Personalized Information Delivery: An Analysis of Information Filtering Methods[*]

Peter W. Foltz
Dept. of Psychology
Univ. of Colorado
Boulder, CO 80309

Susan T. Dumais
Belllcore
445 South st.
Morristown, NJ 07962-1910

Abstract

This research tested methods for predicting which Technical Memos (TMs) best match people's technical interests. Within Bellcore, approximately 150 new TMs are published each month, yet very few are relevant to any single person's interests. In a six month study, 34 Bellcore employees were sent monthly personalized lists of new TM abstracts that were predicted to best match their interests. These predictions were made using two methods for describing technical interests; one based on sets of keywords that the employees provided, and the other using feedback about previous abstracts they found relevant. Two information retrieval methods were tested to make the predictions, one using standard keyword matching, and the other using Latent Semantic Indexing (LSI). All four methods effectively selected relevant abstracts. The best method for filtering used LSI with feedback about previous relevant abstracts. Feedback using previous relevant abstracts provided an efficient and simple way of modeling people's interests. Overall, the filtering methods show promise for presenting personalized information.

Introduction

With the increasing availability of information in electronic form, it becomes more important and feasible to have automatic methods to filter information. Research organizations generate large amounts of information. This information can include departmental and technical memoranda, announcements of meetings and conferences, and minutes from meetings. With this large amount of information, it is difficult to keep employees apprised of all relevant work. Furthermore, only a small fraction of the available information will actually be relevant to any particular employee within an organization that covers a variety of areas. Thus, there is the problem of determining what information is of interest to the employee while minimizing the amount of search through irrelevant information. This research tested several information retrieval methods for filtering Technical Memos.

Filtering information is not a new concept, nor is it one that is limited to electronic documents. When we read standard paper texts, information filtering occurs. We only buy certain magazines, since other magazines may contain information that is redundant with or irrelevant to our interests. In this way, we are filtering out some of the large amount of information to which we have access. Within any particular magazine, we also choose articles that appear relevant to our interests. Thus, information filtering is something that is continually done by people when engaged in any sort of acquisition of information. With the advent of electronic presentation of information, some of that filtering need no longer be done by us, but could be done automatically by the system that presents the information.

While automatic filtering of information sounds like a wonderful vision, there are many difficulties in determining what information a person wants to see. The description of what information is of interest is often referred to as a "user profile". Broadly defined, this can also be thought of as a rudimentary kind of user model. (For a description of much more complex user models in information retrieval, see Belkin, Brooks, & Daniels [5].) There are many problems in developing a good model of a user's interests. For example, a variety of factors could be used to describe a person's interests. Most commonly, people provide a set of words to describe their interests. However, many other sources of information could be used, such as which articles they have read in the past, what organization they work

in, which books they have ordered, etc. While the general topic or content of an article may be important in predicting whether it will match a person's interests, other factors such as familiarity, novelty, importance or urgency may also be useful in predicting what information a person might want to see. In addition there may be interactions between certain factors and specific applications. A factor that provides a good description of interests for world news, may not be effective for describing a person's research interests.

Even with a clear idea of what factors are important for predicting interest, there is no guarantee that those factors can be identified easily. One of the simplest methods of determining whether information matches a user's interests is through keyword matching. If a user's interests are described by certain words, then information containing those words should be relevant. This straightforward keyword matching often fails however. Inappropriate matches can arise because the words people use do not unambiguously reflect the topic or content. A single word can have more than one meaning (e.g., chip), and, conversely, the same concept can be described by surprisingly many different words (e.g., human factors, ergonomics). Furnas, Landauer, Gomez & Dumais [9] showed that two people use the same main word to describe an object only 10 to 20 percent of the time. Bates [2] has reported comparably poor agreement in the generation of search terms by trained intermediaries. (See also Blair & Maron [4].)

Information Filtering Systems and Methods

One of the earliest forms of electronic information filtering came from work on Selective Dissemination of Information (SDI) [11]. SDI was designed as an automatic way of keeping scientists informed of new documents published in their areas of specialization. The scientist could create and modify a user profile of keywords that described his or her interests. SDI then used the profile to match the keywords against new articles in order to predict which new articles would be most relevant to the scientist's interests. While SDI was implemented on a large scale, it was used far less than predicted [14].

Several more recent studies and systems have been developed to test information filtering ideas. Allen [1] conducted a series of experiments to explore user models in predicting preferences for news articles. In one experiment, he predicted which articles a person would read based on previous articles read using a measure of overlap of nouns between the new and old articles. While the predictions were better than chance, the average correlation between the predicted articles and the subjects' ratings of the articles was fairly low (r=0.44). The models were more successful at predicting user preferences for general categories of articles than for specific articles. Predicting what news articles a person will read may be an especially difficult task. News topics vary from day to day, making it difficult to get stable estimates of interest. In addition, external sources of news probably influenced what people read in the experiment. We believe that users' interests for technical literature will be more stable over time.

In Allen's research, the subject's past preferences were used to construct an implicit model for retrieving relevant articles. A somewhat different approach is to let the user explicitly structure the information. The Information Lens system [12, 13] allows users to create rules to filter mail messages based on keyword matches in the mail fields. Since there is already some structure in mail messages, such as sender information in a sender field and keywords in a keyword field, these rules can take advantage of this structure to perform user specified actions on the messages. Thus, a rule may take the form of deleting all messages from a certain person, or labeling messages with certain keywords as urgent. Mackay et al. [12] found that people without much computer experience were able to create their own Information Lens rules to prioritize and filter their mail messages. The largest percentage of rules were created to match on information about the sender and other recipients, while fewer rules were created to match on textual information such as the subject and text body. Although people could create filters, the research does not report on the effectiveness of the filtering methods.

While a variety of information systems have been developed, there has been little systematic evaluation of what features are most effective for filtering. This leaves many unanswered questions, such as: what are the most effective methods for matching a user's interests to information available, how should a user's interests be described, and how will the performance of filtering methods vary in different domains. This research explores the first two of these issues using information retrieval methods as the basis for filtering information.

Information Retrieval

Conventional information retrieval (IR) is very closely related to information filtering (IF) in that they both have the goal of retrieving information relevant to what a user wants, while minimizing the amount of irrelevant information

retrieved [16, 19]. Belkin and Croft [3] identify three primary differences between IR and information filtering. First, user preferences (profiles) in information filtering typically represent long term interests, while queries in IR tend to represent a short term interest that can be satisfied by performing the retrieval. Second, information filtering is typically applied to steams of incoming data, while in IR, changes to the database do not occur often and retrieval is not limited to only the new items in the database. Finally, a distinction can be made between the two, in that filtering involves the process of "removing" information from a stream, while IR involves the process of "finding" information in that stream.

In both information retrieval and information filtering, a textual database can be represented by a word-by-document matrix whose entries represent the frequency of occurrence of a word in a document. Thus, documents can be thought of as vectors in a multi-dimensional space, the dimensions of which are the words used to represent the texts. In a standard "keyword matching" vector system [16], the similarity between two documents is computed as the inner-product or cosine of the corresponding two columns of the word-by-document matrix. Queries can also be represented as vectors of words and thus compared against all document columns with the best matches being returned. An important assumption in this vector space model is that the words (i.e. dimensions of the space) are orthogonal or independent. While it has been a reasonable first approximation, the assumption that words are pairwise independent is not realistic. Several statistical and AI techniques have recently been used to better capture term associations and domain semantics. One such method is Latent Semantic Indexing (LSI). Only a brief overview of the LSI method will be presented here. Mathematical details and examples can be found in Deerwester et al. [6] and Furnas et al. [10].

Latent Semantic Indexing

Latent Semantic Indexing (LSI) is an extension of the standard vector retrieval method which was designed to help overcome some of the retrieval problems described above. In LSI the associations among terms and documents are calculated and exploited in retrieval. The assumption is that there is some underlying or "latent" structure in the pattern of word usage across documents, and that statistical techniques can be used to estimate this latent structure. A description of terms, documents and user queries based on the underlying latent semantic structure (rather than surface level word choice) is used for representing and retrieving information.

The particular Latent Semantic Indexing (LSI) analysis described by Deerwester et al.[6] uses singular-value decomposition (SVD), a technique closely related to eigenvector decomposition and factor analysis. SVD takes a large word-by-document matrix and decomposes it into a set of k, typically 100 to 300, orthogonal factors from which the original matrix can be approximated by linear combination. Instead of representing documents and queries directly as vectors of independent words, LSI represents them as continuous values on each of the k orthogonal indexing dimensions derived from the SVD analysis. Since the number of factors or dimensions is much smaller than the number of unique terms, words will not be independent. For example, if two terms are used in similar contexts (documents), they will have similar vectors in the reduced-dimensional LSI representation. One advantage of this approach is that queries can retrieve documents even if they have no words in common. The LSI technique captures deeper associative structure than the simple term-term correlations and clusters, and is completely automatic.

One can interpret the analysis performed by SVD geometrically. The result of the SVD is a k-dimensional vector space containing a vector for each term and each document. The location of term vectors reflects the correlations in their usage across documents. Similarly, the location of document vectors reflects correlations in term usage. In this space the cosine or dot product between vectors corresponds to their estimated similarity. Retrieval proceeds by using the terms in a query to identify a vector in the space, and all documents are then ranked by their similarity to the query vector. The LSI method has been applied to several standard IR collections with favorable results. LSI has equaled or outperformed standard vector methods and other variants in every case, with improvement of as much as 30%. As with the standard vector method, differential term weighting and relevance feedback can improve LSI performance substantially [7].

Filtering using IR Techniques

In both LSI and keyword vector matching, documents are represented as vectors in a high dimensional space. In keyword vectors, the values on each dimension are determined by which words occur in a document. In LSI vectors, the values are based on a smaller number of statistically derived indexing dimensions. Documents on similar topics tend to be near each other because they share words (in keyword matching) or indexing values (in LSI). This feature is used as the basis for filtering.

In general, the idea for filtering is to create a space of documents, some of which have previously been judged by a user to be relevant to his or her interests. If a new document is close to relevant documents in the space, then it would be considered likely to be interesting to the user. Conversely, if that document is far from relevant documents, then it would be considered not interesting to the user. This same procedure can be used to determine how close any new document is to keywords in the user's profile. For all these comparisons, the only difference between the LSI and the keyword matching methods is that LSI represents terms and documents in a reduced dimensional space of derived indexing dimensions.

Foltz [8] compared LSI and keyword vector matching for filtering of Netnews articles. In an experiment, subjects rated Netnews articles as either relevant or not relevant to their interests. The ratings from the initial 80% of the articles they read were used to predict the relevance of the remaining 20% of the articles for each person. Foltz found that the LSI filtering improved prediction performance over the keyword matching method by an average of 13% and showed a 26% improvement in precision over presenting the articles in the order received as is typically done with Netnews articles.

The goal of the present experiment was to evaluate methods for filtering Technical Memos (TMs). The evaluation compared different methods of matching users' profiles to documents and different ways of profiling the users' interests.

Filtering Experiment

The Domain

Bellcore publishes an average of 150 Technical Memos (TMs) each month. These TMs cover a wide variety of topics ranging from solid-state physics to switching systems requirements. Few TMs in any particular month will be relevant to the interests of a given reader. Because research on similar topics occurs within different groups and in different geographic locations, it is not always possible to keep employees apprised of all of the relevant research. Currently, employees are sent a list of all the TM abstracts published each month. This list is loosely organized by major work group and by the date the TM was received. Thus, an employee must typically glance through all abstracts in order to determine which ones are relevant.

This experiment used information retrieval techniques to provide employees with personalized lists of TM abstracts. Previous filtering research has often applied filtering methods to news sources. However, people's interest in news and current events may vary from day to day based on the events that are occurring. The TM abstracts provide a nice test domain for filtering research because people's research interests are relatively stable over time and there is a steady flow of new TMs each month.

Method

Thirty four Bellcore employees participated in the experiment. The employees had a fairly wide range of interests and positions within Bellcore and were spread across several geographic locations. The employees initially provided a list of words and phrases that described their technical interests. The average list was 24 words long with a range of 6 to 66 words. Over a 6 month period[1], the employees were sent monthly lists of about 20 TM abstracts. For each abstract, they rated its relevance to their interests on a 7 point scale, with 1 being not at all relevant and 7 being very relevant. They also indicated whether they had seen the TM before and whether they would like to order the TM. Each month, the employees were also given the opportunity to update their keyword list by adding or removing words.

The lists of TM abstracts that were sent out each month were based on predicted users interests. Four methods were used to filter new TMs. The four methods were the result of crossing two factors, the first factor being whether the retrieval method used LSI or Keyword matching, and the second factor being whether the profile was based on words and phrases provided by the employee (Word profile), or abstracts that the employee had previously rated as relevant (Document profile). The methods are shown in Table 1.

	Matching Method	
Type of profile	Keyword	LSI
Words	Keyword match- Word profile	LSI match- Word profile

Documents	Keyword match- Document profile	LSI match- Document profile

Table 1. The four filtering methods

The *Keyword match-Word profile* method compared words in the employees' profile to words in the abstracts of new TMs. This method was equivalent to the vector retrieval method [16]. The *LSI match-Word profile* method also compared the words provided by the employee to words in the abstracts but used the reduced-dimensional LSI vector space for the comparison. After the first month, TM abstracts that employees had previously rated as highly relevant to their interests were also used to select new TMs using the two matching methods. In the *Keyword match-Document profile* method, previously rated relevant abstracts were compared to the abstracts of new TMs using the standard vector method. In the *LSI match-Document profile* method, the same comparison was done, except using the reduced-dimensional LSI space. For both document profile methods, the full text of the previous relevant TM abstracts was used for the comparisons[2]. This document profile method is a variant of what is often referred to as "relevance feedback" in the IR literature. Relevance feedback has been shown to provide large performance improvements in simulations [17].

LSI retrieval requires the initial analysis of a corpus of text in order to extract useful statistical relationships between terms and documents. For this study, 6535 Bellcore TMs written between 1984 and 1989 were analyzed. Each TM abstract was automatically indexed. All words occurring in more than one TM abstract and not on a stop list of 439 common words were included. No word stemming or lexical analyses were performed. 16637 words occurred in more than one of the TM abstracts, resulting in a 16637 word by 6535 TM matrix. The frequencies in the word-by-document matrix were then transformed using tf*entropy term weights. The resulting matrix was analyzed using SVD. A 100-dimensional vector space was then used for LSI retrieval. The standard vector retrieval method does not require the analysis of a pre-existing corpus to compare word or document profiles with new incoming TMs. However, since it is known that appropriate term weighting can improve vector retrieval, we used the same tf*entropy term weights for these comparisons. Thus, the same terms and term weights were used for both the standard vector and LSI vector methods. In all cases, the cosine between vectors was used as the measure of similarity.

Figure 1 shows the process used for filtering. New TM abstracts were matched against employees' word and document profiles using the two matching methods. User profiles were represented as several separate points of interest in the vector space. For the word profiles, employees indicated which words or phrases were separate points of interest by putting each on its own line (e.g. "user centered design" and "graphical user interface" are different interests). For the comparisons, each new TM was individually matched against each line in the word profile. The score for each new TM was the cosine between the TM vector and the nearest interest vector. The new TMs were then ranked based on their maximum cosine score. Thus, TMs occurring close to any point of interest for a particular employee would be ranked the highest. For the document profiles, the same type of comparison was done. Each document in the document profile was represented as a separate vector and compared to all new TMs. The new TMs were then ranked based on the maximum cosine for each TM to any document in the document profile.

These comparisons resulted in four rank ordered lists of abstracts, one for each filtering method. The top abstracts from each method were then sent to each employee, who rated them as to the relevance to his or her technical interests. For each method, the top 25% of the abstracts rated 4 or higher were then incorporated into their document profiles[3]. In the first month of the study, only the word profile was used since employees had not previously indicated which abstracts to include in their document profile. In the subsequent months, both the document profiles and the word profiles were used to match against new abstracts.
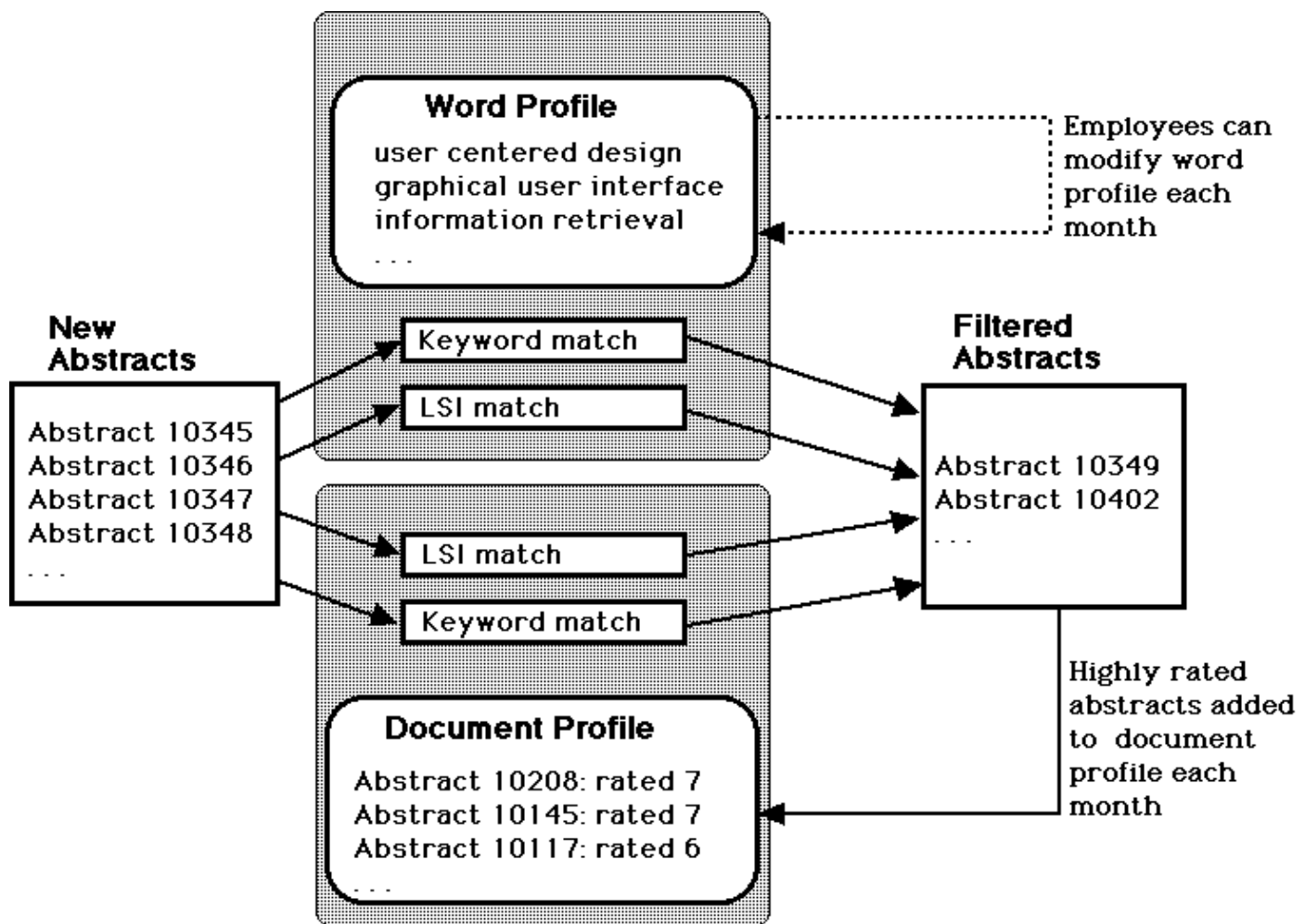
Word Profile

user centered design
graphical user interface
information retrieval

. . .

Employees can modify word profile each month

New Abstracts

Abstract 10345
Abstract 10346
Abstract 10347
Abstract 10348
. . .

Keyword match

LSI match

LSI match

Keyword match

Filtered Abstracts

Abstract 10349
Abstract 10402
. . .

Highly rated abstracts added to document profile each month

Document Profile

Abstract 10208: rated 7
Abstract 10145: rated 7
Abstract 10117: rated 6

. . .

Figure 1. The Filtering Process

For each month, we selected the top 7 abstracts matched by each of the methods. Thus, with the 4 methods, there could be a maximum of 28 abstracts. Twenty eight abstracts represent approximately 20% of the TMs published in any month. We expected fewer than 20% of all the TMs would be relevant to any persons interests. However, since the same abstract could be selected by more than one method, employees typically saw fewer than 28 abstracts. The mean number of abstracts sent by the methods was 17. In addition to the abstracts picked by the four filtering methods, 3 randomly selected abstracts were also included every month. These random abstracts allowed us to estimate how many relevant abstracts were missed and the overall relevance of abstracts for each month. Once the list of abstracts was created, they were printed out in random order and distributed by internal paper mail.

Three weeks after rating the filtered TMs in the last month of the study (July), the 34 employees were also asked to rate the complete set of TMs for the final month. Twenty nine of the employees returned these packets. These exhaustive relevance ratings permitted us to evaluate alternative filtering methods. In addition to the 34 employees who received the filtered TMs each month, two employees received packets containing all of the TMs published in each month. Like the other employees, they had to rate the relevance of all the TMs to their interests. These exhaustive ratings allow us to examine the full range of recall and precision, and to explore alternative filtering methods.

Results

The goals of the analyses were to determine how well the different filtering methods performed and to investigate alternative approaches to improving filtering. The primary data used were the employees' monthly ratings of the filtered TMs.

Comparison of Filtering Methods

The mean rating for the abstracts picked by each of the four filtering methods and by random selection was computed for each month. These results are shown in Figure 2. An analysis of variance was performed on the employees' mean ratings for each method and post-hoc analyses using t-tests were done in order to compare specific methods to each other. The overall ANOVA showed reliable differences between the methods ($F(4,132)=117.5$, $p<.01$). The most noticeable feature of Figure 2 is that the TMs picked by any of the four filtering methods are rated 1.5 to 2.5 points above those picked randomly. All four methods performed significantly better than random selection of TMs (LSI-Doc vs random $t(33)=13.7$, $p<.01$, LSI-Word vs. random $t(33)=12.2$, $p<.01$, Keyword-Doc vs. random $t(33)=12.9$, $p<.01$, Keyword-Word vs. random $t(33)=14.8$, $p<.01$). This indicates that all the filtering methods are succeeding in returning relevant TMs.
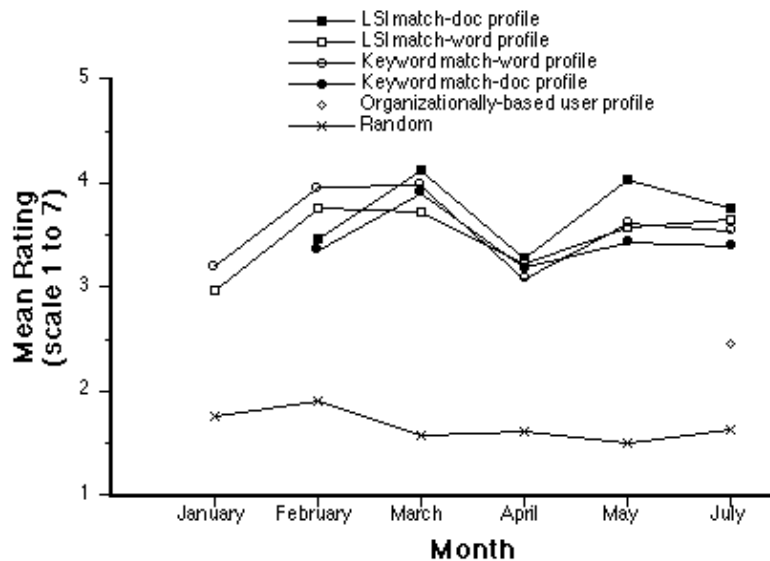

Figure 2. Mean ratings of TM abstracts over the 6 months

The average ratings of abstracts returned by the filtering methods are not very high considering the range of relevance ratings was from 1 for non-relevant TMs up to 7 for very relevant TMs. This is due in part to the fact that a fixed number of TMs were selected by each method regardless of the actual cosine similarity scores. By choosing fewer than the top 7 TMs from each method or only TMs above a cosine threshold, the filtering methods could have been more selective. For example, the mean rating increases from 3.5 for choosing the top 7 abstracts to 4.5 for choosing just the top abstract.

The differences between the four filtering methods are not large. However, after the first month of using the document profiles, the LSI matching using the document profile performs consistently better than the other three methods. Overall, the mean ratings for the four methods were: 3.74 (LSI match-Document profile), 3.57 (Keyword match-Word profile), 3.49 (LSI match-Word profile), and 3.46 (Keyword match-Document profile). Post-hoc analyses were performed on the mean ratings for each employee for each method for the months of March through July. The first two months were excluded from this analysis since there no document profiles in the first month, and there were very few documents contained in the document profiles in the second month. The analyses indicated that TMs selected by the LSI match-Document profile method were rated significantly higher than than those selected by the other three methods (LSI-Doc vs. Keyword-Word $t(33)=2.6$, $p<.02$, LSI-Doc vs. Keyword-Doc $t(33)=3.6$, $p<.01$, LSI-Doc vs. LSI-Word $t(33)=3.2$, $p<.01$).

The relative order of the other methods varied somewhat from month to month and there were no significant differences between the other three methods. The standard error of the means for the methods was fairly high, ranging from 0.13 to 0.16. This high variability is partially due to the fact that preferences between employees varied greatly.

For any given month, some employees found many relevant TMs, while others found very few to be relevant. There is also considerable variability from month to month in the mean ratings. Much of this is attributable to the fact that we always selected the top 7 abstracts from each method. Some months (e.g., April) produced fewer abstracts that were of interest to the employees in our study than others. Additional variability arises because the user profiles are constantly being updated. This effect is particularly salient in comparing document profiles in February and March.

In the first month the document profiles (February), the two methods using document profiles did not perform as well as the word profile methods. We believe this is because there were too few abstracts in the document profile. For this month, the document profiles contained an average of only 1.3 abstracts. By the second month of using the document profiles (March), an average of 4.1 documents were in the document profile and the two document profile methods performed at least as well as the word profile methods. The average number of words in a word profile was 25 for March. This suggests that just a few relevant documents can be as effective as a long list of keywords for describing one's interests, especially when coupled with LSI matching.

Alternative Filtering Methods

The fact that the four methods performed much better than randomly selected TMs is not surprising. An alternative is to compare the IR-based filtering methods with other methods of ranking the documents for filtering. One approach is to filter the documents based on hierarchical organizational distance. Using this method, documents that are written by employees in hierarchically nearby groups would be ranked higher than those written by employees in groups that are more distant. Intuitively, this method makes sense in that organizations are structured so that groups doing similar research tend to be grouped in the same part of the hierarchy. For each employee, a ranking of new TMs based on their distance from the employees' organization can be calculated. Using the full ratings provided by employees in July, the mean rating of the top 7 TMs ranked by organizational distance was computed. This is shown as the single point in July in Figure 2 (mean rating=2.45). Although, better than randomly selected TMs, this method does not perform as well as the IR-based methods. This indicates that employees' interests tend to span the hierarchical structure of the company and are not easily predicted by where they work within the organization.

A second approach is to compare the IR filtering methods with the current Bellcore distribution method. A paper list of abstracts loosely ordered by major work group and date is distributed to employees. All employees get all the abstracts in the same order, so this is not a filtering method, but it does provide a baseline comparison. Using the ordering provided by this method, the mean rating of the top 7 TMs (mean rating=1.66) was almost the same as the rating from the randomly selected TMs (mean rating=1.72). Thus, all filtering methods tested would be improvements over the current distribution method, and over personalized lists based on organizational distance.

Overlap of Methods

Each of the four filtering methods independently picked the most relevant abstracts for each employee based on their profiles. Often more than one method would select the same abstract for the same person. Over all the abstracts selected, 38% were selected for a particular employee by more than one method. Figure 3 shows the mean ratings for the abstracts based on the number of methods that selected an abstract as being relevant for a particular employee. As the number of methods selecting the same abstract increased, the mean relevance rating of the abstract increased. The mean relevance of abstracts selected by all four methods, for example, was 5.04. This finding is consistent with other IR studies. Both Saracevic & Kantor [15] and Turtle & Croft [18] found similar improvements in performance when using multiple representations of queries against a single database.
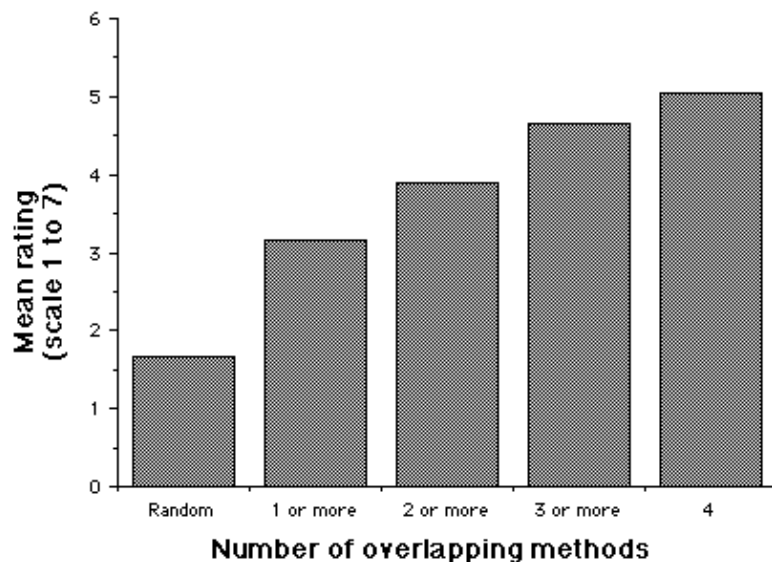
Figure 3. Mean rating of TMs based on the number of overlapping methods selecting an abstract

It is important to note, however, that as the number of overlapping methods increases, fewer documents are matched. 181 (about 5%) of the abstracts returned to users were matched by all four methods. The mean rating for these abstracts was 5.04, well above the means shown in Figure 2. A more reasonable comparison involves examining only the top 5% of the abstracts that would have been returned by each method independently, since this provides a comparable level of selectivity. The mean ratings for the top 5% for each method vary from 4.01 to 4.87 with an average of 4.58. Thus, selecting abstracts that match all 4 methods provides a better filter than simply returning a comparable percentage of abstracts from any single method. Abstracts selected by three or more methods show a similar but smaller advantage. For three overlapping methods (representing 9% of the selected abstracts), the mean rating was 4.65, while selecting the top 9% of the abstracts from individual methods gave a mean rating of 4.54. This advantage disappears when abstracts selected by two or more methods are considered. For two overlapping methods (representing 23% of the selected abstracts), the mean rating was 3.90, while selecting the top 23% of the abstracts gave a mean rating of 4.32. Using multiple methods to increase selectivity appears to be successful only in the most restrictive case (i. e. when 3 or more methods agree). In other cases, simply selecting fewer abstracts from any given method is more effective. Using overlap for selecting TMs has the additional drawback that interesting abstracts that can not be retrieved by some methods (e.g., because they do not share keywords with the word profile) would never be returned.

Alternative Measures of Relevance

The results discussed so far have been based on employees' monthly relevance ratings. Other measures of the relevance of the TMs can be obtained by examining the TMs that the employees ordered or indicated that they had previously seen. Both provide additional, although imperfect, measures of the relevance of TMs. It is expected that TMs that had been seen previously were distributed in some manner so that they would go to people who would find them relevant or useful. In addition, if an employee indicated that he/she wanted to order a TM, it would indicate that the TM was of interest. These measures are less systematic than relevance ratings, since we had no control over how TMs were being distributed outside of the experiment.

The percentage of TMs that were ordered varied from 17.8 to 22 percent, while only 4.5 percent of the randomly selected TMs were ordered. Consistent with the rating data, the LSI match-Document profile method selected the highest number of TMs ordered. The percentage of TMs seen previously ranged from 9.6 to 12.5 percent for the four methods, compared to 0.6 percent for the randomly selected TMs. The Keyword match-Word profile and the LSI match-Document profile methods performed best using this measure. Thus, these two additional measures provide evidence that is consistent with that of the rating data.

Results from Exhaustive Rating of all TMs

While the rating measures discussed above provide measures of the relevance of the TMs that were sent to the employees, they provide little information the ranks of relevant TMs or about the relevance of TMs that were not selected. Using employee's ratings of the randomly selected TMs, we estimated that the filtering methods failed to retrieve 50% of the relevant TMs[4]. More comprehensive performance data can be obtained using exhaustive relevance judgments. We have such data from two employees who rated the complete set of TMs over the six months, and from 29 employees who rated the complete set of TMs in July. One common measure of performance of an IR system is precision [16]. Precision is the percent of retrieved items that are relevant and recall is the percent of relevant items retrieved. This performance measure requires exhaustive relevance judgements and is sensitive to the relative ranks of relevant and non-relevant items.

Figure 4 shows the mean precision at recall levels of 0.25, 0.50, and 0.75 for the four methods for the employees' July ratings. These data are based on the 26 employees who rated at least one TM as relevant for that month. Consistent with the rating data, the LSI match-Document profile method has the highest overall precision, indicating it returns the highest percentage of relevant documents over the three levels of recall. Comparing the two methods that use the word profiles, the LSI matching method is better than the Keyword matching method. Since some of the TMs which an employee rates as relevant do not share <u>any</u> words with their word profile, the LSI match method, which does not depend on exact word matching, will tend to perform better. This sizable difference between LSI and Keyword matching is not seen in Figure 2 because the rating data are based only on TMs selected by each method. Another comparison of interest involves the different profiling methods for a given matching method. For both LSI matching and Keyword matching, using a document profile results in better performance than using a word profile. This advantage is attributable to the richer vocabulary found in the document profiles.
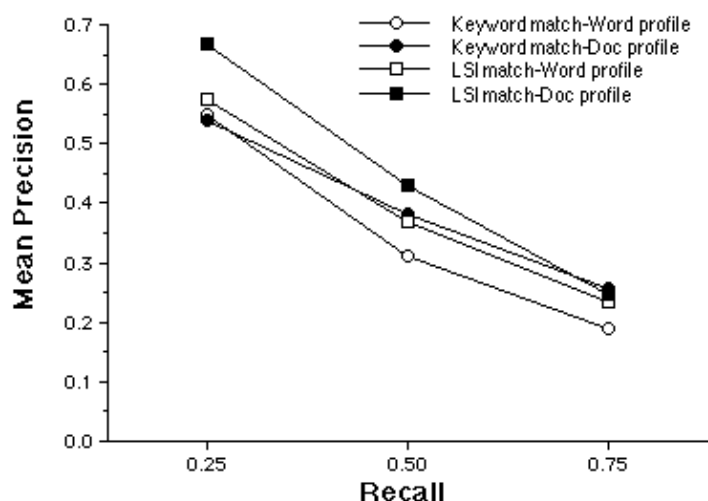

Figure 4. Mean precision for the four filtering methods at three levels of recall (using exhaustive ratings from July)

The complete ratings by the two employees of all TMs published over the 6 months permitted us to test alternative measures of performance. An analysis of the average precision for the four methods was essentially the same as that found for the 26 employees' July ratings. In addition, we explored the effectiveness of the filtering methods with different choices of values for how many TMs to be returned by each method, what rating value would be considered relevant, what percentage of the relevant TMs should be added to the document profile (see footnote 3). Simulations were run varying the values for these three features. Overall, there was very little change in the effectiveness of the filtering methods with different choices of values. This indicates that the initial choice of values made in the experiment did not interact with relative effectiveness of the filtering methods.

Discussion

The TM abstracts provide a nice test domain for information filtering research because people's technical interests are relatively stable over time and there is a steady flow of new TMs each month. Our research compared four IR methods for matching employees' interests to the TMs using two matching methods and two types of user profiles. Overall, the four methods succeed at filtering the TMs when compared to randomly selected TMs, the current distribution method, and an organizationally-based filtering method.

The LSI match-Document profile method proved to be the most successful of the four filter methods. This advantage was observed for all performance measures we examined - mean ratings, mean precision at three levels of recall, and the number of TMs seen or ordered. This method combines the advantages of both LSI and the document profile (which is a kind of relevance feedback). The LSI matching method allows users to retrieve documents that have no words in common with their initial profile. The document profile provides a simple, but effective representation of employees' interests. Indicating just a few documents that are of interest is as effective as generating a long list of words and phrases that describe one's interest. Document profiles have the added advantage over word profiles that users need not generate descriptions of what they like, but can just indicate documents they find relevant.

While the four filtering methods were effective individually, the ratings of TMs increased with the number of methods that matched a particular TM. Each filtering method has slightly different ways of representing and matching an employee's interests. When three or four methods agreed that a certain TM was relevant, the TM tended to receive a higher rating than a comparable number of TMs picked by a single method. So, for highly selective filtering applications, using multiple methods is advantageous.

Future Directions for Filtering Research

In addition to providing a comparison of information filtering methods this research suggests some future directions and issues for filtering research.

One issue is how to determine how many documents to return to a user. With the TMs, in which only about 150 are published per month, it is possible to provide the complete list rank ordered based on some filtering criteria. But, in other cases, this may not be as easy because so much information is available (e.g. news wires). For these cases, some cutoff should probably be used. Many current filtering systems based on keyword matching send just those documents that contain the desired keywords (or some Boolean combination of keywords). Vector methods give graded similarity measures, but most documents will have no similarity (overlap) with a profile. LSI also returns graded similarity measures, but now all documents match a query, just to a greater or lesser extent. We chose a fixed cutoff point (the top 7 documents for each method) for this research because it simplifies the comparison of methods. A cutoff based on the actual cosine similarity score is also possible in practice. In addition, the cutoff for people could vary greatly depending on such factors as: the type of information (e.g. news, related court cases), whether the user needs full coverage of the information, and cost of retrieving or storing the information.

There also remain alternative ways of representing users' interests. While the filtering methods examined in this study used profiles that described an employee's interests, these methods could also use information describing what an employee is not interested in. Providing negative information is a more difficult task, since there is a much larger set of descriptors of things that we are not interested in. However, using a document profile, employees could easily indicate TMs that are not relevant to their interests. In future months, any new TMs that are similar to these non-relevant TMs could also be considered non-relevant. A related issue is how to combine information from profiles. In this study, employees' interests were represented as several different vectors, one for each line in the word profile and each TM in the document profile. The similarity of a new TM was simply its cosine to the nearest interest vector. A more reliable measure of similarity might be obtained by accumulating information from the many points of interest resulting in a measure representing the sum of their interests.

One of the difficulties for users of a filtering system would be determining how well their profiles are actually performing. Certain descriptors in a profile may be effective at matching relevant documents, while others may match fewer relevant documents. One possible way to keep users apprised of the effectiveness of their profiles is to provide information on how and why any document matched against the profile. This would permit users to see what descriptors in the profile are performing effectively and allow them to make changes accordingly. These changes could also be made automatically by monitoring which descriptors match documents and how highly those documents are rated. Descriptors that result in documents that are highly rated could then received increased weight in the ranking of new sets of documents.

This study has examined several methods for filtering technical information. While there still remain additional issues that need to be addressed, the results of this study show promise for developing effective systems for personalized information delivery.

References

1. Allen, R. User models: Theory, method and practice. *International Journal of Man-Machine Studies*, *32*, (1990), 511-543.

2. Bates, M. J. Subject access in online catalogs: A design model. *Journal of the American Society for Information Science, 37*, (1986), 357-376.

3. Belkin, N. J., & Croft, W. B. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, (1992).

4. Blair, D. C., & Maron, M. E. An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM*, *28*, (1985), 289-299.

5. Belkin, N. J., Brooks, H. M., & Daniels, P. J. Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studie*s, *27*, (1987),127-144.

6. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41,*6, (1990), 391-407.

7. Dumais, S. T. Improving the retrieval of information from external sources. *Behavioral Research Methods, Instruments & Computers*, *23*, 2, (1991), 229-236.

8. Foltz, P. W. Using Latent Semantic Indexing for Information Filtering. In *Proceedings of the ACM Conference on Office Information Systems* (April 25-27, Boston, Ma.). ACM/SIGOIS, New York, 1990, pp. 40-47.

9. Furnas, G. W., Landauer T. K., Gomez, L. M. & Dumais, S. T. Statistical semantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal*, *62*, 6, (1983), 1753-1806.

10. Furnas G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th International Conference on Research and Development in IR*. ACM/SIGIR, New York, 1988, pp. 465-480.

11. Houseman, E. M. & Kaskela, D. E. State of the art of Selective Dissemination of Information. *IEEE Transactions of Engineering Writing and Speech III*, 2, (1970).

12. Mackay, W. E., Malone, T. W., Crowston, K. Rao, R., Rosenblitt, D., & Card, S. K. How do experienced information lens users use rules? *Proceedings of ACM CHI '89 Conference on Human Factors in Computing Systems* (April 30-May 4, Austin, Tx.). ACM/SIGCHI, 1989, pp. 211-216.

13. Malone, T. W., Grant, K. R., Lai, K. Y., Rao, R. & Rosenblitt, D. R. Semistructured messages are surprisingly useful for computer-supported coordination. *ACM Transactions on Office Information Systems*, *5,* 2, (1987), 115-131.

14. Packer, K. H. & Soergel, D. The importance of SDI for current awareness in fields with severe scatter of information. *Journal of the American Society for Information Science, 30*, 3, (1979), 125-135.

15. Saracevic, T. & Kantor, P. A study of information seeking and retrieving. III. Searchers, searches and overlap. *Journal of the American Society for Information Science, 39*, 3, (1988), 197-216.

16. Salton, G., & McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.

17. Salton, G., & Buckley, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*, 4, (1990), 288-297.

18. Turtle, H. & Croft, W. B. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, *9*, 3, (1991), 187-222.

19. vanRijsbergen, C. J. *Information Retrieval.* Butterworths, London, 1979.

Acknowledgements