Interactive information retrieval using term relationship networks.

Jim McDonald, William Ogden and Peter Foltz

New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003
{jem|ogden|pfoltz}@crl.nmsu.edu

ABSTRACT

Users have difficulty retrieving information from ad-hoc, textual databases because, by definition, they don't know precisely what's in them. Thus, users submit queries that contain the wrong terms or which don't contain enough information to retrieve all and only those documents relevant to their information needs. Our approach to these problems is to provide users with abstract representations of database content, in the form of Pathfinder networks linking related terms used in the database. This allows users to recognize and select appropriate query terms. The networks displayed to users are derived from textual analyses of documents retrieved from initial queries and, thus, the process can be thought of as a form of relevance feedback. Compared to other relevance feedback methods, however, the network displays can show important relationships between the query terms and terms suggested by the system. In the study to be reported, we compared the performance of two information retrieval systems Zprise, a control system, and ***InfoView***, a system that uses our network displays. Participants used both systems to perform an "aspectual retrieval" task using the six topics. Preliminary results from this study suggest that when participants used ***InfoView*** they took less time to identify topic aspects and were at least as successful as when they used Zprise.

Our approach to interactive information retrieval is based on a simple premise: users of information retrieval systems can't find the information they seek because they don't know exactly what's in the databases they're searching. This tautology can be expanded into a couple of more interesting claims. First, users aren't able to formulate effective queries because they don't know (or can't recall) appropriate query terms. Second, if users are provided with appropriate information about database content, they are able generate (or select) better query terms and subsequently improve retrieval performance. In other words, if users have a good understanding of database content, they do a better job of retrieving the information they seek.

One approach to improving information retrieval is relevance feedback, an inherently iterative technique (e.g., Rocchio, 1971). A user first formulates a query based on her information need and understanding of database content. She then submits the query to the information retrieval system and a list of documents (or document titles) is returned which the system deems relevant to the search request, usually ordered according to relevance. The user examines the document titles or reads the documents in order to determine which are relevant. In its simplest form, the user then modifies her original query by adding or removing terms, then resubmits the query to the information retrieval system. In this scenario query modification is based on an analysis of the retrieved documents and involves the inclusion of terms which are synonyms or appropriate modifiers in the context of the database being searched.

Some relevance feedback techniques attempt to incorporate relevance judgments into the retrieval process by comparing relevant to non-relevant documents automatically. Even these systems, however, require users to examine retrieved documents in some detail in order to determine which are relevant. Such examination can be a time-consuming and error-prone process. Judging relevance may require users to examine a large number of documents, some which may be fairly low on the ranked list returned by the information retrieval system. On the positive side, relevance feedback has been shown to be an extremely effective technique for improving information retrieval performance (Salton &. Buckley, 1990)

Our objective is to achieve the positive results typical of relevance feedback techniques while requiring less effort on the part of the user. Thus, we provide users with information about database content in a form which minimizes time and effort while maximizing content coverage. Our technique is based on the notion that abstract representations derived from statistical analyses of text (documents) can be used to improve queries (and retrieval performance) without requiring users to examine documents or even document titles. There are two components to an information retrieval system which incorporates this approach. First, routines must be provided to analyze database content in order to produce the representations. Second, an interface must be constructed which allows users to interact with these representations in order to improve their queries. Each of these components in turn involves a number of processes, described in the following sections. The *InfoView* information retrieval system used in this study is based on the model described.

Database Analysis

The first step in producing database representations is index-term selection (indexing). In general the goal is to select representative index terms which are capable of discriminating among database topics. Although retrieval itself may require complete indexing, the index terms displayed to users are typically a subset selected to be representative of database content and capable of discriminating among topics. Once a set of index terms has been selected, pair-wise distance estimates are obtained by transforming co-occurrence data. We typically calculate the co-occurrence of index terms within sentence units and transform these data using Dice's coefficient. However, larger units (e.g., paragraphs or documents) and other transformations may be more effective. Finally, networks are produced using the Pathfinder network algorithm which effectively eliminates connections (associations) among index terms until only the strongest or most important associations remain (McDonald, Plate, & Schvaneveldt, 1990; Schvaneveldt, Durso, & Dearholt, 1989).

The User Interface

The Pathfinder algorithm takes as input a matrix of distances estimates between entities, in this case index terms, and produces as output a network specification. For the purposes of information retrieval, the nodes in these networks are index terms and the links represent the strongest associations between the terms based on frequency of co-occurrence. In order for users to interact with these associative networks, graphical representations are produced. Because the database networks are typically very large, often consisting of thousands of nodes, they cannot be effectively displayed. In order to focus interaction within appropriate subsections of the complete network, users submit one or more index terms and portions of the complete index-term network (i.e., subnets) are selected and graphically displayed. The size of these subnets is specified as the

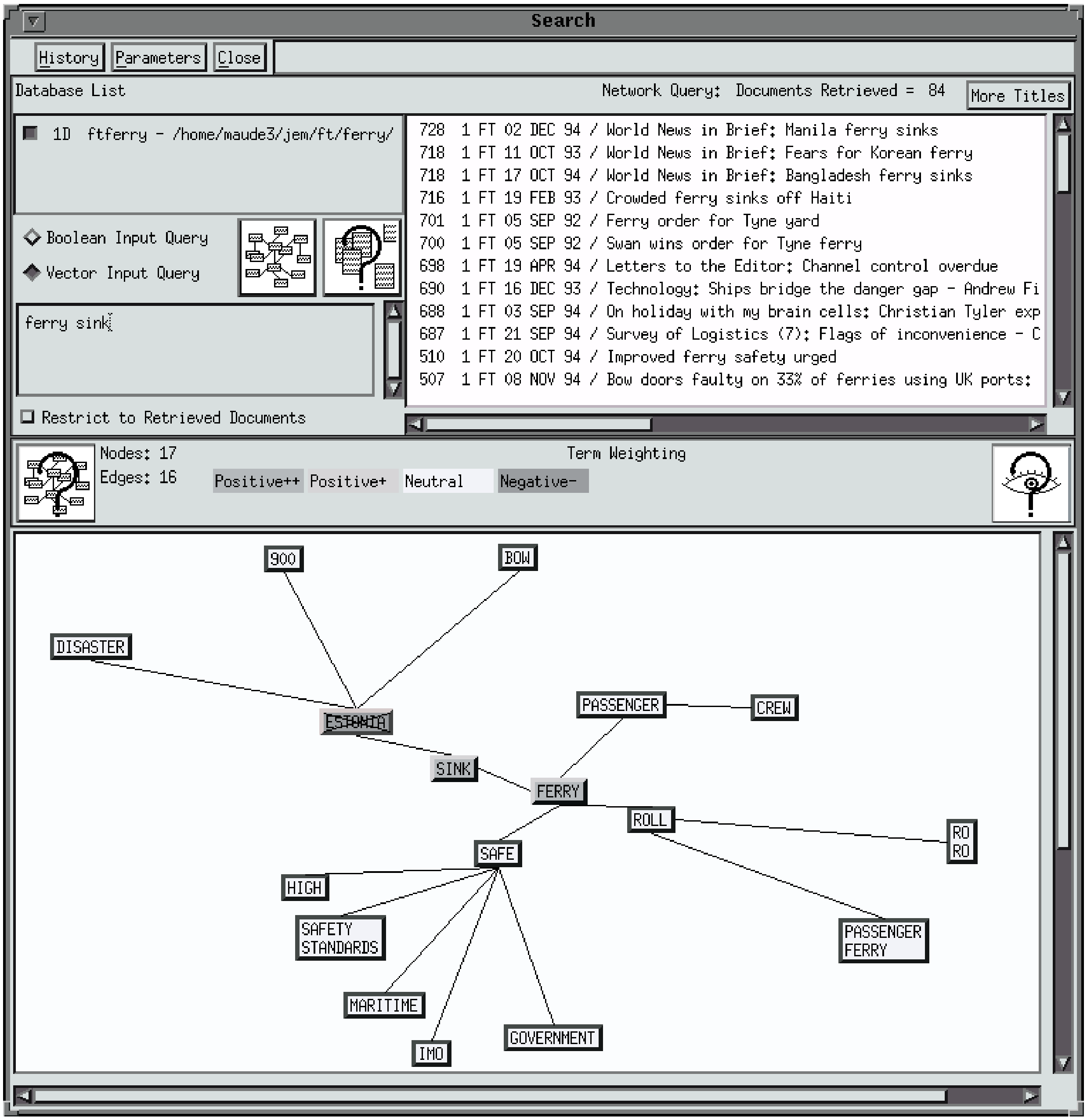

Figure 1. The *InfoView* user interface

number of nodes within a given radius (i.e., number of links) of the query terms.

The networks are used to improve information statements, or queries, by helping users to 1) substitute precise index terms for vague ones, 2) remove or "negatively weight" inappropriate index terms, or 3) add appropriate index terms. The networks can be thought of as "pictures" representing the associations in the database and are to be used to modify queries so as to do a better job of retrieving relevant documents. Users can modify their query by interacting directly with the network display. For example, Figure 1 shows the result of a user interaction with the *InfoView* system during a TREC interactive task. The topic "Ferry Sinkings" requires the user to find as many reports of ferry sinkings where 100 or more people lost their lives. The subnet show that the key terms "ferry" and "sink" are highly associated with the word "Estonia" which was the name of a ferry which sunk, killing over 900 persons thus gaining a lot of publicity. Because most of the reported ferry sinkings are referring to the "Estonia" and the user is trying to find *other* incidents, the user has used the mouse to "negatively weight" the "Estonia" term. The resulting list of document titles (displayed in descending order of estimated relevance in the top right corner of the display) shows the top four relevant documents are indeed about other ferry sinkings.

The *InfoView* model produces networks based on a complete analysis of a specified database. The subnets displayed to users are thus portions of the complete network and are based on the co-occurrences of index terms throughout the collection and across topics. We have observed that networks of this sort, derived from an analysis of the complete database, are not always optimal for information retrieval purposes. This limitation seems to stem from the fact that index terms in such networks become connected because they tend to be associated throughout the collection and across topics. However, what's needed are index terms capable of discriminating between topics, and such associations are often weakened by a global analysis. In this study we modified the *InfoView* model to produce associations based on the co-occurrence of index terms within sets of retrieved documents, rather than across the entire collection. Our TREC-6 Interactive Track effort was designed to test this approach.

TREC-6 Interactive Track Method and Results

Database preparation for the current study consisted of the following steps. First, a subset of documents from the Financial Times database were selected using a Boolean query with the terms in titles for each of the six TREC-6 Interactive Track topics. The sets of documents retrieved were treated as independent databases and Pathfinder networks were produced using the procedure outlined above. When participants were required to perform the aspectual recall task on a given topic, the network for that database was accessed by the *InfoView* system. This is a simulation of the system we propose to build and has several limitations. First, by selecting documents using Boolean search, it is quite likely that some relevant documents were excluded from the individual topic databases and hence were not available to participants, limiting recall. In the next version of our system we will use Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990) to select relevant document sets. Second, although the document sets retrieved for topics were relatively small and more homogenous than the complete collection, further improvements are hypothesized if these sets of documents are grouped or clustered in LSI space and networks formed on these topic sets.

InfoView was used as our “experimental” system in the TREC-6 interactive track according to the method specified for the track. (see http://www-nlpir.nist.gov/~over/t6i/ for the complete specification.) We had four participants each using both *InfoView* and *Zprise* (the control system). Their results were combined with track participants from the seven other interactive track sites. Like all other sites, our experimental, *InfoView* system did not significantly differ with respect to any dependent measure. In general, our participants using InfoView were faster (67 sec) with as least as good of aspectual recall as when they were using the *Zprise* control system.

Future work

Despite the inconclusiveness of the present study we are encouraged by the progress we are making toward better interactive systems and their evaluation. We are in the process of experimenting with the aspectual-recall comparison paradigm. We are currently conducting an extension of the method used in the TREC-6 Interactive track by extending the latin square design to control for task order. Further we intend to look at a new dependent measure we call “aspectual change”. Aspectual change will measure the difference between what people know about a topic before and after an interactive information retrieval session. We believe this could be a very sensitive measure of the effectiveness of an information retrieval system’s user interface.

References

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

McDonald, J. E., Plate, T. A., & Schvaneveldt, R. W. (1990). Using Pathfinder to extract semantic information from text. In R. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization. (pp. 149-164). Norwood, NJ: Ablex.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 24, pp. 249-284). New York: Academic Press.

Rocchi, Jr., J. J. (1971) Relevance Feedback in Information Retrieval. In G. Salton (Ed.) The Smart System - Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, JN, 313-323.

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