

Information Filtering

Jacob Palme

Department of Computer and Systems Sciences
Stockholm University and KTH Technical University
Skeppargatan 73, Stockholm
Phone: +46-8-16 16 67, Fax: +46-8-783 08 29
E-mail: jpalme@dsv.su.se

Abstract: Better ways at finding the most valuable information on the Internet, and to avoid trash, would very much enhance the value of the network. This paper makes an overview of methods and problems in this area, including social filtering, where people help each other with filtering objects on the net.

This document is available at the following URLs:

In HTML format: <http://www.dsv.su.se/~jpalme/select/information-filtering.html>

In PDF format: <http://www.dsv.su.se/~jpalme/select/information-filtering.pdf>

INTRODUCTION

Information Overload, Quality Enhancement

Much of the information on the Internet today consists of documents made available to many recipients through mailing lists, distribution lists, bulletin boards, asynchronous computer conferences, newsgroups, and the World Wide Web.

Common to mailing lists and forums is that the originator of a message need only give the name of one recipient, the name of the group (mailing list, bulletin board, computer conference, forum, closed group, etc.) The messaging network will then distribute the message to each of the members of the group, with no extra effort for the originator. The average effort of writing a simple message is about four minutes, and the average effort of reading a message is about half a minute [Palme 1981], so if there are more than about eight recipients to a message, the total reading time is larger than the total writing time, and if there are hundreds or thousands of recipients, the total reading time caused by the originator is many times larger than his effort in writing the message.

Because of this, Internet users will easily become overloaded with messages [Denning 1982, Palme 1984, Hiltz and Turoff 1985, Malone 1987]. This issue can also be seen as a quality problem:

people want to read the most interesting messages, and want to avoid having to read low-quality or uninteresting messages.

Filtering is tools to help people find the most valuable information, so that the limited time spent on reading/listening/viewing can be spent on the most interesting and valuable documents. Filters are also used to organize and structure information. Filters are, for most users, more important for group messages (messages sent to mailing lists and forums) than for individually addressed mail. Filtering is also needed on the search results from Internet search engines. Future software for the Internet can be expected to employ more advanced and user-friendly filtering functions than today, in order to support less computer-specialist users. Since people download millions of messages and web documents every day, and very often do not immediately get what they would mostly like to get, the gains through better filtering are enormous. Even a filter with a 10 % efficiency gain, the gain would be worth billions of dollar a year.

Before the Internet

Human society has always employed methods to control and restrict the flow of information. When this is done to satisfy the needs of the government, it is named censorship. But most of this control in democratic countries is done to satisfy the needs of the recipients. Publishers, journalists, editors

provide an accepted service of selecting the most valuable information to their customers, the readers of books, journals, newspapers, the radio listeners and the television viewers.

Schools and universities select which information to teach the students based on scholarly criteria. The intention is again to help the customers, the students, to get the most out of a course. Political organizations select what information is discussed in their organizations and distributed to their members.

This control of the information flow is done in the interest of many groups. Politicians want to control what information is given about their activities. The establishment wants to control information flow to protect itself and to control society. The scientific community wants to control information to uphold scientific quality, but has also many times tried to restrict novel research outside of the established paradigms. So control of information flow is not only done to help recipients of information.

What is different with the Internet?

On the Internet, almost anyone can easily and at low cost publish anything they want. This means that a vast amount of information of varying quality is disseminated. There are lots of interesting things, but also lots of trash. (Not that everyone agrees on what is interesting and what is trash, of course.) Can the Internet develop tools to help its users find the most valuable and interesting information? Should this be done, on the Internet, using the same methods as in the pre-internet society, or can novel methods be developed?

Major filtering methods

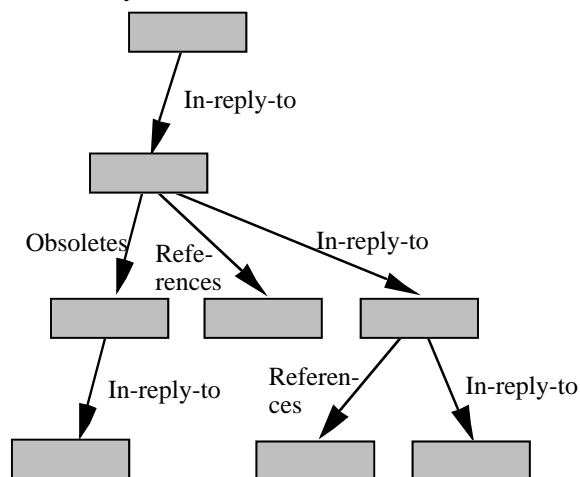
- *Automatic filtering* is where the computer evaluates what is of value for you.
- *Social filtering* (also known as *collaborative filtering*) is tools where other people help you evaluate what is of most value to read. Just like the publishers and organizations did in society before the Internet.

The most successful social filtering system is Yahoo. Yahoo employs humans to evaluate documents, and puts documents, which are interesting into its structured information database. This is very similar to what the publishers, editors,

journalists and organizations did in the world before the Internet.

The simplest and most common filtering is by organizing discussions into groups (newsgroups, mailing lists, forums, etc.) Each group has a topic, and wants only contributions within that topic. Sometimes the right to submit contributions is restricted. A common variant is that only members can submit, and sometimes competence control is done before accepting a new member. Another variant is that special moderators must approve contributions before distribution. The act when a recipient selects which groups to subscribe to, can thus be seen as an act of setting a personal filter.

Another simple and common filtering method is to filter by thread.



A thread is a set of messages, which directly or indirectly refer to each other. People can use threads for filtering by specifying that they want to skip reading of existing and future contributions in certain threads. In Usenet News, this functionality is known under the term “kill buffer”.

Automatic filtering has been successful only with very simple filters. Advanced methods for “*intelligent*” filtering have in general not been very successful. Intelligent filtering is a complex task requiring intelligence which computers are maybe not yet capable of?

FILTERING ISSUES

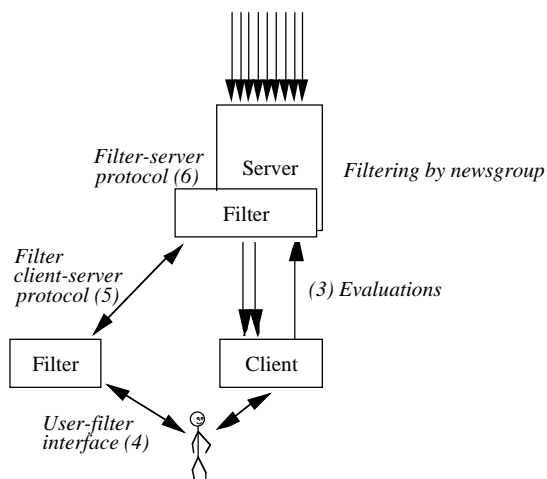
Filtering rules and attributes

Filtering is done by applying filtering rules to attributes of the documents to be filtered. Filtering rules are often Boolean conditions. They are usually put in an ordered list, which is scanned for each item to be filtered. The order of the items in the list can sometimes influence the outcome of the

Filtering of threads

In discussion groups, messages often belong to threads (see above). It may then not be possible to understand a single message without seeing other messages in the same thread. A filter or search facility which only selects certain individual messages, out of threads, might then not satisfy their users. The filter must either select several items in the thread, or at least make it very easy for users, when reading one selected message, to traverse the tree up and down from this message. Filtering in client or server

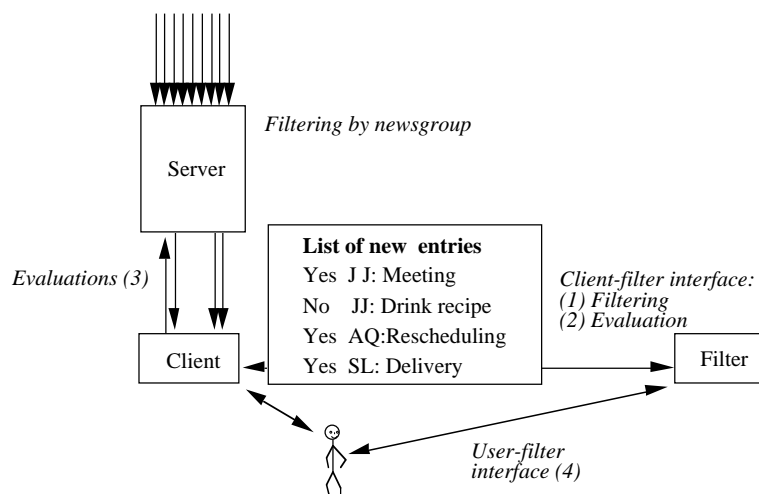
Filtering in clients or in servers



Filtering can be done in servers or in clients.

This figure above shows how a server can filter messages before downloading them to the client. The advantage with this is that filtering can be done in the background, and that messages filtered away need never be downloaded to the client. The disadvantage is that communication between user and filtering system becomes more complex. IETF is currently working on the development of a standard for the user control of server based filtering in a working group on MTA filtering [see IMC 1997].

Alternatively, filters may be part of the client, and apply to sets of documents after they have been downloaded to the client, as shown by the figure below:













Delivery of filtering results

The most common way of delivery of filtering results is that documents are filtered into different folders. Users choose to read new items one folder at a time. Thus, the filter helps users read messages on the same topic at the same time. The user can also have a personal priority on the order of reading news in different folders.

Unwanted messages can be filtered to special “trashcan” folders. User may choose not to read them at all, or to read such folders only very cursorily.

Filtering can also be used to mark messages within a folder. Different colors or priority indications can be put on the messages, or the messages may be sorted, with the most interesting first in the list.

Most services deliver new documents with a list, from which the user can select which items to read or not to read. The user act of selecting what to read from such a list can also be seen as a kind of filtering. The figure below shows an example of such a list, taken from the Web4Groups system [Palme 1997]:

-  [Technical Issues & Mirror Premiere!](#) , by [Roland Alton-Scheidl](#) 15/09/97 14:46 (1)
-  [Re: Technical Issues & Mirror Premiere!](#) , by [Roland Alton-Scheidl](#) 15/09/97 15:26
-  [Re: Technical Issues & Mirror Premiere!](#) , by [Torgny Tholerus](#) 15/09/97 15:59
-  [Integration of Advanced Functionalities](#) , by [Roland Alton-Scheidl](#) 15/09/97 16:58 (3)
-  [time_time_time](#) , by [Andras Micsik](#) 16/09/97 09:36 (2)
-  [Re: Integration of Advanced Functionalities.](#) , by [Alain Karsenty](#) 16/09/97 10:28 (1)
-  [Re: Web4Groups Technical Forum](#) , by [Torgny Tholerus](#) 16/09/97 10:41 (1)
-  [Re: Web4Groups Technical Forum](#) , by [Torgny Tholerus](#) 16/09/97 10:42
-  [Re: Web4Groups Technical Forum](#) , by [Torgny Tholerus](#) 16/09/97 10:42 (1)
-  [Re: Re: Web4Groups Technical Forum](#) , by [Mail Delivery Subsystem](#) 16/09/97 11:38

Intelligent filtering

By intelligent filtering is meant use of artificial intelligence (AI) methods to enhance filtering. This can be done in different ways: AI software can be used to derive attributes for documents, which are then used for filtering, it can be used to derive filtering rules, or it can be used for the filtering process itself. With the machine learning approach, the filter will take as input information from the user about which documents the user likes, and will then look at these messages and try to derive common characteristics of them to be used in future filtering.

Such filtering can be done in the background, behind the scenes, with little or no interaction with the user, or it can be done in a way where a user can interact with the filter and help the filter understand why the user likes certain messages. A disadvantage with much user interaction is that it takes user time, and the whole idea of filtering is to save user time. A disadvantage with very automatic

filtering is that the user may not trust a filter if the user does not understand how it works.

If an AI method is used to derive filtering rules, it might be valuable if these rules are specified in a way which a human can understand and trust. Certain AI methods, the so-called genetic algorithms, are known to produce very unintelligible rules and this may be a reason against using them for information filtering.

Filtering against spamming

Many people want filters which will remove unsolicited direct marketing e-mail messages, so-called spamming. To do this, the filter has to recognize special properties of spam messages, which distinguish them from other messages. Examples of such properties are:

1. A message does not have your name or e-mail address in the message heading, but it does not come from any mailing list, which you subscribe to. Many, but not all, such messages are spams. I personally let my filter mark all

such messages with a blue color, so that I can easily check whether to read them or not.

2. The author or sender of a message has an illegal e-mail address. Many MTAs (mail servers) now stop such messages, and because of this, the spammers have started to use legal e-mail addresses as senders. This is a general problem: If a particular filtering method gets very much used, spammers will change their messages to avoid being filtered.
3. Certain words, such as “money” or “\$\$\$” in the subject. This is not very dependable. It has the same problem as all intelligent filtering, see above.
4. If you often get similar spams, you might be able to recognize special properties of them to use to stop further similar spams.
5. The same message, with identical content, was sent to very many users, or to several newsgroups or mailing lists, at the same time. This method is commonly used for stopping spams in mailing lists and in Usenet News, and it seems to work, but spammers are beginning to learn to circumvent this, too.

None of these methods are very efficient. A social filtering system might be more efficient, see the next chapter.

SOCIAL FILTERING

What is social filtering

By social filtering is meant that some kind of ratings are assigned to documents. The ratings can be compared to the stars (☆☆☆) which newspapers often assign to films, books and other consumer products. But the ratings can also include categorization into subject areas or according to particular scales. Social filtering has some similarities to the filtering done by editors, journalists and publishers, since in both cases humans select the filtering attributes.

Why use social filtering

It is difficult to design automatic or intelligent filtering algorithms which really can evaluate the content of a document and evaluate its value. Humans are more capable of really deciding the value of a document.

Who make the ratings?

Ratings for use in social filtering can be provided by:

Editors, special people with the task of doing such rating. An example is the people selecting which messages to put into services like Yahoo [Yahoo 1998].

Readers, ordinary readers might input ratings on what they read, and these ratings might be collected and put into databases to help other people. Firefly [Firely 1996] and Grouplens [Resnick et al 1994A, 1994B] are systems based on this method.

Authors can provide certain kinds of ratings themselves. The advantage is that authors may be more willing to produce ratings, a disadvantage may be that an author might give too high ratings to his/her own documents. Because of this, author ratings are mostly useful if objective scales are used.

A filter may use an average or median of the ratings put by all who have rated a document. It might be better to use something like the upper quartile, since documents liked very much by a few people may be of particular interest, because they provide new thoughts and ideas. A filter might also base its filtering on the ratings done by other people with similar values, views and knowledge as the filter user. The filtering system might automatically find such people with similar views to the filter user.

Rating collection

A rating system must collect ratings from the people who do the rating. This can be done explicitly, where the user gives a rating command after reading a message. It can also be done implicitly, by studying variables like the time a user has spent on a message, whether the user has written a reply to it, printed it on paper, etc. Some studies indicate that such implicit rating can give as good values as explicit ratings. The advantage, of course, is that people may forget to provide explicit ratings.

Ratings collected in this way can be used for social filtering. But they can also be used as input to intelligent filtering algorithms (see above). And this might be a way of getting people to provide ratings, since people will have a personal gain by

providing ratings: This will make the intelligent filtering for themselves work better.

Spamming of social filtering systems

By spamming is meant ways in which people can cheat the system to force messages on you which you do not want. Most people think of spamming as it is done in e-mail or in Usenet News. But another variant of spamming is performed against Internet search engines. Authors of web documents give faulty keywords to their documents, to cheat the search engine into selecting the document by inserting the most popular search terms, which are known to be words like "sex", "naked", "girl", etc., even if these words are not related to the actual content of the document. Some search engines will first show you documents which contain the search word many times, so spammers may repeat the same word many times in the keyword set. (Keywords are placed in the meta fields of a HTML documents, which is not shown when you read the document with a web browser.)

Search engine providers have developed methods to recognize and dismiss messages with such false keywords. If social filtering systems are used in the future, there is an obvious risk that spammers will try to cheat the system, by entering lots of false positive ratings of their web pages. To stop this, some kind of authentication of raters may be needed.

Privacy issues

If a social filtering data base stores information, for individual raters, of which documents they like and dislike, such storage may be used for infringement of privacy. Possibly, some encryption method might be used to make such invasion impossible or difficult. This will of course depend on trust between user and filtering service. Web search engines today have similar privacy issues: They can store information about what you search for on the web. They already use this information to target selection of banner advertisements – other uses, which you might not like, may also occur.

RESEARCH ON FILTERING

How research on filtering is usually done

There are many research projects on information filtering. Such a project is usually started by some clever computer scientist, who has some novel idea of how to do filtering. He or she often finds that the task of developing a complete filtering system is larger than expected. If there was a standardized architecture for filtering systems, with standardized interface between modules, a researcher might easier be able to reuse existing modules, so that not a whole new filtering system has to be developed, when the researcher only wants to try out some new idea for one particular module.

Evaluation of filtering results

To evaluate a new filtering method, or to compare different filtering methods, one might compare the filtering with manual ratings of documents done by users. A filter which will be good at predicting the ratings done by a user would then be regarded as a good filter. Of course, an intelligent filter should not derive its filtering rules from one set of messages, and then test the filter on the same set. In the most extreme case, if a user found message 1, 3, 17, 32, 36, 53, 55, 58, 72, 76 and 84 best, a genetic algorithm might derive the rule: Select all messages with number 1, 3, 17, 32, 36, 53, 55, 58, 72, 76 or 84. Such a filtering rule would of course be totally valueless. Even if filtering is developed and tested on different sets of messages, there is still a risk that a filtering method is developed which only suits the test subjects. To avoid this, a large and varied set of test subjects should be used.

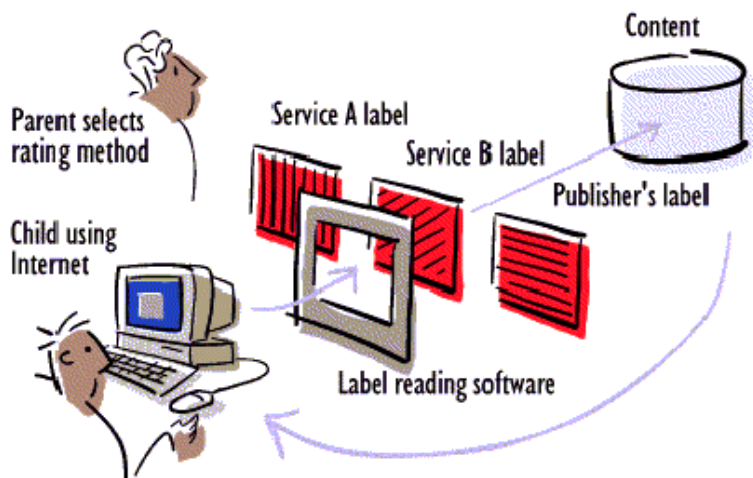
ARCHITECTURE AND STANDARDS

Architectural issues

To reduce the burden of developing and testing different filtering rules, it would be very valuable to develop a standardized architecture and standardized interfaces between the modules. The SELECT EU project [Palme 1998], which will start in the autumn of 1998, will work on this. Some modules which this project will specify are:

- Storage of author ratings
- Storage of personal and social filtering ratings
- User control of filtering rules
- Format and storage of filtering rules
- Filtering agent
- Attribute creators

The PICS standard



Picture from Resnick 1996A.

The PICS standard [Resnick 1996A, Resnick 1996B, Krauskopf 1996] was mainly developed as a tool for teachers and parents to censor the information which children can download from the Internet. But PICS can be useful in other ways. It provided a general-purpose, standardized way of storing and distributing ratings. Users or groups of users of PICS can, within the PICS standard, specify their own rating scales. PICS might thus be useful as a basis for some of the interfaces between the different modules of the filtering infrastructure.

The MTA filtering proposals

Another on-going standards work in the filtering area is the IETF work on MTA filtering [IMC 1997]. IETF is developing a basic standard for controlling server-based filters.

MORE INFORMATION

Overview of research and services

10.1.1 Different approaches

The issue of finding better-quality information on the Internet (in web documents, newsgroup postings, mailing list contributions, etc. below the word "document" is used) has been discussed and tackled in many different ways. A good collection of links to these issues can be found in [Ciolek 1994-1997]. Approaches taken have been:

- *Automatic tools* for finding and correcting technical faults in documents, such as non-working links in WWW pages, were proposed in IETF work in 1994 and are now a part of many web server maintenance tools. Their usage is sporadic and can therefore not assure a general improvement of quality.
- Making newsgroups and mailing lists *pre-moderated*, with a moderator who must accept all contributions before they are sent out, can be an efficient tool in increasing quality. This method has however the disadvantage that interaction is delayed, and that the group depends on the moderator. In practice, it has been found that to ensure continuous flow, there has to be a group of several moderators so that one can replace another who is on travel or ill.
- Another similar method is possible for mailing lists and in most computer conferencing systems but not in Usenet News: *Closed groups* where only selected people are allowed to participate. This requires someone to vet applications for membership and in general closed groups often die out because of too few members and lack of activity.
- *Education* of document authors and maintainers on quality issues is a never-ending work which will surely improve the quality at some places. A related method is to establish rules, procedures or ethical guidelines for documents and try to get them generally accepted. Such

work is surely valuable, but if Internet is to stay a medium where anyone can put up anything they want, no full solution to the quality problem.

- There is a large and rapidly increasing set of journals on the Internet, where contributions are selected in similar ways as in ordinary journals, for example scientific journals with *peer review* processes.
- Much work in different places has been spent on developing so-called *subject trees* or *subject structures*, i.e. maintained and well-organized databases of links to high-quality documents. Most well known is the Yahoo service [Yahoo 1998]. Some Internet search services have started to provide quality evaluations or reviews (Magellan from McKinley [Magellan 1997], Excite, OCLC's NetFirst, SBIG's [see Koch 1996A]), and the DESIRE telematics project [Koch 1996B] has as one of its major goals to develop quality assured collections for different subject areas. Another example is The Argus Clearinghouse (which started at the University of Michigan but is now a commercial company) which provides labels on Internet subject structures with descriptions and manually set quality ratings, in many ways similar to the quality labels specified the Centre for Information Quality Management. Such databases are developed and maintained by time-consuming human work, which limits their size and scope. The largest, Yahoo has for example less than a hundred thousand documents compared to tens of millions of documents in the largest Internet search servers. They are also not suitable for transient information, such as mailing list, computer conferencing and netnews contributions.
- One problem with the Internet is that documents come and go, and even valuable documents disappear. To solve this, some libraries have started scanning the net and *archiving copies* of documents available on the Internet for future retrieval. Another method is the work in IETF of developing URIs (*Uniform Resource Identifiers*), which are meant to be document references which do not have to change as rapidly as the currently used URLs (Uniform Resource Locators). Special URI servers are

meant to translate a URI to a URL when a document is to be retrieved.

- The PICS (Platform for Internet Content Selection) [Resnick 1996A, Resnick 1996B, Krauskopf 1996] of the World Wide Web Consortium has developed a standard protocol for content labels (labels with information about the quality of information resources), how to embed them in other Internet protocols and how to run label bureaus (service organizations providing labels). The primary incentive for PICS was the protection of children from unsuitable information, but the PICS protocols can be used to convey many kinds of quality labels, and SELECT may decide to use the PICS protocol for some of its modules.
- The Centre for Information Quality Management set up by the UK Online User Group of the Library Association has worked on specifying a format for *quality labeling* of databases. Quality labeling is a format for a producer of a database to specify the characteristics of his database in unbiased ways, similar to consumer product standards for consumer information labels.

10.1.2 Existing rating and filtering services and research projects

Many research projects are going on or finished in the area of information filtering.

- Patrick van Bommel at the University of Nijmegen maintains good overview pages of ongoing research at [Bommel 1997].
- Sepia Technologies, Inc in Quebec, Canada, has developed a collaborative filtering system for movies, music and books, see [Sepia 1995].
- Surflogic LLC in San Francisco has developed Surfbot, a web browser plug in which will search for and filter information on the net according to a users needs.
- The Department of Computer and Systems Sciences at KTH and SU has just finished a research project INTFILTER on intelligent filters. The result of this project can be found in [Kilander 1997]. A new EU project SELECT will start in the autumn of 1998 [Palme 1998].
- The most well known application of social filtering, Firefly [Firefly 1997], a commercial company which keeps a database of ratings of

movies, music and other information. A user can connect, input his favorite movie or music, and be told which other movies and music were rated highly by people with similar tastes as the user.

- The MIT Media Laboratory has a project on filtering agents led by professor Pattie Maes. They are also studying social filtering.
- The MIT Center for Coordination Science has developed GroupLens, a social filtering system for Usenet News [Resnick et al 1994A, 1994B].

References

Bommel 1997: *Internet filtering references* at <http://www.cs.kun.nl/is/research/filter>

Denning 1982: *Electronic Junk*, Communications of the ACM no. 23 vol. 3, March 1982, pp 163-165.

Firefly 1996: Personalize your Network at <http://www.firefly.com/>.

Hiltz and Turoff 1985: *Structuring Computer-mediated Communication Systems to avoid Information Overload*, by S.R. Hiltz and M. Turoff. Communications of the ACM, Vol. 28 No 7 July 1985, pp 680-689.

IMC 1997: *IETF MTA-filters Mailing List*. <http://www.imc.org/ietf-mta-filters/>.

Karlgren, 1994, Jussi: *Text genre recognition using discriminant analysis*. International Conference on Computational Linguistics, 1994. <http://www.sics.se/~jussi/cmplglixcol.ps>.

Kilander 1997, Fredrik, Fåhræus, Eva and Palme, Jacob: *Intelligent Information Filtering*, http://www.dsv.su.se/~fk/if_Doc/juni96/ifrpt.ps.Z.

Kilander, 1995, Fredrik: *A Brief Comparison of News Filtering Software*. http://www.dsv.su.se/~fk/if_Doc/Comparison.ps.Z.

Koch 1996A *Internet Search Services*, by Traugott Koch at the Lund University Library, in German at <http://www.ub2.lu.se/tk/demos/DO9603-manus.html>, in English at <http://www.ub2.lu.se/tk/demos/DO9603-meng.html>.

Koch 1996B *DESIRE: Development of a European Service for Information on Research and Education*. <http://www.ub2.lu.se/desire/>

Krauskopf 1996: *Platform for Internet Content Selection Version 1.1: PICS Label Distribution - Label Syntax and Communication*

Protocols, By T. Krauskopf, J. Miller, P. Resnick and W. Treese. URL

<http://www.w3.org/pub/WWW/PICS/labels.html>
Magellan 1997 *Magellan Internet Guide* at <http://www.mcinley.com/>

Malone 1987 et al: *Intelligent Information-sharing systems*, by Malone, Grant, Turbak, Brobst and Cohen. Communications of the ACM, May 1987, Vol. 30, No. 5, pp 390-402.

Palme 1981: *Experience with the use of the COM computerized conferencing system*, DSV, Stockholm University, 1981, re-published 1993, Available via gopher from dsv.su.se.

Palme 1984: *You have 134 Unread Mail! Do You Want To Read Them Now?* by Jacob Palme. In Proceedings of the IFIP Wg 6.5 Working Conference on Computer-Based Message Services, 1984.

Palme 1994, Jacob, Karlgren, Jussi and Pargman, Daniel: *Issues When Designing Filters in Messaging Systems*. Computer Communications 19 (1996) 95-101.

http://www.dsv.su.se/~fk/if_Doc/JPfilter-filer/IssuesDesFilter.ps.Z.

Palme 1997A: *Filtering and Collaborative Filtering*, Notes from the DELOS Workshop, Budapest, November 1997. <http://www.dsv.su.se/~jpalme/select/delos-filtering-notes-nov97.htm>.

Palme 1997B: *Non-Simultaneous Web-based Groupware* at <http://www.dsv.su.se/~jpalme/w4g/web4groups-summary.html>.

Palme 1998: *Choices in the implementation of rating* at <http://www.dsv.su.se/~jpalme/select/rating-choices.html>.

Pargman, 1994, David et al: *How to Create a Human Information Flow*. http://www.dsv.su.se/~fk/if_Doc/JPfilter-filer/HumaneInfoFlow.ps.Z.

Resnick 1996A: *PICS: Internet Access Controls without Censorship*, by P. Resnick and J. Miller, Communications of the ACM, and <http://www.w3.org/pub/WWW/PICS/iacwcv2.htm>

Resnick 1996B: *Platform for Internet Content Selection Version 1.1: Rating Services and Rating Systems (and their Machine Readable Descriptions)*, by J. Miller, P. Resnick and D.

Singer. May 1996. URL

<http://www.w3.org/pub/WWW/PICS/services.html>

Resnick et al 1994A: *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*. Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, Pages 175-186, and at URL

<http://ccs.mit.edu/CCSWP165.html>.

Resnick et al 1994B: *Roles for Electronic Brokers*, by Paul Resnick, Richard Zeckhauser and Chris Avery, Twenty-Second Annual Telecommunications Policy Research Conference,

October 1994, URL

<http://ccs.mit.edu/CCSWP179.HTML>.

Sepia 1995: *Collaborative Filtering – The SEPIA Suggestion Box*,

http://www.sepia.com/suggestion_e.html.

Tzolas 1994, I and Hussain, F.P: *Word-statistical categorization of texts for filtering of electronic messages*

http://www.dsv.su.se/~fk/if_Doc/JPfilter-filer/OrdStatUppsats.ps.Z.

Yahoo 1998: *Yahoo*. <http://www.yahoo.com/>.