# Information Retrieval in Mail Archives

No Author Given

No Institute Given

**Abstract.** Mail archives are mail boxes storing large activity over a long period, and they are the subject of our research. Mail archives, including attached files, store relevant information since they keep memory of the user flow of information with the environment. Since the size of mail reservoirs frequently reaches the order of Gbytes, information retrieval schemes on Mail Archives should benefit from the reach structure embedded in the relation between messages. We propose a web-like structure for Mail Archives and we present first steps towards the design of efficient search engines to recover mail information that may use the structure embedded in Mail Archives. The application developed to transform mailboxes into the HTML Mail Archive is offered to the community to collect a corpus of actual usage of mailboxes storing large volumes of messages.

## 1 Introduction

The increasing use of e-mail as a standard tool for communication gives a relevant role to mail reservoirs. The messages are normally stored in mail boxes of client software connecting regularly with mail servers. As the use of e-mail spreads, the activity of a regular working week for a person relying on e-mail for information exchange can be of the order of a hundred messages. This activity would add up to some thousand messages if a complete year is considered, and the storage for a complete year of every letter, with the attached files, may build up a reservoir of hundreds of Megabytes. Even if clearly irrelevant letters are deleted regularly from the mail box, the size of mail boxes grows to large volumes of information over a long period.

Some users consider mail boxes like gateways from which information needs to be classified and stored away regularly. But some users do like to keep the messages in the mail boxes to leave unspoiled all the structure inherent to the relation among messages. Such structure does retain relevant information since it keeps the threads followed in the conversations. We call "mail archives" to mail reservoirs where mail activity over long periods is stored, leaving intact the structure inherent to the mail exchange and prepared to browse and retrieve the information. Up to now, most work in mail archives has addressed the activity of mail lists but in our opinion mail activity of individuals is reaching a volume to deserve attention. No need to say that the authors started this work prompted by their personal wish of having at hand a suitable and efficient way to store and dig the large volume of information associated to their own mail.

If mail is the common gateway of our computer with the outside world, the information stored in mail archives gets naturally filtered with respect to relevance. In our opinion, mail archives may become in the future the main reservoir where information is stored and retrieved in personal computers. Furthermore, mail archives may also play an important role in enterprise knowledge management if employees' mail is aggregated in a common reservoir to keep track of commercial relations. When used at work, mail boxes have also a record of relations with clients, supervisors, colleagues, etc. and attached files can be documents having the same value as documents faxed or mailed over regular mail.

Some work has been done previously on mail archives and information retrieval. Mailcat [16] is a Lotus Notes add-in that helps the user to classify e-mail by presenting the three folders that are more likely to be appropriate, using an adaptive classifier. Another e-mail classifier, based in rules, is shown in [11].

There are several commercial products, either stand-alone applications or plug-ins for other mail user agents, that can help the user in the task of searching huge amounts of e-mail, such as EMO [6], EZDetach [17] and 80-20-Retriever [1]. In the area of visualization, TimeStore [15] presents mailboxes in a tabular form with dates and persons as the axes. There exists, also, software to translate the archive of an e-mail discussion list to HTML, such as MhonArc [10].

We consider mail archives independently of mail clients, as a stand alone reservoir, where recent mails are added from time to time, but in which the underlying structure is used to improve the possibilities of browsing and searching for information retrieval. From this point of view, a mail archive needs to have:

- a standard format to be stored and used over long period of times, a format as independent as possible from changes in hardware or operating systems.
- a structure and organization that facilitates browsing to look for the information using all kind of relevant indexes.
- an efficient search engine to retrieve the information from message bodies and attached files. The search engine should benefit from an analysis of the overall set of messages.

In the present work we propose a web format for mail archives where the hypertext provides a suitable way for browsing during information search. A Java application has been developed to automatically transform a standard Unix or Mozilla compatible (like Netscape) mailbox into the proposed web structure for mail archives. Also, the problem of building an efficient search engine is approached.

A first step towards an efficient search engine is to gather information on the common structure of mailboxes. Efficient World Wide Web search engines use the structure of the Web to rank documents. The study of the structure of the Web has allowed the development of efficient algorithms for document retrieval that use information embedded in the relation among web pages and web sites. Efficient algorithms for documents retrieval over mail archive may also use the relations between messages and their authors. However, no work has been done yet on the analysis of the structure of mails reflected by the actual content of mailboxes.

There are no widely available corpora of e-mail messages at this time, and e-mail messages are very different from other documents, as they are more similar to an hybrid between speech (conversation) and text [19]. In the last few years, a corpus of spam messages has been assembled by the Spam Archive [18].

The main reason for this absence is the confidential character inherent to mail messages. Web sites are publicly exposed to the community but, mail boxes are kept well protected by the owners. So, although it seems unreasonable to build corpora of e-mail messages, it seems feasible to build a corpus of files with "anonymous" data related to the structure of mailboxes. The Java application developed in this work is offered to be downloaded and to be used to convert mailboxes into the proposed web solution. The application does produce also a plain text file with a register for every mail in the original mailbox but without any kind of personal information. Users downloading the application are requested to send back the files in order to build such a corpus.

This communication presents a concrete proposal for organization of mail archives as a set of html pages with hyperlinks, a plan to build up a corpus of data with useful information about mail usage, and a first step towards the development of an efficient search engine to retrieve information on mail archives. The communication is organized as follows, Section 2 provides a framework of relation between mail archives and the problem of information retrieval. In Section 3, the proposed web-like structure for mail archives is analysed. In section 4, algorithmes to automatically cluster the messages are presented. Section 5 is devoted to discuss the structure underlying mailboxes, and Section 6 to describe the details of the application developed to transform mailboxes into the web-like mail archive. Finally, Section 7 is the discussion and it includes the lines guiding the planned work towards the development of useful mail archives.

## 2 Searching information in mail reservoirs

Common Mail User Agents (MUAs) such as Mozilla Mail or Microsoft Outlook, usually provide some search capabilities, but the underlaying search engines disregard important information such as attachments and relationships between people. Some of the tasks of mail information retrieval are:

*Ranking* The TF-IDF ranking in the vector space model (see [3]) is not appropiate for e-mail messages as very short messages have high rankings -the *term frequency* of terms in the query will be high- but are not relevant as they have little informational content. Also, ranking should consider the relative importance of persons that are involved on a message with respect to the user.

*Grouping* In the same way as search engines such as Google [9] group the results from the same site, we would like to group results from the same person, or from the same cluster of persons. Also, sometimes a single message is not important, and a complete discussion thread is the relevant content unit.

*Filtering* A mail message has several fields that could be used for filtering, such as date, persons, domain names or clusters of persons.

*Browsing* Keyword–based search is usually only the first step towards finding a relevant piece of information, as users tipically use search results as a starting point for browsing [12]; as far as we know, there are no tools available that allow a user to browse e-mail messages except for normal indexing by address or date.

## 3 A mailbox as a Web

We envisage a mail archive as a web structure, where the content of the user mail-box is recorded as a set of html pages with hyperlinks. These pages and links will be all inside a single site, but it can be compared to the whole web if we substitute web sites by users. Then each site (user) has its own pages (a home page with user information and a page for each message that has been sent by that user) which contain hyperlinks to other sites (home pages of users who have received that mail), a site also includes a "Bookmarks", "favourites" or "my links" page, (which is a list of hyperlinks, one for every message received by the site owner, and pointing at his mail page). This means that every message has its own page ("Mail page") included in the site of the user who sent it ("Mail author"), and contains links to users who have received that mail and is pointed by the users who received that mail. The "Mail page" contains not only the text but also the files attached to the message and the links to users to whom it was sent. The "Received Page" contains a list of links to the messages received by that user.

This parallelism between both structures, allows to adapt some of the search and ranking strategies developed for web search engines to find and rank the result of a search in mail archives. It also allows to evaluate the *value* or relevance of every user with respect to the mailbox owner. However, some differences arise:

- All links are symmetric, as for each mail there are the same outbound links pointing to users who have received the mail as inbound links from those users.
- Every "mail page" of a user not being the centre (the mailbox owner) has a link to the central user, because only mails sent to him are in the mailbox. If a mail does not include the owner, means that is a mail from a distribution list and should be considered separately .
- The size of a site (number of pages) gives an indication of the user activity, but the value is partial: it is about activity with the mailbox owner. A single mailbox is used to build the mail archive. However, the value is useful since relevance is also judged from the same point of view: that of the mail archive owner.
- The number of inbound or outbound links indicates the degree of *connectivity* of that user with other actors appearing in the mailbox.

- The links between two users indicates how many mails share with the owner of the mailbox but does not indicate the traffic between them, as they can share other messages without sending copy to the central user. However, this information is crucial in order to rank users with respect to the central user.
- The network structure is biased as only contains the information seen from one mailbox. The resulting network is different from the global one that could arise considering all the mailboxes from all the users/actors appearing once in the mailbox under scrutiny. The only information we have at hand from those networks are messages sent to or copied to the "central" owner. For example, if two users appearing in the mailbox are lovers, and send love messages between them, they will not sent a carbon copy to the owner of the mailbox. Thus, from the point of view of the centre mailbox, these two users may be softly related; a conclusion surely rejected after analysis of any of their mailboxes. Whereas the world wide web has a unique structure, every mailbox gives a different view of the relationships between the users appearing in a mailbox.

A rank function must take into account these particularities. Making a parallelism with Google[9] PageRank[1] , the site with higher page rank is the centre. And all their pages have a high rank The rank of a given user appearing in the mailbox is a weighted sum of the rank of "mail pages" that point to him. "Mail pages" are important if mail is sent to important people, as it receives a link from the "received pages" of users to whom the mail is sent. A "received page" has the same rank as the user it belongs to. A user is important when it receives many mails from important people (and the centre is the most important one).

## 4 Aggregation of related messages

The complete set of messages in a user mailbox defines a complex network of relationships between the members of his address book. The aggregation of nodes in such a network is a way of simplifying our representation of the mail archive for browsing purposes. The grouping of nodes can also help to evaluate the relevance of messages in the search engine since some messages can inherit their rank from the users in the same group. A detailed analysis of the relation between nodes in the network can benefit from the results delivered by the search of web communities [8] and it will be used once the corpus of mail usage is more advanced. As a first approach we set up a framework to use clustering algorithms to group messages and users.

Every mail that is sent to, or with copy to, different users indicates some kind of relationship between them. Thus we could say that every mail defines a group of users with something in common. One way to simplify this network is to perform clusters of users that may be seen as a group from the mail-centre.

[1] PageRank is trade mark of Google

*Clustering of messages* A first approach could be the clustering of messages, and then for each cluster we could find the users involved in the messages that belong to it. Any clustering algorithm needs to define a measure of distance between the individuals to be clustered. As the objective is to use mails to cluster users, the distance between two mails must be proportional to the similarity in the set of users acting in both messages. A metric space of dimension $U$ can be defined, where each mail is represented as a vector $V_i$ with $U$ components, one for each user in the mailbox; this vector is:

$$V_i = u_1 u_2 \dots u_U \tag{1}$$

$$u_i = \begin{cases} 0 \; if \; i \notin mail \\ 1 \; if \; i \in mail \end{cases} \tag{2}$$

The distance between messages, computed as the Euclidean distance, gives the square root of the number of users differing from one message to the other.

Once the distance measure is set, most of the cluster algorithms can be applied to messages. The centroid of a cluster will be a $U$ dimensional vector indicating the relationship of each user to the cluster, and thus each user can be assigned to the cluster with a maximum ratio of pertinence.

When applying this clustering technique two main drawbacks arise: first the dimensionality, we are faced with a matrix $M \times U$, and secondly there is a second task: extracting groups of users from groups of mails

*Clustering of users* A better approach is to cluster users directly, but a metric space must be defined to apply, for example, k-means. We can transpose the matrix and use a $U \times M$ matrix. With this approach the size of the matrix is the same, being still an expensive task to compute these distances.

Finally we can compute the correlation between users and apply clustering techniques that use correlation. To do so we need a measure of correlation $c_{i,j}$ between two users $i$ and $j$. A simple one can be:

$$c_{i,j} = \frac{m_{ij}}{m_i + m_j - m_{ij}} \tag{3}$$

Where $m_i$ is the number of messages where the user $i$ is a *ToUser* or *FromUser* while $m_{ij}$ is the number of messages where both users appear as sender or addressee. This correlation formula is symmetric, and $c_{ii} = 1$ . The correlation between two users is determined by the user who has more messages.

Using this (or any other) correlation measure we can apply the McQuitty's elementary linkage analysis [14]. With this simple clustering method, we attach users to the one with whom they are more related, cutting clusters at local minimums. This algorithm has the advantage of having a number of clusters that depends only on data, and does not need, like dendograms or k-means a prior knowledge. This allows to any user to run the algorithm without no knowledge of the clustering algorithm. Using K-means, the user of the clustering technique needs to find the correct K, number of clusters, or in a dendogram the cutting height.

Some modifications have been done to the basic McQuitty algoritm. First, only users "highly" correlated with the group are taken into account to include new users. If the maximum correlation of the group is $C_1$ then, users with the maximum correlation with a group member being under $C_1 - 0.30$ are not considered "highly" correlated but included in the cluster. A user $i$ with a correlation coefficient $C_{ij}$ with any group member $j$ such that $C_{ij} > 0.9 * \underset{k \neq i}{Max} (C_{ik})$ is included in the cluster as a "guest", but can be included in other clusters.

The algorithm has been applied to a reduced set mailboxes with more than 30,000 messages, and the mailbox owners have collaborated to grade the quality of the clusters delivered. The results have been found "sensible" by mail owners, since every person included in the same cluster had a "reason" to joint the party. However, quite often the testers found that there were some other persons that should have been invited to the same party. Once the corpus of mail usage builds up, the algorithm will be tuned and used in order to group messages for simplification of browsing through the mail archive and to improve relevance in the search engine.

## 5 The structure embedded in mail archives

Mailboxes are a partial view of the world. They have messages sent to a single user we call the mailbox centre. There is a set of mailbox "users", being them those individuals sending directly the message or copying it. For all purposes, in what follows, we shall not consider differences about a user being included in the "to" or in the "cc" field. We understand that quite often, we sending mail to a group of people, receivers are written randomly in any of the fields. We make the distinction between senders and receivers with the obvious meaning: a "sender" is a user appearing in the "from" field, and a receiver appears in the "to" or in the "cc" field.

Mailbox users can be ranked by their presence in the mail box as sender or/and receivers. Normally, we should expect the top sender or receiver to be the centre, but this is not always the case. In general, the centre does receive more messages than it sends, and it should be frequent to find mail boxes with a top sender not being the centre. However, the top receiver should be always the centre. We are using such criteria to select the mailbox owner when getting anonymous data from the application to transform the mailbox into the mail archive.

In a mailbox with $N$ mails, there is a number of messages $N_m > N$, since we consider a message to be a communication between one sender and one receiver. For the purpose of describing the structure, we define a (sparse) matrix $U$ with element $u_{ij}$ being the number of messages sent from user $i$ to user $j$. The dimension of the non-symmetric matrix $U$ , is $n \times n$, being $n$ the total number of users (the set union of senders and receivers sets). If we consider the mailbox to be a graph, the number of connections between users is $N_c$ and it is the number of non zero elements in $U$. We use $n_s$ for the number of senders and $n_r$ for the number of receivers. It is expected to find $n_r >> n_s$ since in the mailbox there will be many users included in distribution lists that never send a single message to the centre.

It is interesting to consider a subset of the mailbox where all "silent" users are removed: the users that never sent a mail to the centre. With this mailbox subset a new matrix of active users, $W$, shall be considered and the dimension reduces to $n_w \times n_w$. The number of connections and messages are $N_{cW}$ and $N_{mW}$, respectively. In matrix $W$, the number of receivers, $n_{rW}$, is equal to the number of users $n_W$, since it contains bi-directional activity.

Although the mailbox stores message conversation with the centre, it also reflects conversations among other users being copied to the centre. In order to find out the rank of users it is interesting to study a different matrix, $D$, in which the central user is not considered. Activity included in $D$ does not include the messages sent or received by the mailbox owner. For this subset, the number of connections, messages, sender and receivers are $N_{cD}$, $N_{mD}$, $n_{sD}$ and $n_{rD}$, respectively. The activity reflected by $D$ does allow to build a graph in which all links to or from the central user node are not considered. The number of links in this graph, $N_{cD}$, is smaller than $N_c$, but it remains being large. A step further towards identifying *hubs* and *authorities* among mailbox users is to consider a more reduced matrix $E$, extracted from $D$ by elimination of silent users in the subset. That is, elements $e_{ij}$ of matrix $E$ have sent at least one mail and are iteratively found from elements in matrix $D$: users that have sent messages to a user that never has sent one, are also removed. The activity matrix $E$ has a dimension $n_E$ much lower than the original one, but does reflect conversations being observed by the mailbox owner in which he, or she, does not participate directly. Top active users derived from matrix $E$ are the best estimation of *authorities* that can be done from the partial knowledge of the outside world provided by the mailbox of a single user. A top activity user in $E$, not well ranked in $U$, would correspond to a user with whom many of the user related to centre do exchange messages although the direct activity with the centre is not relatively high. For the subset considered in $E$, the number of connections, messages and receivers are $N_{cE}$, $N_{mE}$ and $n_{rE}$ respectively.

The relatively large activity remaining in $E$ was not expected and it was checked not to be an artefact coming from the fact that a mailbox owner can use more that one address to send or receive mails. In the small subset considered in this preliminary study, mailbox owner were asked to identify their one activity specifying the addresses from where they sent or received mail and the information was considered to calculate matrix $D$ from matrix $U$.

In the following table, some values of the activity for a reduced set of six mailboxes is presented. Although one is tempted to derive conclusions from them, the set is too reduced and the values are given as a mere illustration. Whenever the corpus of mail usage information under recollection will get a sensible size to perform statistics, we will be able to derive general conclusions from this kind of analysis.

| user | VL | CC | CR | FC | MO | JH |
|---|---|---|---|---|---|---|
| $N$ | 8047 | 6786 | 4871 | 13176 | 2989 | 12464 |
| $n$ | 3771 | 2481 | 2919 | 1400 | 4344 | 2512 |
| $N_m$ | 42671 | 10965 | 15628 | 28006 | 10922 | 25066 |
| $N_c$ | 18284 | 3603 | 4507 | 3439 | 3559 | 5005 |
| $n_s$ | 479 | 663 | 1016 | 694 | 642 | 720 |
| $n_r$ | 3378 | 2081 | 1951 | 837 | 1597 | 1969 |
| $n_W$ | 472 | 585 | 933 | 341 | 445 | 535 |
| $N_{cW}$ | 3379 | 1437 | 2040 | 1974 | 1174 | 2016 |
| $N_{mW}$ | 17198 | 7426 | 9254 | 17221 | 4944 | 11425 |
| $n_{rW}$ | 325 | 398 | 239 | 183 | 172 | 288 |
| $N_{cD}$ | 17739 | 2014 | 3618 | 3071 | 3223 | 4993 |
| $N_{mD}$ | 36536 | 3138 | 13084 | 20656 | 9375 | 21478 |
| $n_{sD}$ | 262 | 248 | 421 | 405 | 467 | 720 |
| $n_{rD}$ | 3364 | 1347 | 1940 | 836 | 1594 | 1968 |
| $n_E$ | 232 | 128 | 250 | 341 | 236 | 535 |
| $N_{cE}$ | 2244 | 360 | 1059 | 1974 | 793 | 2016 |
| $N_{mE}$ | 9639 | 814 | 6408 | 17221 | 3324 | 11425 |
| $n_{rE}$ | 208 | 104 | 195 | 183 | 146 | 288 |

Being mail interchange a manifestation of social activity, the structure embedded in mailbox should reflect the same kind of properties found in other social networks. In the case o Web search, this structure has opened the way towards new specifics algorithms of information retrieval [13][5]. We have checked if a scale free[4] ordering in the activity of mail box users can be expected. And the answer is positive.

As one suspects while using regularly e-mail services, a reduced set of people is sending most of the received messages. In all the data we have analysed, the ordering of activity in users, considered as senders, receivers or both, is such that the number of users $u$ (senders or receivers) exchanging $m$ messages decreases as:

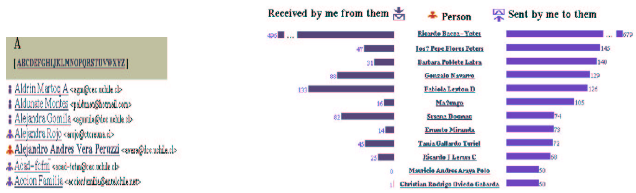
$$u = \frac{A}{m^{\alpha}}$$

with $\alpha$ bigger than one, showing up a scale free structure. Although, again, one is tempted to conclude general $\alpha$ exponents for the mail activity with such reduced set of data, the conclusions will be deferred to calculations over a larger set of data once a larger corpus of mail activity information is collected.

## 6 Visualization and interfaces

We have developed a prototype that converts a mailbox with standard RFC 822 [7] e-mail messages to a personal web site suitable for browsing and searching via a Java applet using Lucene [2]. This prototype ask the user for the folder in his filesystem where its mailbox is stored, and generates a set of static web pages that form a web site. No client-server activity is required.

The generated web site has one page for each message, and three main indices: persons sorted by relevance, persons sorted in alphabetical order, and a calendar.

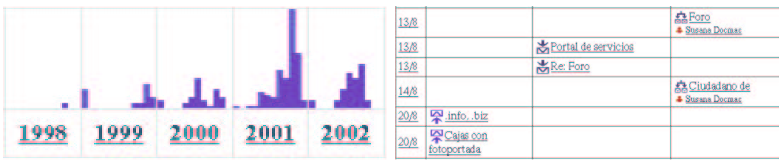



Person Index (alphabetical)    Person Graph (by relevance)

**Fig. 1.** Persons are shown in two indices. The person index (left) shows persons in alphabetical order, with icons representing their relative importance. The person graph (right) shows bar representing incoming and outgoing total traffic.

To sort persons, we consider the exchange of messages with the owner of the mailbox as an indicator and relevance. We use different icons and colors to highligh people that receives 50% and 80% of messages, as shown in Figure 1.

Also, there is a page for each person and a page for each week. In the page of each person, a monthly activity graph is shown and all the messages associated to that person, as shown in Figure 2.




Monthly activity    Message list

**Fig. 2.** For each person, a monthly activity graph (left) is shown, and messages are shown in a three-column layout of sent, received, and copied (right).

## 7 Discussion

The aim of this paper is to recall the attention of the information retrieval community towards the problem of Mail Archives and to communicate the availability (www.XXX.XX/XX) of an application converting mailboxes into useful mail archives. If both objectives are achieved, we expect collaboration of users sending back to the authors the mail activity data file produced by the application. As far as we are successful building a corpus of mail activity data, researchers will have at hand a source of information very useful in order to design efficient algorithms for information retrieval on mail archives.

The web like mail archive we propose has some useful features in order to browse and search for information. Besides that, the fact of being delivered as a tree of html pages, does allow for it being used remotely since mail archive owners can directly serve it (to themselves) remotely using a web server. Also, it can be stored in a CD or DVD and then, can be consulted from any computer using just the browser. Work to check the validity of search engines specially designed to search in the web over the proposed html mail archive is under progress

We have also given a first analysis of the structure of mailboxes and of clustering capacities over mail archives. All the results are pointing to the possibility of building efficient search engines for mail archives. In order to proceed further, more data are needed and we hope to have them at hand once the intended corpus builds up.

Finally, organizational memory is a broad term applied to all the knowledge of an organization. In [20], a series of components of this memory are listed: individuals, culture, transformations, structures and ecology; individuals are the most important source. E-mail plays a very important role here, because it keeps a record of the communication between individuals, and as far as e-mail usage increases mail archives will store structural data relevant to advance in the knowledge of social organization. If the scale free structure of the users activity gathered in our first analysis is confirmed, studies on mail archives will deliver to a broader community scale exponents, maybe universal, to be used for research in many branches of social sciences.

## 8 Acknowledgements

## References


1. 80-20 software: Mail Retriever. Commercial Product (2002) `http://www.80-20.com/products/retriever/`
2. Apache Software Group: Jakarta Lucene text search engine. GPL Library (2002) `http://jakarta.apache.org/lucene/`
3. Baeza-Yates, R. and Ribeiro-Neto, B: Modern Information Retrieval. Addison-Wesley-Longman (2001).
4. Barabasi A., Albert R.: Emergence of scaling in Random Networks" Science **286** (1999) 509–512.
5. Brin S., Page L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proc. 7th Int'l World Wide Web Conference (WWW7), Computer Networks and ISDN Systems, **20(1–7)** (1998) 107–117.
6. Caelo Software: Nelson E-Mail Organizer. Commercial Product (2002) `http://www.caelo.com/`
7. Crocker, D.H.: Standard for Arpa Internet Messages. IETF Standard (1982) `http://www.ietf.org/rfc/rfc0822.txt`

8. Flake G. W., Lawrence S., Giles C.L., Coetzee F.M.: Self-organization and Identification of Web Communities. IEEE Computer 66, March 2002.
9. Google Search Engine (2002) `http://www.google.com/`
10. Hood, E.: E-Mail messages to HTML conversion. GPL application. (2002) `http://www.mhonarc.org/`
11. Itskevitch, J.: Automatic Hierarchical E-Mail Classification Using Association Rules. Master Thesis, Simon Fraser University (2001) `http://fas.sfu.ca/pub/cs/theses/2001/JuliaItskevitchMSc.ps.gz`
12. Jansen, B.J. and Pooch, U.: Web user studies: A review and framework for future Work. Journal for the American Society of Information Science and technology. **52(3)** (2002) 235–246.
13. Kleinberg J.M.: Authoritative Sources in a Hyperlinked Environement. J. ACM. **46 (5)** (2000) 604–632
14. McQuitty, L.L.: Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. Educational and Psychological Measurement **17** (1957) 207–229.
15. Kelvin, Y. and Baecker, R. and Silver, N. and Long, B.: A Time-Based Interface for Electronic Mail and Task Management. Proceedings of Human–Computer Interfaces (1997) 19–22
16. Segal, R. and Kephart, J.O.: MailCat: An Intelligent Assistant for Organizing E-Mail. International Conference on Autonomous Agents (1999) 276–282. `http://www.research.ibm.com/people/r/rsegal/mailcat.ps`
17. TechHit: EZ-Detach for Outlook. Commercial Product (2002) `http://www.techhit.com/ezdetach/`
18. The Spam Archive. Website. `http://www.spamarchive.org/`
19. Uhlíøvá, L.: E-mail as a new subvariety of medium and its effects upon the message. In S. Mejrková & P. FrantiĐek (Eds.) The Syntax of Sentence and Text (1994) 273–282.
20. Walsh, J.P. and Ungson, G.R.: Organizational memory. Academy of Management Review **16** (1991) 57–91.