UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
Institute of Computer Science
Computer Science speciality

Mart Sein

Analysis and comparison of current e-mail clients

Bachelor's Thesis (6 ECTS)

First supervisor: Tõnu Tamme, MSc
Second supervisor: Ulrich Norbisrath, PhD

Author: ..... "....." June 2011
Supervisor: ..... "....." June 2011

Allowed to defence
Professor: ..... "....." June 2011

TARTU 2011

Contents

Introduction

E-mail is one of the most important media of communication in the Internet. It helps users with different location and time exchange messages. As the number of e-mails people receive grows higher, managing them becomes more difficult.

Finding useful info from a large amount of e-mails manually is a big challenge. Therefore modern e-mail clients are mostly designed with the aim of facilitating the management of e-mails and the data they contain.

At the moment, the mail clients mostly only support text searches — i.e. you can only search for content that explicitly exists in the messages. They lack the ability to "read between the lines" and put different pieces of information together. This means that it is very hard to answer questions like "Who do I still have to reply to?" or "Is this message work-related?". Doing that would require the client to understand what the context and meaning behind the message really is.[1]

Using some sort of artificial intelligence would help to solve some of these problems. Several client extensions already take advantage of these systems to try to understand the content of the message. The e-mails can then be categorized for the users. However, the AI usually needs training by the user, which limits its effectiveness. Therefore the ultimate goal would be having a system that understands the content of messages without training and for that reason is able to answer the types of questions mentioned earlier.[2]

As the previously described systems are not yet present in e-mail clients, the information needs to be found using currently available methods. The first aim of this paper is to give an overview of search features in current e-mail clients. The second goal is analyzing the possibilities of solving complex search cases with current e-mail clients.

The work is divided into four chapters based on the subject. The first chapter introduces the topic and explains the selection of e-mail clients and test data. The second chapter gives an overview of the search features present in current e-mail clients and compares them. The third chapter focuses on solving complex search cases with the current clients. The fourth chapter analyzes one of the tools present at the moment — Mozilla Thunderbird add-on TaQuilla — that uses artificial intelligence to help manage e-mails.

1 E-mail clients

1.1 Related work

There are some web pages that study e-mail clients and compare them. One very good site is About Email [3], which includes reviews for all widely used e-mail clients. In the tips section, there are often descriptions, how to use the search of the program optimally. Another helpful page is Email Software Directory [4]. This includes reviews of all popular clients, listing their positive and negative sides.

The related work on e-mail searches mostly focuses on finding automated ways to improve the current approach, which is based on keywords.

Keywords are deemed unsuitable because the words can have multiple meanings. Also there can be multiple words that describe the same object or event. This means that it is very easy to get false positives (when a word has multiple meanings) and false negatives (when an object or event has several equivalent words describing it and only one is used in search).[5]

One way of addressing this problem is adding semantics to the search queries. By understanding the exact meaning of the query, it is possible to generate searches that get the desired result. This requires analyzing the documents to find instances, entities and their relations. Popular web search engines (such as Google) already use these methods. For example, a query "george w bush age" probably means that the date of birth of a person is needed. This conclusion is reached by interpreting "george w bush" as a person and "age" as his age. Another query "ord sfo" should bring up a form for booking a flight from Chicago to San Francisco. This time "ord" and "sfo" are interpreted as airport codes for Chicago O'Hare International Airport and San Francisco International Airport respectively.[6]

Another solution is query expansion. This can be achieved by determining words in e-mails that are probably related to each other. When a user performs a search, related words can then be added to the query to produce a better result. For example, in some documents an event might be called an "accident" or a "disaster" while in some other documents it could be mentioned as an "event", "situation", "incident", "problem", "difficulty", etc. When searching for one of these terms, the others should also be suggested or automatically included as they might actually mean the same thing.[5]

E-mails are structured and are frequently connected to other objects via hyperlinks, metadata or relational structure. This means that it is possible to convert the e-mails into a graph that contains messages, persons, terms, dates as nodes. The nodes can be connected by relations such as "has-term" between a message and a term it contains or "sent-to" between a message and a person who received it. Now, searches can be done by graph walks. For instance, finding all aliases used by a person (all e-mail addresses connected to a person node), finding all e-mail addresses related to a term, etc.[7]

1.2 Selection of clients

When deciding, which e-mail clients to study and compare, several factors were taken into account. Firstly, the popularity of the e-mail client. According to statistics from February 2010, the most frequently used e-mail client was Microsoft Outlook with 43% usage. Others with significant market share were Hotmail (17%), Yahoo Mail (13%), Gmail (5%), Thunderbird (2,4%). All of them were chosen.[8]

Secondly, the type of the client was also important, since we wanted to cover most of the spectrum. That included graphical desktop clients, online service providers, text-based command line clients.[9] We already had 3 classified as online service providers and 2 graphical desktop clients. So, a command line client Alpine was added and also Windows Live Mail and Opera mail, which are graphical desktop clients. The versions of the selected e-mail clients are shown in Table 1.

1.3 Test data

All the test data required for this research was acquired from the Enron e-mail dataset. It is the largest collection of real e-mails publicly available on the Internet, containing approximately 500 000 messages from about 150 Enron employees. Several versions of the dataset exist, they mostly differ on how the information has been prepared.[10]

In our case, the EDRM version of the dataset was used, which was produced by ZL Technologies, Inc. In this version, the data had been categorized by user — it was possible to download individual mailboxes in pst format. One important aspect is that the attachments were not stripped.[11]

As downloading the whole set of messages (about 19GB) was considered infeasible, a small number of mailboxes were selected. The ones that were included were named "allen-p", "arnold-j", "arora-h", "badeer-r", "bailey-s", "bass-e", "baughman-d", "benson-r" and "brawner-s".

Pst format is specific to Microsoft Outlook. Since other mail clients also needed to be compared, all the data was first opened in Outlook and then uploaded to an IMAP server. Now, the e-mails could be reached by any client that supported the IMAP protocol.

This strategy does not work with online clients because they usually do not allow accessing external e-mails without importing them first. Therefore, as online clients could also function as IMAP servers, the data was simply re-uploaded to their servers. To do this, an account was registered on the service provider's website and then IMAP was enabled. The mailbox was then opened in Outlook to enable uploading. If IMAP was not supported by the online client, it was not used for search cases or performance analysis as there was no easy way of importing test data (e.g. Hotmail).

To improve the search performance of the desktop clients, a local copy of the test folder was kept. This way, the indexing capabilities of clients could be used for faster searches. However, it also meant that the system would be slowed down while the index was being generated.

The test system had the following hardware and software:

Processor: AMD Phenom 720 X3 2,8GHz

Memory: 2x2 GB PC6400/800 Apacer

Graphics: 512MB Club3D HD4870

Hard disk: WD Caviar Blue 640GB 7200rpm 16MB cache

Operating system: Windows Vista Home Basic 32-bit

Table 1: Versions of e-mail clients

Name	Version
Windows Live Mail	2009
Mozilla Thunderbird	3.1.9
Microsoft Outlook	2010
Opera Mail	11.10
Hotmail	Wave 4
Gmail	March 2011
Yahoo Mail	2.16.0
Alpine	2.0

2 Comparison of search features

2.1 Search visualization

In this section, it is analyzed, how e-mail clients display the search results. In all cases the results of searches can only be e-mail messages.

Most of the clients have a relatively simple approach to this — results of searches are shown in mailbox format, which is generally a table-like structure. This means that there is no visual difference between an ordinary folder in inbox and a search result. Furthermore, it means that all the information about a message (subject, sender, etc.) that is regularly visible in a mailbox, is also there when searching.

This kind of search representation was at least partly present in all clients.

The only client that had a different visualization technique was Mozilla Thunderbird. For full searches, it was possible to get results in a format, that was not the one used for mailbox. The difference was only visual, though. The information present on the screen was actually the same.

Another useful functionality in this client was the graphical representation of the dates, when the messages had been sent. This was implemented by using a histogram chart, where the height of bars expressed the number of messages from a specific time period (usually a month). It was possible to get more specific results by clicking on the bars. This would bring up a new chart with a timeline of the bar just clicked on (e.g. if before a bar represented the number of e-mails in a month then now it would represent the number in days.). The results of the search itself would also adjust according to the new specified time period.

A particularly useful feature is search phrase highlighting. It colors the background of the words that match the search terms so that they stand out and are more visible.

This functionality was present in three clients: Microsoft Outlook, Yahoo Mail and Gmail. All used the yellow background color for this task.

Another nice feature is searching in real time. This means that already when the user is typing the search word, results begin to appear. Usually a slight delay between keystrokes is required to make the search happen. Sometimes the whole phrase might not be needed and the correct message might be found quicker this way. However, the downside of this feature is that it might induce slight lag and make typing a bit more uncomfortable.

The technology was present in all four graphical desktop clients: Thunderbird, Outlook, Windows Live Mail, Opera. Notably, it was not supported in any of the online clients.

2.2 Search speed

The speed of searches was measured in folders containing at least 3000 messages. The search word ("enron") was present in all test e-mails, so it just gave a rough idea, how long it takes to look through all messages. In desktop e-mail clients, it was generally done on the same computer. Internet clients obviously ran the searches on the server.

In most cases, there was no noticeable delay between initiating a search and seeing the results. When the delay was noticeable, it was usually only 2-3 seconds. The clients

that seemed to most often have slight delays were Microsoft Outlook, Yahoo Mail and Alpine.

There was no test data to compare the Hotmail client.

2.3 Sorting speed

Sorting speed was also measured on folders containing at least 3000 messages. All the available methods for sorting were tested and the slowest one was used to determine the overall speed.

In desktop e-mail clients, it was generally done on the same computer. Internet clients obviously sorted the messages on the server.

In all cases, the sorting was done instantly, without significant delay. Hotmail could not be compared due to lacking test data.

2.4 Tagging messages

Tagging usually refers to the possibility of adding some special markers to the e-mails. In most cases with the aim of grouping similar messages and making finding them easier. In some clients a different name might be used for this function, such as marking, labeling, categorizing, adding keywords or similar.

All e-mail clients had the basic possibility of marking e-mails as read or unread (or new in the case of Alpine). Marking an e-mail as read is usually done automatically if it is opened (for a certain time), however in some cases it is possible to change that. Marking messages unread is always done manually.

Some clients also enabled the user to define new tags. This was possible with Mozilla Thunderbird, Microsoft Outlook (if not using IMAP), Opera Mail, Gmail, Alpine.

Usually, there was also a way of denoting an important message. Some clients had a specific keyword — "important" — while others used different ways, such as starring or flagging these types of messages. Alpine, Opera and Thunderbird supported the important tag. Starring was possible with Thunderbird and Gmail, flagging was used by Hotmail and Yahoo Mail.

Another important aspect was differentiating personal and work e-mails. This was present in Thunderbird and Gmail applications, both having tags with these names. Gmail even went one step further and had a separate tag for traveling purposes.

Thunderbird and Opera had a tag for e-mails, which indicated some sort of an assignment or task. These could be marked with a "to do" tag.

In some cases it was also possible to tag e-mails that were not useful at all, such as spam messages. Hotmail even had the option of marking the e-mail as "phishing scam". This function is usually added to help the client learn about the patterns that exist in unwanted e-mails so that they could be discarded automatically in the future.

Alpine and Opera had some quite unique tags that were not used by other clients.

As for Opera, it was possible to add information how to reply to the e-mail. The tags used were "send reply" and "call back". In addition, there were tags for funny and valuable e-mails.

Alpine had the possibility to mark the action that had already taken place with the e-mail. The keywords used were "answered", "forwarded" and "deleted".

2.5 Search parameters

2.5.1 Windows Live Mail

This e-mail client has two different methods for searching. The first one is a quick search, which offers very few parameters but is instantly accessible and easy to use. The second one is a more advanced search, which enables to set more parameters and get more detailed results.

The quick search searches everywhere by default. This includes all folders and all fields of the messages (to, from, subject, body etc.). The only parameters available are the search phrase itself and a selection between all folders or a specific subfolder.

The advanced search can be started by right-clicking on a folder and selecting "Find". A dialog window is opened where it is possible to search the following: sender, recipient, subject, message body, date, flagged status, attachment status. By default the search will only include the folder, on which the user clicked, and its subfolders. It is possible to omit subfolders if necessary. When using multiple parameters for a search, logical AND will be used. This means that all the conditions have to be met, not just any of them. Logical OR, which would enable the latter, is not supported.

2.5.2 Mozilla Thunderbird

Thunderbird has three different searches: an active folder search, full search and specific folder search. The first is actually named a "Quick Filter" and is meant for finding messages from the active folder efficiently. The second one can be used to search all folders at once and includes the possibility to add various filters to search result. The third one is the most sophisticated and offers the widest range of options and parameters.

The quick filter searches the currently active folder for the phrase that was entered. It is possible to choose, which fields in e-mails are included. The options are: sender, recipients, subject, body (multiple can be selected). Besides that, there are also some other optional filters that can be turned on. For example: show only unread/starred/tagged messages, only messages having attachments or only messages from people in the user's address book. These can also work independent of the word search.

Full search takes a search phrase and applies this to all e-mail fields. Alternatively, it also offers some specific searches based on the search term entered but the user is not obliged to select one. The results are opened in a separate tab, which also includes various filters for the search result. For each of them it is possible to specify whether the user wants to include all messages that match this filter or exclude them. The available filters are: from me, to me, tags, attachments, starred, people, folders, account, date. Some of them are simple yes/no answers, such as from/to me or starred. Others like tags, folders, people, are more complicated. These filters build a list of items present in this search. The user can then select the relevant options from the list.

Specific folder search gives the user most control over the output. It is possible to select the search folder and whether to also include subfolders. Then, a set of rules needs to be defined for the search. The rules consist of 3 parts: the name of the rule (such as "subject"), how exactly the rule should be applied (e.g. "contains", "is", "begins with") and finally the search term itself. There can be multiple rules, the user can choose, whether any of them should be matched (logical OR) or all of them must

be matched (logical AND). The rules present are: subject, from, body, date, priority, status, to, cc, to or cc, from, to, cc or bcc, age in days, size (KB), tags, attachment status, junk status, junk percent, junk score origin.

2.5.3 Microsoft Outlook

Outlook has two e-mail searches: a simple quick search and an advanced find feature. Besides that, it also includes the possibility to find messages from people that are in some ways related to a specific e-mail (sender or receiver).

The quick search is initiated by moving the cursor inside the search box, which is situated above the messages. By default, it will include only the current folder but it is also possible to add all subfolders or even all items altogether. Simply typing in a search phrase will result in a full search, which will include all fields in the messages. When typing, options are displayed below the search box to only include some of the fields (such as subject or sender). Other options include specifying category, attachment status, time sent, recipient, flag, etc. To determine, which fields need to be searched, specific keywords are used (e.g. from("some name"), subject("some title")). The user is not expected to know these by heart, they can be inserted by clicking on relevant buttons or icons. The user can join these search words by using logical operators (AND, OR).

The other way to have specific fields searched is by adding search boxes for each of them. In addition to previous options, it is possible to search for attachment content, message size, received date, modified date, sensitivity, importance, receivers of (blind) copy, phrase in message body. All of the terms in the boxes need to be matched in a message for it to be included in the search result.

The advanced find has many of the same features that the quick search has. The main benefit it brings is adding more conditions to the fields searched. For fields, which could be classified as text strings, this includes conditions like “contains”, “is exactly”, “doesn’t contain”, “word starts with”, “phrase matches”, “is empty”, “is not empty”. For example, this means that for phrases the search term does not have to be a complete match, you can also choose to include just the beginning of the word or any part of the word. It is also possible to exclude some kind of phrases and be case sensitive in the search.

For dates there is a myriad of options including “any time”, last/next/this day (or week, month), “on”, “on or after”, “on or before”, “between”, “exists”, “does not exist”. For criteria, which only have a small set of possible values, the options are “equals”, “not equal to”, “exists”, “does not exist”.

The people search is automatically conducted when an e-mail is opened. It shows all people that are connected to this message (sender and receivers). By clicking on a person, the user can then find other messages or messages with attachments that are related to this person.[12]

2.5.4 Opera Mail

With Opera two searches are possible: a quick search and a more advanced one.

The quick search just requires typing in a phrase. This phrase is then searched in all the fields of the messages. By default only the active directory is included but it is also possible to search all messages.

To get more advanced search features, labels need to be used. After creating a new label, the rules need to be described. The fields to search from include subject, sender, recipient, copy receiver, reply-to address, message body, whole message, newsgroups header or any header. The conditions for the phrase are: contains, does not contain, matches regexp (regular expression). All of the rules need to be matched in a message for it to be included in the search result.

Finding attachments is made easier by having labels automatically in place for different types of them. There are separate labels for messages with document attachments, picture attachments, video attachments, music, archives.

2.5.5 Hotmail

Hotmail also has two types of searches — a quick search and an advanced search.

The quick search takes a keyword and searches for it in all fields in the messages. All the folders are included in the search. It is possible to use words "from:", "subject:" or "to:" to limit the search only to these fields in the message. Logical operators are supported.

In addition to the features in quick search, the advanced search allows to select a folder, choose a time period when the message was sent and determine whether a message has attachments. Logical operator OR is not supported here.[13]

2.5.6 Gmail

Gmail provides two ways to search for e-mails.

The first one is a distinct search box, which is mostly meant to be used with search operators combined with keywords. It searches all fields of all e-mails by default. When the user opens a view of a label, the condition to include only results that have this label is automatically added.

The search operators that can be used are: from:, to:, cc:, bcc:, subject:, label:, has:attachment, list: (search for messages from a certain mailing list), filename: (complete or partial attachment name), in: (specifies a search directory), is:read, is:unread, is:starred, after:, before: (some date), deliveredto: (to find forwarded e-mails). In addition to that, logical OR and exclusion (marked with a hyphen) can be used. If no logical operators are placed between different keywords or search operators, logical AND is used.

The second one provides a better user interface for dealing with search operators. The users just have to fill out relevant parts of a form, which has the following fields: from, to, subject, search directory, contained words, excluded words, has attachment, date range. All fields that have been filled out, have to be matched.[14]

2.5.7 Yahoo Mail

Yahoo mail has a very simple search method consisting of one distinct search box. This can be used to enter keywords and search shortcuts (operators). By default, all fields of all messages are searched.

The search shortcuts (operators) can be used to search only a specific part of the messages. They can roughly be divided into two classes: header shortcuts and attachment shortcuts. Header shortcuts can be used to find some specific header, such as to:, cc:, bcc:, from:, subject. Attachment shortcuts can be used to search for properties of

an attachment. The properties that can be found are: attachment name, attachment count, attachment type, attachment language, text inside attachment.

Of logical operators, AND is automatically placed between separate keywords or search shortcuts. It is also possible to use exclusion (a hyphen before a search term). Logical OR is not supported.

After conducting an initial search, the user can narrow down the results using some options that were not present before. These include folder, date and status of a message.[15]

2.5.8 Alpine

In Alpine, the search can be performed by selecting messages. Therefore it can only be used on an active directory. Selection is based on criteria defined by the user. Possible selection options are: message id, date (on, before, since), subject, from, to, cc, participant (to, from, cc), recipient (to, cc), body, all text (body and headers), status, size in kBytes, keyword, rule (defined in some filter), thread. Only one selection option can be applied at a time. It can narrow the results (if we use logical AND) or it can widen the results (if OR is used). The exclude operator also exists.

2.5.9 Summary

It can be said that the e-mail clients have very much in common.

Nearly all of them offer two types of searches: one with just a simple search box and another one with more advanced features. To extend the functionality of the simple search box, many clients use predefined operators to specify, which part of the message to search.

All of the clients search all fields by default, if none are specified by the user. Which directories are searched, is not so universal. Some clients search the active directory, some search all messages. However, this can usually be easily changed.

Of the fields that can be searched, the headers are most common and are present in all clients. These include to, from, cc, subject, body, date.

The support for finding messages with attachments is quite different. Nearly all clients support the simplest way of just determining whether an attachment is present. Searching for attachments with a specific file type or name is present in about half of the clients. More advanced options such as searching inside the attachment, determining the language and counting attachments are rarer. An overview of attachment search possibilities is given in Table 2.

Another type of common search criteria is tags (or labels, categories, stars, flags) — ways of marking the message and its state. Usually, if a client supports adding these to messages, it also allows to search for them. A special kind of marking — whether a message has been read or not — is present in about half of the clients.

As for logical operators (AND, OR, NOT), they are not evenly supported. Logical AND is explicitly used in many cases in all clients when multiple search terms are specified. It can be used to narrow down the search results. Logical OR was not supported in 2 clients. Excluding some results (operator NOT) was present in all clients. In most cases it was possible to use only one operator at a time, making the user unable to combine them.

Table 2: Support for searching attachments

attachment ...	... exists	... type	... name	... inside search
Windows Live Mail	+	-	-	-
Thunderbird	+	-	-	-
Outlook	+	-	+	+
Opera	-	+	-	-
Gmail	+	+	+	-
Hotmail	+	-	-	-
Yahoo Mail	+	+	+	+
Alpine	-	-	-	-

2.6 Sorting possibilities

Sorting refers to changing the criteria based on what the messages are currently ordered.

By default the messages were ordered by date in all clients. Normally, there were also other options, except for Gmail, which only had sorting by date.

All clients that supported other sorting options had the possibility to order by subject and sender. Sorting by e-mail size and having attachment were also quite widespread. The former was supported by 6 clients and the latter by 5 clients (out of 7). Support for previously explained sorting options is shown in Table 3.

Some other options that were present in more than one client were sorting by tag (or label, flag, etc) and whether the message has been read or not. In the case of Microsoft Outlook, it was possible to add sorting criteria by yourself.

Yahoo, Hotmail and Opera supported only the basic options, which were described before. Thunderbird, Outlook and to a certain extent Alpine and Windows Live Mail had some unique sorting options. Some examples of these are: sorting by account, importance, priority, thread, etc.

Table 3: Support for sorting

Sorting option:	Date	Subject	Sender	Size	Attachment
Windows Live Mail	+	+	+	+	+
Thunderbird	+	+	+	+	+
Outlook	+	+	+	+	+
Opera	+	+	+	+	+
Gmail	+	-	-	-	-
Hotmail	+	+	+	+	-
Yahoo Mail	+	+	+	-	+
Alpine	+	+	+	+	-

3 Search cases

3.1 Finding all messages having a certain type of attachment

This search focuses on the e-mail client's ability to differentiate between various types of files that can be added as an attachment. The types might be documents (pdf, xls, doc, etc), pictures (png, jpg, gif, etc), videos (mpg, mov, wmv, etc) and so on.

Generally it was possible to find certain types of attachments by using simple text search over the whole message. Using the file extensions as keywords, there is little chance of false positives, as they are quite unique.

For the purposes of this search case, a scenario for finding videos was used. In real life this might have multiple uses. Firstly, videos take up a lot of room in the mailbox and might be considered for deletion if disk quota on a server is reached. Secondly, videos are usually attached to more informal messages and therefore this might be a good way to find messages sent for fun and entertainment purposes.

If the client did not support searching by attachment type, the following extensions were searched for in the messages: .avi, .mpg, .mpeg, .wav, .mov, wmv. These should cover most of the widely known video formats.

The tested e-mail clients performed adequately in this test. 4 clients were able to present sufficient results: Mozilla Thunderbird, Microsoft Outlook, Gmail and Alpine.

Out of these, Thunderbird provided the most accurate information with all 21 matches found containing videos. The downside was the long search time (around 20 seconds).

Outlook was able to perform the search faster (about 2 seconds) but at the cost of many messages that did not meet the criteria. However, all the messages with videos were also present.

Gmail was able to find 6 unique messages with videos attached. With conversation view turned off, this turned into 58 messages with a lot of duplicate entries. The search speed was also the fastest.

Using Alpine with Gmail IMAP yielded the same result — 58 messages. Determining the time for search was not possible due to the unique search/select functionality. All formats had to be searched individually, every succeeding search broadening the previous selection.

The reasons of other clients failing to find videos can be divided into 2 main groups: not having a logical OR expression and inability to detect the correct file type.

Not having the OR expression proved problematic for Windows Live Mail and Yahoo Mail clients. This meant that it was only possible to search for one extension at a time, which is pretty inconvenient and time consuming. Also we never get the results on a single screen.

Inability to detect the correct file type was the problem for Opera Mail. This client has a built-in functionality for finding videos as a label/filter. It was unable to find any videos, probably because in e-mail headers the type of video was given as application/octet-stream. It was able to detect documents and pictures, though.

3.2 Finding all messages sent on a certain date

The purpose of this search case was to find out the capabilities of e-mail clients when needing to find messages that have been sent during a specific time period.

A date (and time) is added to every e-mail header as it is sent and forwarded through mail servers. In some cases, the date for sending and receiving an e-mail could be different (e.g. when the sender and receiver are in different timezones). We focused on the date, which is added when the mail is sent. It is a field named "date" in an e-mail header.

In all the tested e-mail clients, it was possible to sort the messages according to date. This means that one method for finding relevant messages is simply sorting e-mails and then manually searching until the correct time period is spotted.

Most clients also included a proper search for message dates. There were various approaches to how it was actually implemented:

1. find all messages between two dates
2. find all messages sent before/after a date
3. find all messages sent on a certain date
4. specify a certain date and an interval (+/-) in days to search

Of the clients that had a date search functionality, most used only one of those options. When multiple approaches were used, 2. and 3. were most frequently used together.

The scenario for this search case was finding messages sent on a specific date — 11 September 2001. This was the time of terrorist attack on the World Trade Center and Pentagon in the United States. We were interested in the content of e-mails sent on that date, whether they contained references to the event or not.

The results of searches were generally positive. All 5 clients that had the ability to search for date were able to find the e-mails we were looking for.

The best performers were Mozilla Thunderbird and Windows Live Mail. Both of them found all the messages without delay.

Microsoft Outlook and Alpine also found all the e-mails but the search took a bit longer. 6 seconds in the case of Outlook and 3 seconds in the case of Alpine. For Alpine, it has to be noted that after the first search (3 seconds) the subsequent ones were instant.

For Gmail, the minimum time period for a search was 3 days. This was achieved by selecting a date (11 September 2001) and an interval of one day. Therefore it found some extra messages that were sent either 1 day before of after the needed date. Despite that, all the required e-mails were also present, and there was no noticeable delay during the search.

There were two clients that did not have a date search functionality — Opera Mail and Yahoo Mail. With Yahoo Mail, using date as a search criteria became possible after some search had already taken place. Therefore it could only be used to narrow down the results.

3.3 Finding duplicate e-mail messages in folders

The goal of this search case was to find out, whether current e-mail clients have a built-in function of detecting duplicate mail.

There is not a single definition, what constitutes a duplicate e-mail message. Most third party duplicate removers enable the users to define the criteria themselves. The typical fields to compare can include subject, date, body, from address, to address, size.

There are numerous reasons, why an e-mail might be received multiple times. These can include forwarding loops, messages downloaded from a POP3 server more than once, interruptions between a mail server receiving a message and local delivery, etc.[16, 17, 18]

The results of this search case revealed that none of the tested clients currently have a duplicate finding functionality. The only option for finding and removing duplicates is installing third party programs or add-ons.

3.4 Finding attachments exchanged between persons

This search case had the aim of determining the e-mail clients' ability to find messages containing attachments that have been sent between two persons.

That means searching for all e-mails that have the user as sender and the second person as recipient or the other way round. In addition, it needs to be checked, if the message also contains attachments. Preferably, the attachment check should be part of the search. If this is not possible, sorting can also be used.

The scenario of this search case was finding all attachments exchanged between John Arnold and Ina Rangel. Arnold was a natural gas trader while Rangel was a lead assistant.[19, 20]

In this scenario 3 clients were able to find the needed e-mails.

The best performer was Microsoft Outlook that found all the messages matching the criteria in about 1 second.

Finding the e-mails was also possible with Mozilla Thunderbird and Opera Mail. There was no noticeable delay during the search. However, in both cases sorting had to be used to isolate messages with attachments.

There were two main reasons why this search was not possible with the other clients.

The first one is the lack of the logical OR operator to enable matching the same name for both sender and receiver. Yahoo Mail and Windows Live Mail belong in this group.

The second reason is inability to search or sort by attachments. Without this capability, it is very hard to find attachments from a large amount of e-mails. Alpine represents this group.

The search also failed with Gmail, although this client theoretically should support this scenario of finding messages. The reason for failure was the fact that from and to fields in most of the e-mail headers had gone missing during upload. This made it impossible to search for sender or recipient.

3.5 Creating user profiles

The purpose of this search case was to extract a user profile from a set of e-mails. As the word “user profile” is not universally defined, we had to define it ourselves.

A “user profile” is a set of personal information that can be represented as a frame. The following 10 properties were selected: name, e-mail, phone number, address, occupation, employer, birthday, education, marital status, hobbies. It is not required for all of the properties to be present, some can be missing if they cannot be found from e-mails.

In this search case the goal was to construct user profiles for the owners of two mailboxes, which were named “arnold-j” and ”allen-p”.

3.5.1 Name

Finding the full name of the owner could be achieved by reading the headers of e-mail messages. Besides the sender’s and receiver’s e-mail address they usually also contain names. Given a large volume of e-mails, it is highly probable that at least some will contain it.

Opening a random e-mail from the sent folder provided the results: the sender’s name was marked as John Arnold and Phillip Allen respectively.

3.5.2 E-mail

Discovering the e-mail address is usually a trivial task, as this should be present in all e-mail headers. Therefore only one e-mail would be required to get the information. However, a person can have multiple e-mail addresses and collect them in a single mailbox.

A search for all e-mail addresses used to send messages in the sent folder revealed two of them in the first case and three in the second case. The e-mail addresses were John.Arnold@enron.com and jarnold@enron.com for the first mailbox and Phillip.K.Allen@enron.com, pallen@enron.com and pallen70@hotmail.com for the second.

3.5.3 Phone number

Phone numbers have to be searched from message bodies as they are not required in headers. Sometimes a message body can have a signature or footer (especially for formal addresses used for communicating with customers) in which a phone number is included.

In this case, the owners of the mailbox did not have a footer or signature in the message bodies, therefore a text search using the word "phone number" was used for the mailboxes. This yielded a result of about 50 messages for "arnold-j" and looking through them it was found that the cell phone number was: 713-557-3330. For "allen-p", the result was 36 messages but none contained the actual phone number. Instead, a search term "cell phone" was used. This matched 43 messages, one of which contained the cell phone number 713-410-4679.

3.5.4 Address

Address (i.e. the location where a person lives) is also not present in headers. This might be found in a message body, such as a signature or a footer mentioned in the previous section. Using the search term "address" is not viable as it can be part of many message bodies due to similarity with the word e-mail address. Therefore some different methods have to be used.

One option is using the phone number acquired previously as the search term. If phone is present, sometimes people also add other details, such as address. This proved to be right as the phone number search returned 19 messages for "arnold-j" and some of them also had an address: 909 Texas Ave #1812 Houston, TX 77002 (10.05.2001) and 2000 Bagby #15403 Houston, Texas 77002 (19.11.2001). Possibly the person moved between these dates. For "allen-p" the search returned 24 e-mails. An address was present in one of them: 8855 Merlin Ct, Houston, TX 77055. In addition, the search revealed two previously unknown phone numbers.

3.5.5 Occupation

Occupation in this context might be just a job title, a profession or a similar name. This needs to be searched from the body as it is not present in headers. A job title is also frequently present in footers or signatures.

As the owners of the mailboxes did not use a signature or footer, a simple text search was conducted. This included the words profession, occupation, job, position. None of them provided an actual job title. The only thing we found is that both of them might be involved in natural gas trading.

3.5.6 Employer

This is closely related to the previous property. Now the main interest is the company, for which the person works. In some cases the solution might be extremely simple and just checking the domain part of the e-mail address might suffice.

That also applies to these mailboxes as the address has a domain name: enron. Therefore it is likely that the persons in question work for Enron as one would not get such address otherwise.

3.5.7 Birthday

Generally finding the birthday should be done using a text search. The word "birthday" is not overly used in e-mails and usually results in a small number of messages to look through.

In the mailbox of "arnold-j", 19 e-mails had the keyword. None of them provided a definite answer, though. It is assumed that the birthday is 9 March but it is never confirmed in any message. As for "allen-p", only 3 e-mails contained the word birthday. These were received on 25 and 27 April and contained happy birthday wishes. So the birthday is probably around these dates.

3.5.8 Education

For education, the main interest would be university name and possibly the name of the degree obtained. This could be revealed by looking at the domain names of

the persons the owner is communicating with or just by text search involving words university, alumni, etc.

Searching for words "university" and "alumni" revealed that John Arnold was an alumnus of Vanderbilt University as he received their alumni newsletter. There was no information about the degree received anywhere. Information about the education of Phillip Allen was not found.

3.5.9 Marital status

Marital or relationship status should show whether a person is currently single or has a wife/girlfriend (husband/boyfriend). To extract this info, a text search containing these keywords should be used.

According to the results (29 messages) John Arnold had a girlfriend at some point of time. This was mentioned by himself and also some of his peers. There was no sign of him ever having a wife, so he was probably not married.

Phillip Allen mentions having a wife in one of his messages. So he was probably married.

3.5.10 Hobbies

Understanding the hobbies of a person is very difficult with only a few searches. As they can be so varied, there are no easy search words to point out. However, they should mention free time, evenings, game, ticket etc. One good way is also looking through messages labeled "personal" as they are likely to contain references to possible interests.

It seems from this that John Arnold was a sports fan as he frequently attended American football, soccer, baseball games.

For Phillip Allen, the information was harder to find as he had a separate e-mail address for personal matters. Still, it looks like he was also interested in sports, basketball and American football to be exact.

3.5.11 Results

Using e-mail searches, it was almost possible to complete the predefined profile of a person.

Out of 10 properties, only 2 were of dubious value for John Arnold. For the rest, the information from the e-mails was adequate enough to be quite sure of their correctness.

For Phillip Allen 1 property could not be found and 3 were uncertain. The rest were probably correct, though.

The results of the searches, formatted as tables, are shown in Tables 4 and 5.

Table 4: Profile of John Arnold

Property	Value
Name	John Arnold
E-mail address	John.Arnold@enron.com, jarnold@enron.com
Phone number	713-557-3330
Address	909 Texas Ave #1812 Houston, TX 77002 ; 2000 Bagby #15403 Houston, Texas 77002
Occupation	Gas trader (exact job title unknown)
Employer	Enron
Birthday	9 March (uncertain)
Education	Vanderbilt University
Marital status	Had a girlfriend
Hobbies	Attending sports games

Table 5: Profile of Phillip Allen

Property	Value
Name	Phillip Allen
E-mail address	Phillip.K.Allen@enron.com, pallen@enron.com, pallen70@hotmail.com
Phone number	713-410-4679, 713-853-7041, 713-463-8626
Address	8855 Merlin Ct, Houston, TX 77055
Occupation	Gas trader (exact job title unknown)
Employer	Enron
Birthday	25 April (uncertain)
Education	Unknown
Marital status	Married
Hobbies	Sports

3.6 Summary of search cases

We tested the possibility of solving five complex search cases with the current e-mail clients. We found some clients that were able to solve search cases 3.1, 3.2, 3.4 and partially 3.5. As for 3.5 (user profiles), there were minimal differences between the clients as all the searches were text-based and included keywords. This is well supported by all clients we looked at. For the other three cases, there were bigger differences.

The performance of a client was rated as good (G) if the results of the search case were exactly as expected and adequate (A) if there were some deficiencies but the result was still achievable with a single search. If the result was not possible to achieve with the client, the performance was rated as not adequate (N). The comparison of clients for search cases is given in Table 6.

Table 6: Solving search cases with e-mail clients

Client \ Search case	3.1	3.2	3.4
Windows Live Mail	N	G	N
Mozilla Thunderbird	G	G	A
Microsoft Outlook	A	G	G
Opera Mail	N	N	A
Gmail	A	A	-
Yahoo Mail	N	N	N
Alpine	A	G	N

4 Automatic tagging by TaQuilla

TaQuilla is an add-on for the Mozilla Thunderbird e-mail client. It can be used to automatically tag existing and incoming messages by performing Bayesian analysis of the content.

The add-on supports all 5 tags that are included in the mail client: important, work, personal, to do and later. In addition, it is possible to specify for which folders to calculate the results.

Each message will get a score of 0 to 100, depending on how closely it matches the content of previously tagged e-mails. The score is used to determine whether to mark the message with the tag (if the score is higher than a certain threshold — by default 50) or not.[21]

For the experiment, automatic tagging of personal e-mails for "arnold-j" was tested. Other tags were disabled and the default threshold value (50) was used.

Firstly, the AI of the add-on needs to be trained by manually tagging some messages, which could be considered personal. It should enable the program to define some rules for automatic detection. About 50 e-mails that were not work related and contained references to personal life were tagged that way.

The program was then asked to calculate the results and tag messages. As there were no negative examples, all e-mails ended up being tagged personal. To fix this problem, the tags were removed from about 50 e-mails that were mostly work related and clearly not personal. After re-calculation nearly all messages lost their private tags, including many, which should have had it.

The process of manually adding positive examples and then removing negative ones was repeated several times. For each cycle, the detection rate of the program seemed to increase, although not as fast as maybe expected.

After about 10 cycles and 3 hours, the add-on had reached a quite acceptable detection rate. Most of the personal e-mails were found and there were not many false positives.

To determine an approximate detection rate, a small sample of 100 e-mails was chosen randomly. Each e-mail was then checked to see if the tag had been applied correctly. It appeared that out of 100 e-mails 83 were correct and 17 were not. This means that based on this sample, the detection rate was 83%.

Therefore, investing time and effort in teaching this add-on eventually pays off.

Conclusion

In this paper, an overview of the search features in some of the most popular e-mail clients was given. This included comparing their search visualization techniques, search and sorting speed, message tagging, search and sorting parameters.

Conducting some complex searches was also tested with the same e-mail clients. The results showed that finding messages with certain types of attachments, messages sent on a certain date and messages with attachments exchanged between two people were possible. On the other hand, creating user profiles was only partly possible and finding duplicate e-mails was not possible.

Finally, we also tested an application, which automatically tags messages. It took a long time to set up properly but in the end managed an acceptable detection rate for personal e-mails.

In general, all the goals that were set in the beginning were met. For future work, the selection of e-mail clients could be widened because none of Mac or Linux clients were tested in this paper. Furthermore, it would be sensible to include more add-ons in the future because they have interesting features that the clients do not support yet.

Meiliklientide analüüs ja võrdlus

Mart Sein

Bakalaureusetöö (6 EAP)

Sisukokkuvõte

E-mail on tänapäeval üks olulisemaid kommunikatsioonivahendeid Internetis. See võimaldab teateid vahetada erinevates asukohtades ning ajavööndites viibivate inimeste vahel. E-mailide hulga kasvades muutub nende haldamine üha raskemaks.

Suurest hulgast e-mailidest vajaliku info leidmine ei ole kerge. Seetõttu on tänapäeval enamasti meiliklientidesse sisse ehitatud vahendid, millega e-mailide haldamist ja neist info leidmist lihtsustatakse.

Hetkel võimaldavad meilikliendid enamasti vaid tekstiotsinguid. See tähendab, et võimalik on leida ainult infot, mis otseselt kirjas on. Meilikliendid ei suuda ridade vahelt lugeda ja erinevaid infokilde omavahel seostada. See tähendab, et väga raske on vastata küsimustele nagu "Kellele ma veel pean vastama?" või "Kas see teade on seotud minu tööga?". Sellistele küsimustele vastamine nõuaks programmilt kirja konteksti ja tähenduse täielikku mõistmist.

Kuna hetkel on meiliklientides kirjadest aru saamise tehnoloogia alles lapsekingades, tuleb eelnevalt mainitud küsimustele vastuseid leida praegu kasutada olevaid meetodeid kasutades. Käesolevas töös selgitati välja, millised otsinguvõimalused tänapäevastes meiliklientides leiduvad. Analüüsiti otsingutulemuste visualiseerimist, otsingu ja sorteerimise kiirust, kirjade märgendamist ning otsingu ja sorteerimise parameetreid.

Neid võimalusi kasutades prooviti lahendada mitmeid komplekssseid otsingujuhtumeid: mingit kindlat tüüpi manuste leidmine, e-mailide duplikaatide leidmine, kindlal kuupäeval saadetud kirjade leidmine, kahe inimese vahel saadetud manustega kirjade leidmine, kasutajaprofiilide koostamine. Uuriti, millised otsingujuhtumid on praeguste klientidega lahendatavad ja millised mitte. Leiti, et komplekssete otsingute teostamine programmide poolt pakutavate vahenditega on võimalik, kuid eeldab üsnagi suurt tööd inimeste poolt. Eelkõige on probleemiks õigete otsinguvahendite ja fraaside leidmine ning see, et suur osa tööst tuleb teha manuaalselt.

Lisaks uuriti üht hetkel saada olevatest rakendustest e-mailide automaatseks märgendamiseks, milleks oli Thunderbird meilikliendi lisa TaQuilla. Meid huvitas selle veaprotsent ning treeninguks kuluv aeg. Tulemused olid meie jaoks rahuldavad: kuigi treening võttis mitu tundi, tundis programm ära rohkem kui 80% isiklikest kirjadest.

References

[1] S. Michel and I. Weber. Rethinking Email Message and People Search. 2009.

[2] M. Dredze, H. M. Wallach, D. Puller, T. Brooks, J. Carroll, J. Magarick, J. Blitzer, and F. Pereira. Intelligent Email: Aiding Users with AI. http://www.cs.jhu.edu/~mdredze/publications/nectar_intelligent_email.pdf, 2008. Last visited 24.05.2011.

[3] About Email. http://email.about.com/. Last visited 24.05.2011.

[4] Best Email Client Software. http://emailsoftwarepro.com/. Last visited 24.05.2011.

[5] V. Rangan. Discovery of Related Terms in a corpus using Reflective Random Indexing. http://www.umiacs.umd.edu/~oard/desi4/papers/rangan.pdf, 2011. Last visited 24.05.2011.

[6] R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu. Using Structured Queries for Keyword Information Retrieval. http://www.almaden.ibm.com/cs/projects/avatar/pubs/semsrchtechreport07.pdf, 2007. Last visited 24.05.2011.

[7] E. Minkov and W. W. Cohen. Improving Graph-Walk Based Similarity with Reranking: Case Studies for Personal Information Management. http://www.cs.cmu.edu/~wcohen/postscript/tois-ghirl.pdf, 2010. Last visited 24.05.2011.

[8] Email client market share statistics. http://litmus.com/resources/email-client-stats. Last visited 24.05.2011.

[9] List and types of email clients. http://en.wikipedia.org/wiki/List_of_e-mail_clients. Last visited 24.05.2011.

[10] Enron email dataset. http://www.cs.cmu.edu/~enron/. Last visited 24.05.2011.

[11] EDRM Enron email data. http://edrm.net/resources/data-sets/edrm-enron-email-data-set-v2. Last visited 24.05.2011.

[12] Using Outlook search. http://office.microsoft.com/en-us/outlook-help/find-a-message-or-item-by-using-instant-search-HA001230585.aspx. Last visited 24.05.2011.

[13] Hotmail search. http://www.freeemailtutorials.com/windowsLiveHotmail/hotmailTipsAndTricks/hotmailEmailSeachFindEmails.php. Last visited 24.05.2011.

[14] Using Gmail advanced search. http://mail.google.com/support/bin/answer.py?answer=7190. Last visited 24.05.2011.

[15] Yahoo mail search. http://help.yahoo.com/l/us/yahoo/mail/classic/manage/manage-01.html. Last visited 24.05.2011.

[16] Outlook Duplicate Causes and Prevention. http://www.anti-dupe.com/products/prevent_outlook_duplicate.aspx. Last visited 31.05.2011.

[17] Why do I get duplicate emails? http://kb.mediatemple.net/questions/774/Why+do+I+get+duplicate+emails%3F. Last visited 31.05.2011.

[18] Indiana University Knowledge Base. http://kb.iu.edu/data/ahph.html. Last visited 31.05.2011.

[19] John D. Arnold. http://www.forbes.com/lists/2007/54/richlist07_John-Arnold_2ATC.html. Last visited 24.05.2011.

[20] Ina Rangel. http://www.linkedin.com/pub/ina-rangel/8/9ab/a9a. Last visited 24.05.2011.

[21] TaQuilla homepage. http://mesquilla.com/extensions/taquilla/. Last visited 24.05.2011.