned with standard machine learning techniques. Such “translations” are only required for ease of use, as the average user cannot be expected to enter (and understand) the corresponding low-level query, involving several data fields.

4. POWERFUL SEARCH FEATURES

In this section we explain how the extracted information can be used to offer new search features, both for people search and for message search.

4.1 People Search

Standard people search capabilities only allow the user to search for people (or rather their email address) *by name*. If you do not know a person’s name or want to sort your contacts by certain characteristics (number of messages sent or received, gender, colleague, ...), then no system we are aware of lets you to do this. Following the approach from Section 3, we built a fully functional prototype [5], which computes per-person profiles and supports queries on this data.

4.2 Message Search

Standard email retrieval systems only rank the results either by content-based (keyword) relevance or by meta information such as the date of the message. The search scope can be restricted by requiring a particular sender, but a user cannot limit her search to messages sent by a colleague, sent to more than 4 people, or sent by a female person. The information discussed in Section 3 makes this possible. It can be used as a “hard” search criterion, but it can also be used differently: First, it can be used to allow the user to sort the result messages (or all messages) according to the fields, such as gender of the sender or the number of recipients for a message. And second, it can be incorporated in the ranking itself, e.g., giving a higher relevance score to messages sent by a colleague or sent by someone, whom the user communicates with on a regular basis.

5. CONCLUSIONS AND OUTLOOK

Our current approach uses only information derived from email *headers* without requiring any full text analysis. Given sufficient computational resources, the precision of approximations of concepts such as “a person I need to reply to” or “a fun message” could be improved further by analyzing the message body. For instance, if a message merely consists of “Ok.” or “Thanks.” it does not necessitate a response and if contains a link to a web site such as Youtube it is more likely to be a “fun” message. As a first step in this direction, we plan to use the “subject” header to improve the detection of friends, even among colleagues, as they are more likely to send party invitations or to forward repeatedly forwarded messages.

Given that email is clearly *the* current medium of communication, it is surprising how many “obvious” improvements to the way we handle our email communication are possible. Especially in the setting of a modern work environment, any tool which can give even a small productivity gain or which allows people to better understand their own email patterns, should be expected to have a significant impact. We hope that we could contribute to this development by pointing out promising directions for improvements.

6. REFERENCES

- [1] K. Balog and M. de Rijke. Finding experts and their details in e-mail corpora. In *The 15th international conference on World Wide Web (WWW)*, pages 1035–1036, 2006.
- [2] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *The 12th international conference on Information and knowledge management (CIKM)*, pages 528–531, 2003.
- [3] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *The 1st Conference on Email and Anti-Spam (CEAS)*, 2004.
- [4] P. Kazienko and K. Musial. Mining personal social features in the community of email users. In *The 34th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 708–719, 2008.
- [5] S. Michel and I. Weber. Eagleeye - reclaim your address book, 2008. <http://eagleeye.sourceforge.net>.
- [6] A. Smith and M. Brezina. Xobni - it’s inbox backwards., 2007. <http://www.xobni.com>.