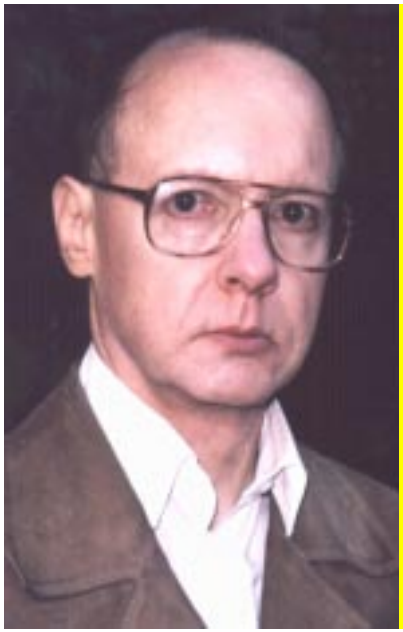
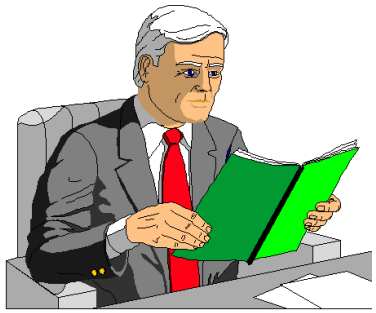
Filtering and Rating (=collaborative filtering or social filtering)

http://cmc.dsv.su.se/select/



by:
Jacob Palme
e-mail:
jpalme@dsv.su.se
web: http://www.dsv.su.se/~jpalme

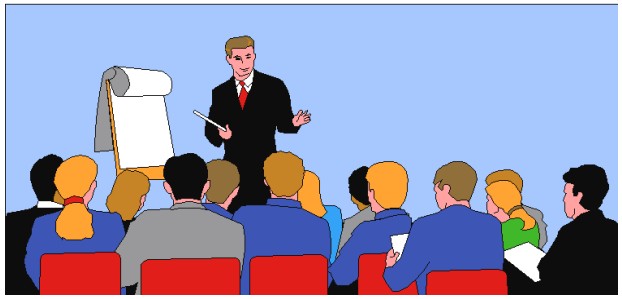
Who controlled distribution of information before the Internet?



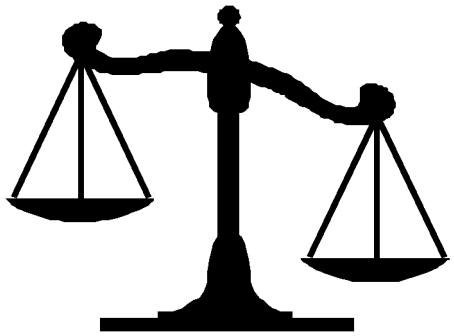
Publishers

News-papers

Schools and universities



Societies



Govern-ment, law

Was there no free speech then?

Sure, there was free speech. But free speech was controlled, channelled and organised.

In whose interest was it controlled, channelled and organised?

- Politicians
- Establishment
- Scientific community
- Readers

What is different with the Internet?

Anyone can easily at low cost publish anything they want.

Is this good or bad?

Both!

But everyone cannot read everything. The systems for control of free speech which we had before the Internet were in many ways tools which aided people in selecting the most valuable information.

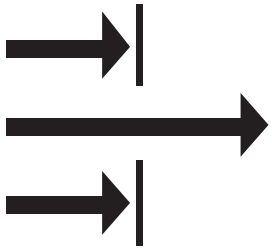
Newspapers and magazines selected the most interesting news. You chose to subscribe to the paper which selected according to your tastes.

The same with television, books.

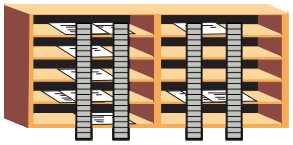
The same with societies: You chose which society to join, and in that way selected what information you would get and could yourself disseminate.

The quality of the information on the Internet is very varying. 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.)

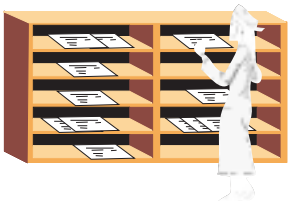
What is filtering?



Filtering is tools to help you find the most valuable information, so that the limited time you have for reading can be spent reading interesting information and avoiding trash.



Automatic filtering is where the computer evaluates what is of value for you.



Social or collaborative filtering is tools where other people help you evaluate what is of most value to read. Just like the publishers and organisations 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 data base. Just like the publishers and organisations did in the world before the Internet.

Is filtering successful?

Automatic filtering is successful only with very simple filters.

Examples of partially successful filters:

- Filtering by mailing list/newsgroup.
- Filtering by topic (thread).

But filters which automatically in a more intelligent way finds what is of most interest to you have not been very successful.

Why?

Maybe because filtering is a complex task requiring intelligence which computers are not yet capable of?

The most successful social filtering service, Yahoo, employs people to select the best, just like publishers did before the Internet.

Another important social filtering service is mailing lists, where only members can contribute (or at least should be able to) and various tools are used to keep out disruptive people.

How can social filtering work on the Internet?

Several different ways:

- People are employed to make the selection (like Yahoo).
- More or less closed groups for people with common interests.
- Peer groups helping each other find the best.
- Data bases where anyone can put their ratings:
- You get the highest rated documents by all raters.
- You get the highest rated documents by people with similar values as yourself.

How is research on filtering usually done?

A clever computer scientist develops his wonderful new ideas of how filters should work. His/her results very seldom result in products used by other people.

Is there another way of organising research?

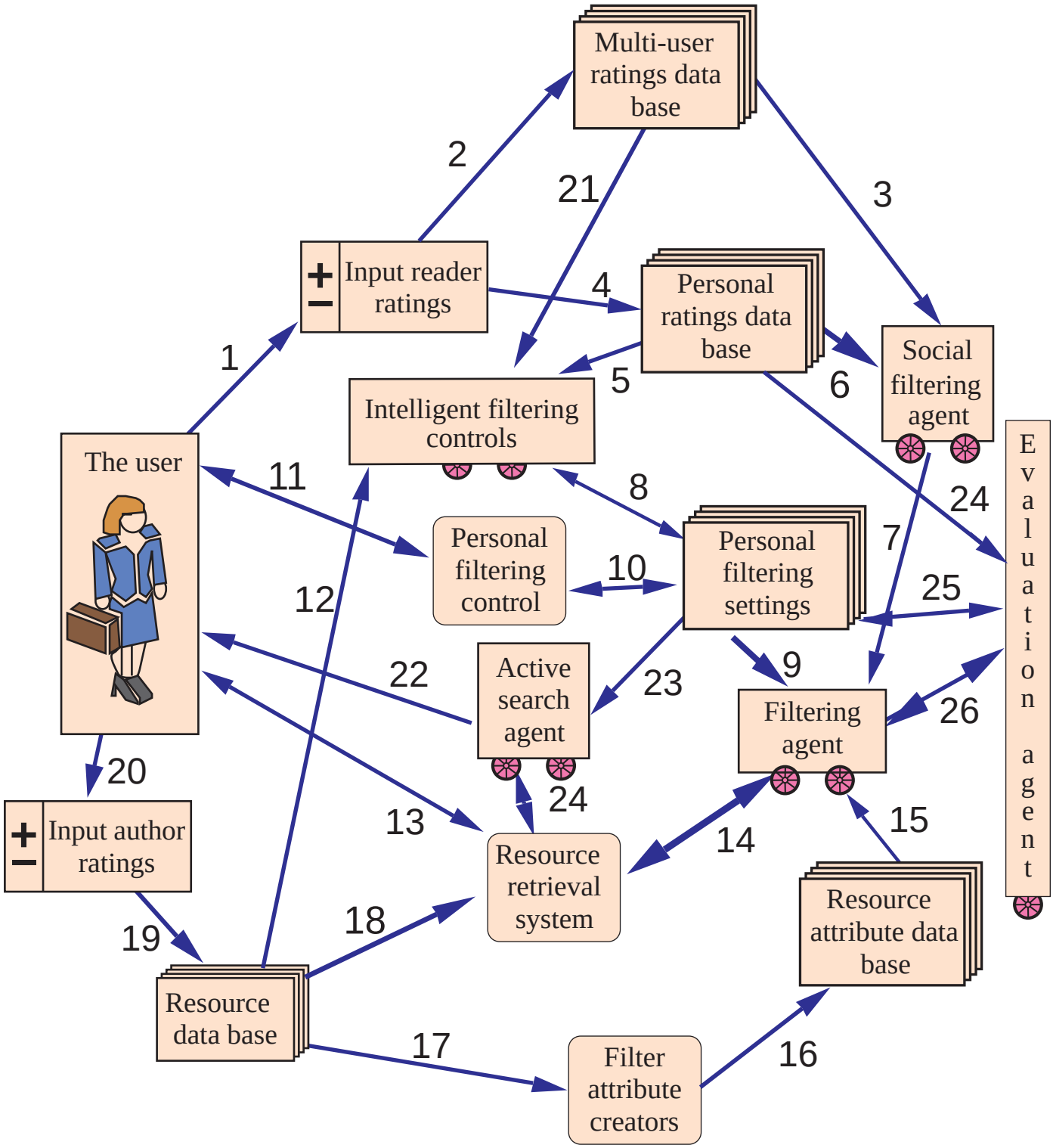
Yes: Specify a general-purpose architecture, so that many different developers can develop and test their filters.

An architecture is a defined set of modules and interfaces, where there can be different versions of each module, which can still co-work with the other modules.

Why should we specify such an architecture?

1. Developers need not develop the whole filtering system, with all its components, which usually is very time-consuming. They can develop only the module where they have bright ideas, and these modules can be tested in the general-purpose architecture.
2. Different filtering methods and modules can be tested and evaluated against each other within the general-purpose architecture.

Proposed architecture for multi-user filtering and rating system for SELECT project, November 1997 by Jacob Palme.



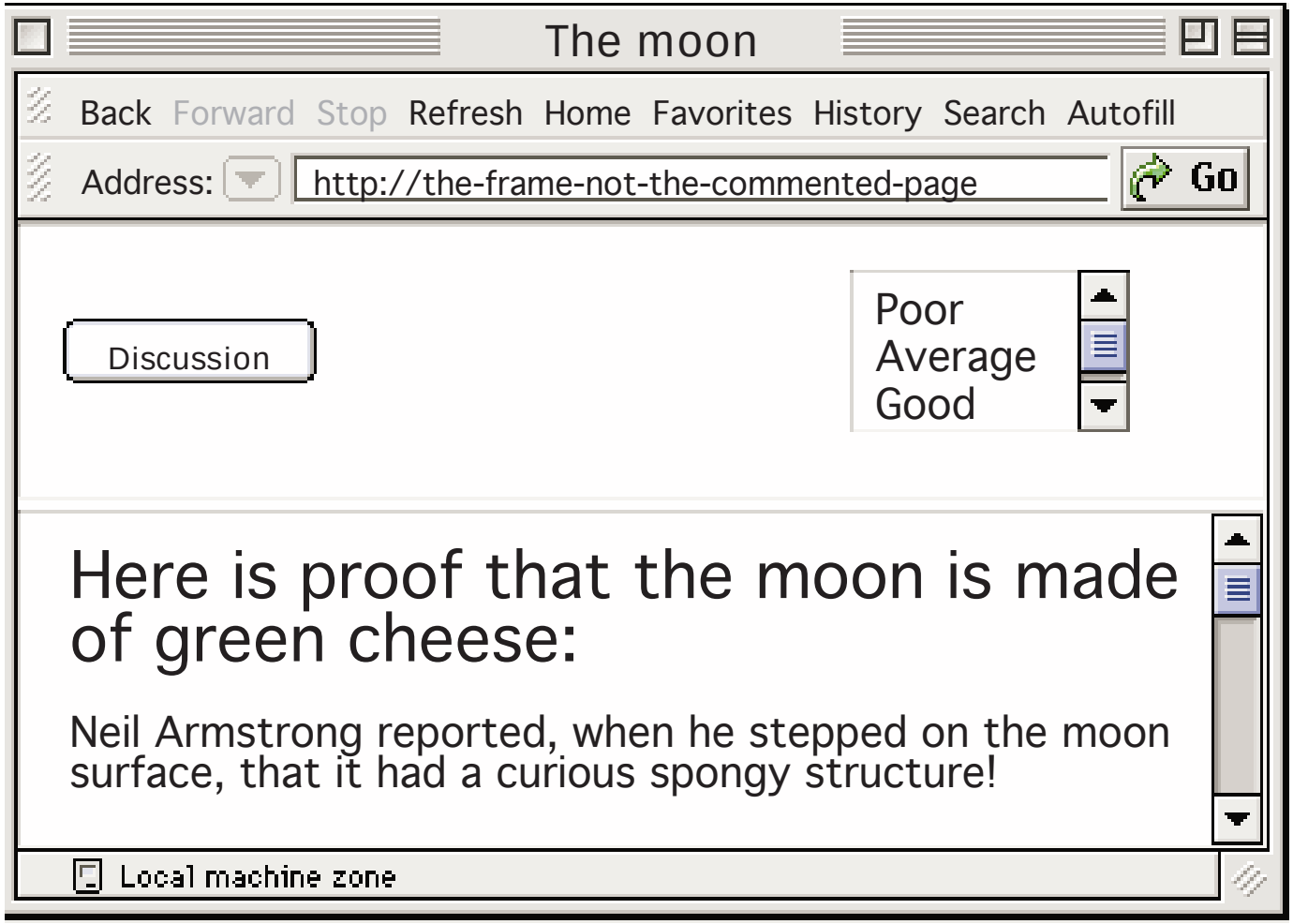
Legend: Data bases; Ratings; Agents; Processes

SELECT project primary tasks:

- Collect a set of rated documents to use for experiments.
- Implement automatic filtering attribute creators.
- Implement a social filtering system for web pages and Usenet News.

Select issues:

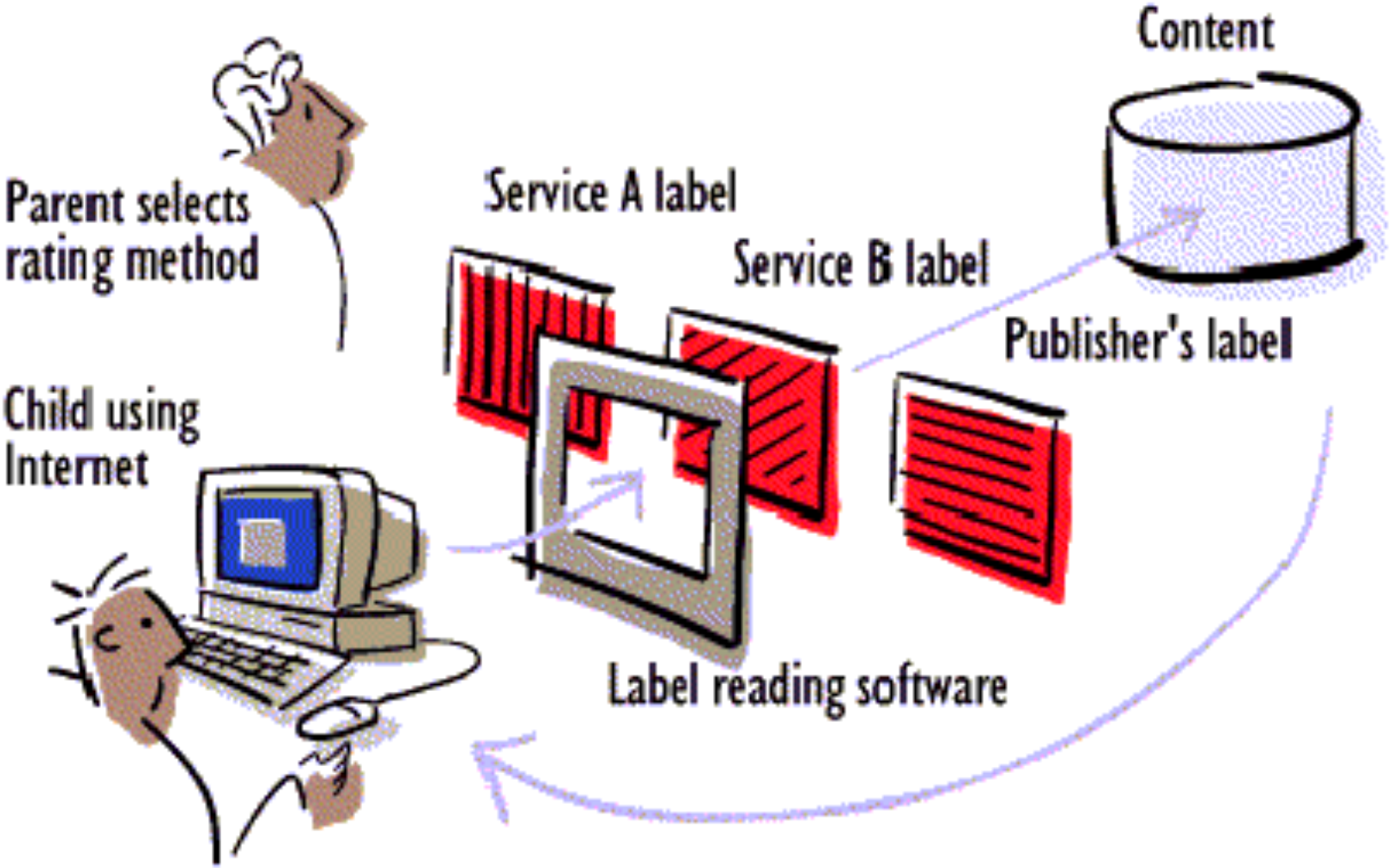
- How to specify individual interest profiles.
- Which rating scales to use.
- Adjustment to needs of different user communities.
- Rating for everyone or for specialist groups.
- Whose ratings to use: Those from anyone, from automatically derived peer groups, from manually selected peer groups, from experts or appointed reviewers.
- Questionnaire at URL:
  http://www.sztaki.hu/servlets/voting/SELECT



Use of frames, inserted by a proxy server, to provide ratings or find talkback forum.

Mockups of ratings specification page:

The PICS (Platform for Internet Content Selection) standard can be used for more than parental censoring of what their children can download



Use of existing standards: Resource delivery to users

Delivery of documents to users is normally done using HTTP and FTP (normal Web lookup), SMTP, POP and IMAP (e-mail delivery) and NNTP (news server-server and client-server protocol). It is an advantage if we use the same protocols as in other Internet usage as much as possible, since this makes it easier for other people to extend their software (news clients, mail clients, Web browsers) to use our filtering services.

If we want to use only one protocol:

IMAP	Powerful protocol, can be used to retrieve mail, news and Web pages, has better news control than HTTP, but not widely adopted.
HTTP	Users can use ordinary Web browsers, for example using an extension of the Web4Groups Web gateway as delivery tool.

Use of existing standards: Other uses

HTTP	Can be used for many needs as the protocol to rapidly get and put small units of information. Can be modified, for example change “HTTP/1.0” in the first line to “SELECT/1.0” for our own protocol, and use only a limited set of all facilities HTTP can provide.
PICS	Can be used to specify rating systems and permitted values. Can also be used as a format to convey ratings. *(More info about PICS on the next slide)*

Use of existing standards: PICS

PICS	Can be used to specify rating systems and permitted values. Can also be used as a format to convey ratings. For example: PICS already specifies formats to put PICS labels in the <HEAD> of HTML documents and in the Heading of RFC822 e-mail messages. PICS also specifies formats to get PICS labels from PICS servers, so-called label bureaux. The current PICS user interfaces are oriented toward forbidding the user to see forbidden texts, which is not so suitable for general social filtering needs. But PICS is still a good choice if we modify this part of PICS to suit our needs.

Spamming problem

Spamming	Forcing or cheating people into getting information they do not want.
Mail spamming	Sending unsolicited ads in e-mail, usually with falsified sender information, often misusing mailing lists.
Search engine spamming	Giving a Web document incorrect search key information to cheat search engines into showing this Web document before others.
Social filtering spamming	Cheating social filtering systems into believing that your documents have been highly rated by users.

This is a difficult problem. Possible solutions:

1. Require raters to get an account, do not allow them to use the account until the password or account name has been sent to them by e-mail.
2. Only allow use of ratings for members of special societies.
3. Employ people who are paid to do the rating.
4. Require all raters to get a cryptographic certificate and use this when submitting ratings, but this might make it too difficult to input ratings.

More to read in the full paper at:

http://www.dsv.su.se/~jpalme/select/information-filtering.html

- Relations between filters and other net agents like search engines, groupware, e-mail software
- Architecture of filter in relation to other net agents
- Protocols to communicate to and from filters
- Delivery of filtering results
- Intelligent filtering
- Administration of social filtering
- Overview of research on filtering