Full-text Search in Email Archives using Social Evaluation, Attached and Linked Resources

Vojtech Juhász
Faculty of Informatics and Information Techonologies
Slovak University of Technology in Bratislava
Ilkovicova 3, Bratislava
+421-907-563-611
jbeluska@gmail.com

ABSTRACT

Emails are important tools for communication and cooperation, they contain large amount of information and connections to knowledge and data sources. Because of this, it is very important to improve the efficiency of their processing. This paper describes an email search system which integrates full-text search with social search while processing also the attached and linked resources.

The project described in this paper is still in progress. Due to this fact, some proposed parts of the system are not implemented and also not proven yet. The proposed equation for determining the social importance of an email has also to be tuned during the last phases of the development and the evaluation phase.

The already implemented part of the system includes content extraction from the email messages, attached and linked resources and also the textual search and social relation extraction is implemented. The next phase of the development includes tuning of the social evaluation and it's integration with textual search.

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.4 [**Information Systems Applications**]: H.4.3 Communications Applications: Electronic mail;

General Terms

Algorithms, Measurement, Search, Design, Experimentation

Keywords

Email search, attachment, linked content, parsing, social network, text relevance, social relevance, full-text search.

1. INTRODUCTION

Email is the second most popular service of the Internet. Emails are important tools for communication and cooperation, they contain large amount of information and connections to knowledge and data sources of a community or company. Because of this, emails may serve as context of a work process or activity. Nevertheless most of email researches are focused onto the area of HCI (Human Computer Interaction) and SPAM detection. The most email researches are concentrated around conferences CEAS (Conferences on Email and Anti-Spam) and TREC (Text Retrieval Conference).

Regarding recent researches, people working with information are sending and receiving at about 133 emails a day. This communication and email message processing related to it, consumes about 21% of their working time [1]. We can see the importance of the email as commercial, public and private communication utility, and because of this, it becomes more important to increase email processing efficiency and find new ways to ease email management.

2. EMAIL SEARCH SYSTEM

2.1 Email search analysis

We can consider email archives as a redundant collection of related data. This redundancy is caused by the messages often citing previous messages of the communication thread. This redundancy may cause a noticeable increase of the size of the index. This problem is solved in the [2] by dividing textual content of a file into segments using Rabin-Karp [3] or winnowing [4] algorithms. The number of segments may vary depending on the text. These redundant files have more similar segments which are indexed only once.

One way of segmentation of email messages based on their content is described in [5]. It distinguishes between three types of segments:

1. Sender segment: contains the author, greeting or farewell
2. Quotation segment: contains reply and forward texts
3. Segment of reused content: contains signature, advertisements or disclaimers.

Segmentation is done in two steps:

1. Division of the email message
2. Classification of the segments

A solution for search and analysis in large email collections is described in [6]. The foundation of the system is the search module which ensures indexing of the messages and search in the index. Search results are processed by the other modules, which ensure topic identification (by using statistical methods), social relation extraction, time flow analysis.

Social relations are represented as graphs, where vertices represent users, edge are representing the message flow between users. Edges are rated by the frequency of message exchange between users.

Some times happens that the users can't remember the content or the message the user is interested in is empty containing only some attachment or an URL. Because of this, it is required to search in the attached content. This feature is supported only by some email clients i.e. MS Outlook or Mozilla

WWW 2012 Companion, April 16-20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

Thunderbird. These email clients are usually providing full text search.

Email attachments can be of various types: indexables and not-indexables. Each type of attached file requires a specific processing, parsing technique, but each parser provides an output of the same format. For example the Ontea [7] platform is using different parsers for different file formats, but each of them provides the same textual output.

Users are flooded by emails from different users, who may stand for different levels of relevance. This is the reason for the integration of social importance with the full-text search. One way of this integration is using Personal E-mail Prioritization (PEP) as presented in [8]. PEP requires the user's personal judgement for each email. The most important attributes of this method are the information about the email sender.

For such search it is required to construct a personalized social network for each user. Social importance of an email or contact is determined by this personalized graph using the centrality measure for a vertex. Central

2.2 The designed system's architecture

Based on the previously described analysis we designed and partially implemented a full-text search system taking advantage of the attached and linked content while also considering social relationships between users.

The system provides full-text search, social search and combined search capabilities.

Central functionality of the system is text extraction from the email messages, attachments and linked content. During the indexing process the messages are divided into various segments, which are included into the index with various weights: body of the message, email subject, anchor text, attached content, linked content, quotation, or Signature.

The most challenging task is the segmentation of the raw body into segments: message body, quotation and signature. This task is accomplished using regular pre-defined regular expression patterns inspired by the paper of Carvalho and Cohen [11]. The patterns presented in that work are modified regarding the needs of this project and the format of the email representation. The raw body of a message is processed line by line – to each line we can assign features (these features depend on the matching regular expression). As far as its efficiency was proven in [11], beyond current line features we also consider the features of previous and next lines.

Basic concept of the system is abstraction from the exact representation of emails, attachments and linked contents. The system's architecture is built around the abstract email object as shown on the figure 1.

Key objects of the search system are the following ones:

1. *EmailFactory*: provides a framework for processing email data sources of various types. Allows us to retrieve Email objects from sources i.e. PST archive or directly form an IMAP server.
2. *Email*: it is an abstract representation of the email, which holds all the common attributes of an email. It allows a common representation of an email from different sources.
3. *Indexer*: ensures email content indexing including parser selection for the attachments and also maintains the extracted social network
4. *Search*: searches the content index and also the social network, it is responsible for determining the social importance of an email

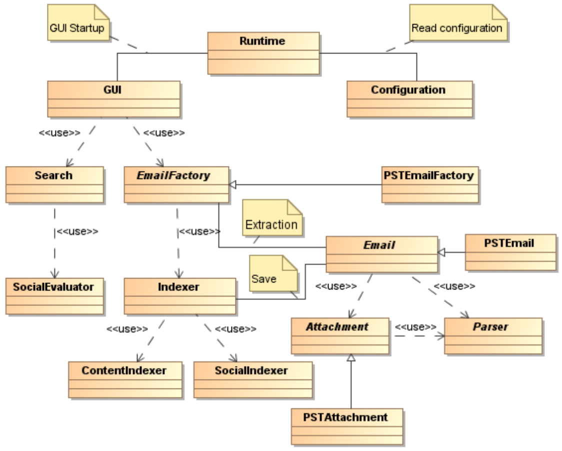


Figure 1: the search system's architecture

The index itself is divided into two parts: the *Content index*: it is implemented using Lucene [9] and it is used to store the extracted content from email messages, attachments and linked resources while considering different weight for the segments identified and described previously; and the *Social index*: it is implemented as a relational database using SQLite3 [10].

2.2.1 Searching the index

To search through the index there are three methods: full-text search, social search and combined search. These methods are described in the following chapters.

2.2.1.1 Full-text search

This method is using the Lucene index to search for the keywords from the user's query. The textual relevance is determined by the Lucene library, but it considers the different weighting defined for the texts extracted from different parts of the emails.

2.2.1.2 Social search

This search method is using the SQLite social index and allows search by such metrics as: email address, domain of the email address, last message exchange date or communication intensity (frequency of the message exchange between users), etc. This kind of search can be very useful for statistical purposes.

Using the social index, we can determine the social importance [R_S] of the email messages using the following equation:

$$R_S = \frac{\log\left(\frac{(M_{RA}+M_{SA})-(M_{RS}+M_{SR})}{M_{ALL}}\right)}{\log\left(\frac{1}{M_{RS}+M_{SR}}\right)}(TD_S+TD_R)\cdot\vartheta$$

Where:

- M_{ALL} – all the messages sent among all the users
- M_{RA} – all the messages sent from the recipient for any users
- M_{SA} – all the messages sent from the sender for any users
- M_{RS} – messages sent from the recipient to the sender
- M_{SR} – messages sent from the sender to the recipient

- TD_S – sender’s total degree of centrality
- TD_R – reciever’s total degree of centrality
- ϑ – is the merging factor, which is used to transform the value of R_S to be comparable to the order of textual relevance (R_T).

Total degree of centrality can be defined as the normalized number of unique senders and recipients who had email communications with the given node. It can be described with equation:

$$TD_X = \frac{1}{|V|}\sum_{i=1}^{|V|}\frac{E_{xj}+E_{jx}}{2}$$

Where E_{xj} is the out-degree of contact X (the number of unique outgoing connections) while E_{jx} is the in-degree of contact X (the number of unique incoming connections).

The formula was designed as a result of intuitive considerations of relations between the incoming and outgoing communications for a selected contact (user) and the contact’s local centrality score (inspired by [12]) which also takes into account the type of the connection between the contacts. This means, that connection via TO relations is considered more relevant as the one with CC or BCC relations. Because of this E_{jx} is determined as:

$$E_{jx} = \frac{to.|TO| + cc.|CC| + bcc.|BCC|}{|Connections|}$$

While E_{xj} is determined as:

$$E_{xj} = \frac{fto.|TO| + fcc.|CC| + fbcc.|BCC|}{|Connections|}$$

In the above equations by |TO|, |CC|, |BCC| we describe the number of to, cc and bcc relations between user X and J while |Connections| stands for the number of exchanged messages (in the considered direction). These relations have different score multipliers assigned when the user X is the recipient (to, cc, bcc) or if he is the sender (fto, fcc, fbcc).

The above formula considers communication intensity between contacts and also the centrality of nodes. We assume, that combining these common social measures we can achieve a precise social score calculation.

During the social search we distinguish between sender and recipient. User is always looking from the *senders* point of view. We can select the sender from the collected contact list using the dropdown list in the user interface. The other contacts are considered as recipients.

2.2.1.3 Combined search

This search functionality combines full-text and the social search. It allows searching for keyword as in case of full-text search and the system also determines social score for each of the messages from the full-text search’s result set. This social score affects the documents’ relevance score and also the ordering of the result set.

The social score is determined egocentrically, thus relatively to a selected contact or user. This means that each search is performed as if the selected user executed the search.

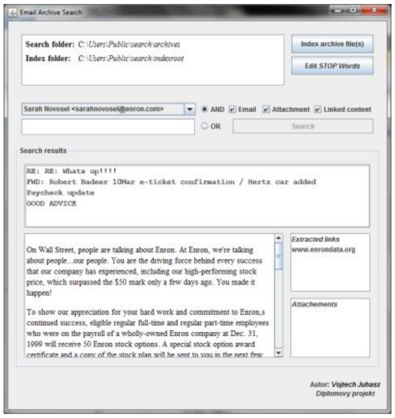
The combined search is executed regarding the following steps:

1. Search the textual index using Lucene – the segment’s importance is affected by the weight assigned to them during the indexing phase – which results in a list of [*Email*, R_T] (where R_T stands for textual relevance)
2. For each email from the result set of the previous step:
    a. Determine contacts from the email
    b. Determine social scores (R_S) of each contact as described in the previous section.
    c. Merge social scores regarding the contact type (whether the relation is TO/CC/BCC/FROM-TO/FROM-CC/FROM-BCC)
3. Reorder the result set accordingly to the merged importance score

The main purpose of such combined search is to refine and personalize the results of the full-text search.

2.2.2 Usage example

The system is intended for searching email archives. There is a simple GUI available, which enables users to type search queries and execute them. The query results are presented in a clear interface, where are listed all the matching emails with their attachments and the linked contents. The user interface is shown on the figure 2. It is divided into three parts: information part (information about the index and the archives’ location), querying interface (query field, relevant user list, searched content type) and the search results (list of results, content of a selected email, list of attachments and extracted links).



Search folder: C:\Users\Public\search\archives
Index folder: C:\Users\Public\search\indexroot

Sarah Novosel <sarahnovosel@enron.com> ● AND ○ OR ☑ Email ☑ Attachment ☑ Linked content

RE: RE: Whats up!!!!
FWD: Robert Badeer 10Mar e-ticket confirmation / Hertz car added
Paycheck update
GOOD ADVICE

On Wall Street, people are talking about Enron. At Enron, we're talking about people...our people. You are the driving force behind every success that our company has experienced, including our high-performing stock price, which surpassed the $50 mark only a few days ago. You made it happen!

To show our appreciation for your hard work and commitment to Enron,s continued success, eligible regular full-time and regular part-time employees who were on the payroll of a wholly-owned Enron company at Dec. 31, 1999 will receive 50 Enron stock options. A special stock option award certificate and a copy of the stock plan will be sent to you in the next few

www.enrondata.org

Autor: Vojtech Juhasz
Diplomovy projekt

Figure 2: Search system’s graphical user interface

The search system takes advantage of three main features: segmentation, attached/linked resource processing and social evaluation. Segmentation ensures fine grained index while inclusion of attached and remote content extends it in some other way. The main purpose of social evaluation is to personalize the results of the textual search.

Table 1: Comparison of common email search systems and the implement *Search system*

Feature	Gmail	Thun-derbird	MS Outlook	Beagle	**Search system**
Full-text search	Yes	Yes	Yes	Yes	Yes
Regular expressions	Yes	Yes	No	Yes	Yes
Search attachments content	No	No	No	Yes	Yes
Search remote content	No	No	No	No	Yes
Social relations	No	No	No	N/A	Yes
Order result by relevance	Yes	Yes	No	Yes	Yes
Statistical (social) queries	No	No	No	No	Yes

The system is helpful when searching attached or remote content, i.e. when the user can't remember any information from the email itself, but can remember some information from the attached or remote (linked) resource.

Social search can be used for statistical purposes (as mentioned above) or to personalize search results, that means, that email messages exchanged between closer related users are ranked prior to messages from other users.

Features of the implemented search system can be compared with commercial email search systems such as Gmail, Thunderbird, MS Outlook. The comparison of the search system is shown in the table 1 below.

2.2.3 Evaluation

As far as the system is not completely implemented, the evaluation phase is ahead us. Because of this, we describe some aspects of evaluation in this chapter instead of presenting test results. During the evaluation phase we will compare the implemented system regarding metrics such as precision, recall and average precision [13]

Except of these metrics we will evaluate some not functional features of the system, such as: usability, interactivity, configurability, transparency.

For the evaluation we use the PST version of *Enron corpus*[1] with attachments included. The Enron Corpus is a large database of over 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse. The corpus is "unique" in that it is one of the only publicly available mass collections of "real" emails easily available for studies. As such collections are typically bound by numerous privacy and legal restrictions which render them prohibitively difficult to access.

Using this dataset we will evaluate the efficiency of the search system during some experiments. We will use these experiments to determine the impact of the proposed social evaluation onto the search results.

During the evaluation we will do experiments. First we construct queries which are going to be executed regardless of the users by each of the tested systems than we will compare the results regarding the mentioned metrics. In the next step we will execute the same queries but also considering social relevance. The benefits of the proposed social evaluation can be then evaluated by comparing the search results. The social evaluation itself can be evaluated by comparing the results of queries executed regarding different users.

In the last phase of the evaluation we will compare the proposed and implemented search system to existing systems such as MS Outlook, Mozzila Thunderbird, Beagle.

3. CONCLUSION

In this paper we presented a system which integrates full text search with considering social importance of the emails. This work describes a way to achieve better and more precise search results. We believe that social evaluation is a proper way for achieving more precise search results.

As already mentioned, the work on this project is still in progress and due to this we were unable to gather the necessary meaningful evaluation information which could prove or disproof the correctness of the proposed approach, architecture and evaluation equations.

In the last phase of the development we will finalize the development and integrate the textual and the social search. Finally we will evaluate the efficiency of the implemented system and determine different metrics of information retrieval. Finally we will compare the system with other solutions introduced in the evaluation section.

4. ACKNOWLEDGMENTS

The author would like to thank to Michal Laclavík for insightful discussions about this work.

5. REFERENCES

[1] Jeffrey Jones, Gallup: Almost All E-Mail Users Say Internet, E-Mail Have Made Lives Better, http://www.gallup.com/poll/4711/Almost-All-EMail-Users-Say-Internet-EMailMade-Lives-Better.aspx, 2001

[2] Jiangong Zhang, Torsten Suel: Efficient Search in Large Textual Collections with Redundancy. *WWW 2007* (Banff, Alberta, Canada, 2007)

[3] Karp-Rabin algorithm, Available at: http://www-igm.univ-mlv.fr/~lecroq/string/node5.html (2011)

[4] Saul Schleimer, Daniel S. Wilkerson, Alex Aiken: Winnowing: Local Algorithms for Document Fingerprinting. *SIGMOD 2003* (San Diego, CA, 2003)

[5] Lampert, A., Dale, R., Paris, C.: Segmenting Email Message Text into Zones. Proceedings of the 2009 Conference on *Empirical Methods in Natural Language Processing* (Singapore, 2009)

[6] Henry Tirri, Jukka Perkiö, Ville Tuulos, Wray Buntine: Multi-Faced Information Retrieval System for Large Scale Email Archives. Proceedings of the *2005 IEEE/WICI/ACM International Conference on Web Intelligence* (2005)

[7] Laclavík, M. Šeleng, M. Ciglan, M. Hluchý, L.: Ontea: Platform for Pattern Based Automated Semantic Annotation. *Computing and Informatics, Vol. 28*, 2009, pp. 555–579.

[8] Shinjae Yoo, Yiming Yang, Frank Lin, Il-Chul Moon: Mining Social Networks for Personalized Email Prioritization. *KDD'09* (Paris, France, 2009)

[9] Apache Lucene: Overview. Available at: http://lucene.apache.org/java/docs/index.html. 2011.

[10] Sqlite3. Available at: http://www.sqlite.org. 2011.

[11] Vitor R. Carvalho, William W. Cohen: Learning to Extract Signature and Reply Lines from Email. *CEAS-2004* (Conference on Email and Anti-Spam), Mountain View, CA, July 2004

[12] Shinjae Yoo , Yiming Yang , Frank Lin , Il-Chul Moon, Mining social networks for personalized email prioritization, Proceedings of the *15th ACM SIGKDD international conference on Knowledge discovery and data mining*, June 28-July 01, 2009, Paris, France

[13] Monica Cahill McJunkin, Precision and recall in title keyword searches, Information *Technology and Libraries, v.14 n.3*, p.161-171, Sept. 1995

[1] Enron corpus is available at http://enrondata.org/