Intelligent Filtering; Based on Keywords Only?

Ann Lantz
Dpt of Computer and Systems Sciences
Stockholm University
Electrum 230, S-164 40 Kista
Sweden
alz@dsv.su.se

Fredrik Kilander
Dpt of Computer and Systems Sciences
Stockholm University
Electrum 230, S-164 40 Kista
Sweden
fk@dsv.su.se

ABSTRACT
Empirical studies on users of Usenet News indicate that features other than keywords also plays an important role in the user's decision to read or reject a message. The IntFilter project aims to develop a message classification mechanism which use such extra features while minimizing the burden for the user who has to provide and maintain long-term queries.

KEYWORDS: Text classification, information filtering, computer-mediated communication, machine learning.

INTRODUCTION
Keywords and phrases are in common use in information filtering ([4, 5, 10]), but other criteria could be usable, especially in combination with machine learning techniques. By studying how users of Usenet News select and rate information, we have searched for additional filtering criteria. The term *filtering* here include the concept of selection, as well as rejection.

USER STUDIES
In a first pilot study [2] the aim was to investigate how experienced users of Usenet News selected their information and how they reacted towards assisting the system in different ways i. e. by rating messages, write keywords, use rules or by letting the system take part of the users profile. Four subjects were instructed to perform a set of tasks in Usenet News and think aloud during the session. After the session the subjects were interviewed.

The results pointed at the fact that the users selected their information by taking notice of the information in the message header; mainly keywords they saw in the subject field, and partly by browsing through the text, where keywords and general layout were used for selection. In an unfamiliar newsgroup the selection took more time, apparently since the keywords were unfamiliar. The subjects were positive to rating messages, adding keywords and also to use rules. On the question if the system should be allowed to take part of the users profile, one subject expressed the feeling of need for privacy i. e. not wanting the system to know all of the groups he was entering.

In a second pilot study [3] the aim was to investigate the criteria for selecting information used by experienced readers of Usenet News. Three subjects were instructed to select ten messages from five newsgroups each and to think aloud while they read. They were asked to read the messages and place each in one of three categories (read now, read later, discard).

The subjects used depth and amount of cited text, keywords, patterns in the message, questions, the length of the text, alphabetical signs or numbers to form their decisions. These criteria have been divided into two groups; pattern recognition and demands reading.

A questionnaire was distributed in a Swedish Usenet News newsgroup. The aim with the study was to answer the following questions:

- Do users of Usenet News experience a need for filtering?
- Which criteria do the users adopt to find or avoid information in Usenet News?
- In what way are the users willing to support the system?
- How important do the users regard the information presented in Usenet News?

Twenty subjects responded to the study and the preliminary results showed that:

- The need for filtering was experienced by 19 of the subjects and they did not express a need for filtering information in the newsgroups not subscribed to.
- The contents of the subject field in the message header was the base for what the subjects thought the messages were about, but other factors (e. g. genre) were also considered.
- To avoid information the subjects again referred to the subject field but the length of the message was also important.
- The subjects were mainly willing to support the system by writing rules but also by writing keywords; to allow the system to take part of their user profile and to rate a message as interesting or not.
- To take part of the information in the news system was rated as important or very important by 17 of the subjects.

TECHNICAL WORK
In order to facilitate automatic recognition of messages that the user is likely to regard as interesting,

some kind of feedback model is required. The feedback model enables the user to provide the system with information about preferred messages. One of the most uncomfortable ways to do this, is to force the user to formulate such preferences in terms of a rule system with rigid and narrow semantics. Popular readers for Usenet News commonly employs such rules. With them, the burden of creating operational and non-conflicting rules is put on the human.

Schemes involving machine learning has been proposed where the user supplies either an explicit or implicit rank on each message. The rankings are then treated as classification labels and inductive processes are employed to distill classification rules. Explicit ranking of messages, however, distracts the user from the primary task of reading. The burden is increased by inductive machine learning techniques which commonly require many correctly and consistently ranked examples and counter-examples. Implicit ranking does not put any extra strain on the user, but must make a number of ad hoc assumptions with respect to the behaviour of each user. Such assumptions include the expected reading time per lines of text, that cited text is not read in detail, that the user is actually reading and not answering the telephone and that the user does not reject relevant messages by accident.

In response to the empirical studies by Lantz mentioned above, we are of the opinion that the user is willing and able to supply an automatic classification system with a small number (five or less) of *prototypical* messages for each conceptually distinct and interesting topic. We also believe that the user is prepared (although not required) to review the installed prototypes in order to delete text, add important fragments or emphasize critical parts of the existing text.

Our approach to the information filtering problem is to view it as an information retrieval process, where the user has a number of long-term queries. Each query corresponds to a topic the user is interested in. The document collection at any time, is the stream of unread messages provided by the subscribed newsgroups. This view has earlier been presented by Belkin and Croft [1].

Our work is similar to that of Sheth [9], Morita and Shinoda [6], and Yan and Garcia-Molina [11]. Like them, we have decided to use a simple statistical similarity measure for queries and documents. We have evaluated two such measures, the well-known *cosine* measure (Salton [8]), and *sub-string indexing* (Morita and Shinoda [6]).

However, we also want to take advantage of the structural information contained in each message. We believe that a message filter (in contrast to an *information* filter) should be able to gain discriminative power from the physical appearence of a message. In quantitative terms: the amount of cited text, the layout of the text, the proportion of misspelled words and the presence of common features such as punctuation and capitalization.

AN INTEGRATED SYSTEM

In a previous phase, we developed a filtering system called GHOSTS which was intimately coupled to the newsreading client NN. GHOSTS used a rule production system, boolean matching and required the user to maintain an explicit set of rules. GHOSTS has since been abandoned in favor of a collection of smaller and more independent software modules for classification and learning. We plan to integrate these modules with the computer conference system KOM [7]. The KOM system will not only offer users a combination of message sources beside Usenet News, but also provide the necessary interface, and database functions.

REFERENCES

1. N. J. Belkin and B. W. Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, December 1992.
2. A. Lantz. How do experienced users of the system usenet news select their information? Technical report, Dpt of Computer and Systems Sciences, Stockholm University, 1993.
3. A. Lantz. Useful criterias for intelligent filtering? Technical report, Dpt of Computer and Systems Sciences, Stockholm University, 1994.
4. T. W. Malone, K. R. Grant, F.A. Turbak, S.A. Brobst, and Cohen D. Intelligent information.sharing systems. *Communications of the ACM*, 30(5):390–402, 1987.
5. T.W. Malone, K-Y. Lai, and C. Fry. Experiments with oval: A radically tailorable tool for cooperative work. In *CSCW Proceedings*, pages 289–297, 1992.
6. M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 272–281. Springer-Verlag, 1994.
7. J. Palme and T. Tholerus. Superkom — design considerations for a distributed, highly structured computer conferencing systems. *Computer Communications*, 15(8):509–518, October 1992.
8. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
9. B. Sheth. A learning approach to personalized information filtering. Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 1994.
10. B. Sheth and P. Maes. Evolving agents for personalized information filtering. In *Proceedings of the ninth IEEE Conference on Artificial Intelligence for applications*, 1993.
11. T. W. Yan and H. Garcia-Molina. Index structures for information filtering under the vector space model. Technical report, Dpt. of Computer Science, Stanford University, Stanford CA 94305, 1993.