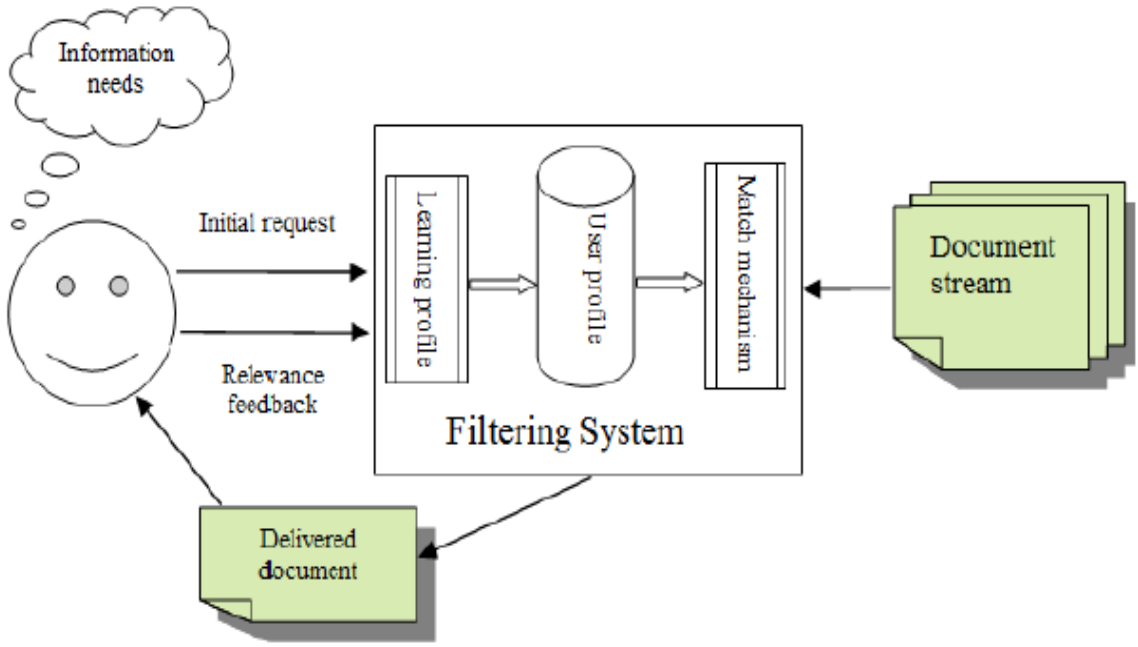
Document Filtering

Document filtering is an application that stores a large number of queries or user profiles and compares these profiles to every incoming document on a feed. Documents that are sufficiently similar to the profile are forwarded to that person via email or some other mechanism.



Major differences compared to a search engine:

1. Generally, filtering systems calculate similarity between the profile and each incoming document, and retrieve documents with similarity higher than a threshold. However, many systems set a relatively high threshold to reduce retrieval of non-relevant documents, which results in ignorance of many relevant documents. On the other side, search engines do a pretty good job is retrieving relevant documents.

2. The first major difference between a search engine and document filtering is that, in document filtering, once the user

profile is created, the documents are fed to the user, maybe via email. An example of this can be RSS feeds. In case of Search Engines, the user needs to enter the query and relevant documents are presented to the user.

3. Document filtering uses the publish-subscribe scheme, where the users publish their profile and then subscribe to receive relevant documents based on their profiles. The user unsubscribes by giving negative feedback about the document it had received which could be because the user is no longer interested in receiving documents of that type.

4. Another important difference is that, search engine uses ranking algorithms like page rank, to rank the documents and document filtering relies on user's feedback to rank the documents.

5. Search engines use index to store the path of the documents and provides a list of documents based on user query, whereas document filtering store profiles and feeds documents to the users.

6. Since search engine use index, the response time is faster than that of document filtering.

Usefulness of ranking in a filtering application:

1. Ranking can be achieved based on the feedback received from

the users. When a user receives a particular document based on his profile, he can either give a positive or negative feedback. Documents that get negative feedback, are less relevant to the user and the user profile can be updated to reflect that he is not interested in similar documents when the rank of such documents reaches below a particular threshold.

Reference:

1. http://dl.acm.org/citation.cfm?id=345573