21

Visual interface for textual information retrieval systems

A. Veerasamy, S. Hudson and S. Navathe
Georgia Institute of Technology
College of Computing, 801, Atlantic Drive
Georgia Institute of Technology,
Atlanta, Georgia 30332-0280, USA.
Email: {veerasam, hudson, sham}@cc.gatech.edu

Abstract

A prototype user interface implementation for text information retrieval system is described. Using a visualization scheme, the interface provides visual feedback to the user about how the query words influence the ranking of retrieved documents. The interface also helps the user in constructing complex structured queries by simple drag-and-drop operations. An intuitive model where the user classifies the information provided to him/her as being positive and negative aids him/her in supplying rich relevance feedback information to the system. Our prototype interface has been built on top of INQUERY [Callan *et al.*, 1992]. Preliminary experience with the interface shows it to be a valuable tool in aiding the interactive search process between the user and the system. To test the effectiveness of the interface, we plan to conduct studies on users with real information need searching a large corpus of articles.

Keywords

Visualization of results, visual query languages, query processing, information retrieval

1 USER INTERFACE ISSUES FOR INFORMATION RETRIEVAL SYSTEMS

User Interface issues and interaction techniques for full text information retrieval systems have in general received much less attention than system issues like document representation and retrieval algorithms. We have developed an interface that facilitates the user in visually constructing powerful queries for ranked output retrieval systems. The interface includes a scheme for visualizing the query results in a form that enables the user to see

S. Spaccapietra et al. (eds.), *Visual Database Systems 3*

the relationships between the query results and the query. While a majority of online library catalog systems use a boolean model of retrieval, a vast majority of existing experimental information retrieval systems retrieve a ranked set of documents in decreasing order of relevance in response to a free-form textual query. In ranked output systems, the documents and the queries are modeled by a set of weighted index terms. The index term weighting function for the documents primarily takes into consideration

- the frequency of occurrence of the index term in the document,
- the number of documents in the corpus containing that index term.

The effectiveness of a retrieval system is measured by two metrics: recall (the ratio of the number of relevant documents retrieved to the total number of relevant documents in the corpus) and precision (the ratio of the number of relevant documents retrieved to the total number of documents retrieved). The reader is referred to [Belkin and Croft, 1987, Rijsbergen, 1979, Salton and McGill, 1983] for a comprehensive description of evaluation metrics of information retrieval systems, document representation and retrieval techniques.

While processing a free-form textual query, most ranked output Information Retrieval systems automatically extract index terms from the query and weight them. The weighted query index terms are then matched against the weighted index terms of documents to retrieve a ranked set of documents in decreasing order of relevance. Each document is weighted, the higher the weight of a document, the more likely it is to be relevant to the query. Most of the existing library information systems (On-line Public Access Catalogs, OPAC) follow a boolean retrieval model. In this model, the documents retrieved in response to a boolean query are not ranked. If a document satisfies the boolean query specification, it is retrieved. Compared to boolean systems, ranked output systems are a significant improvement since the query can be in a free-form text as opposed to a strict boolean syntax. Also, the retrieved documents are ranked, thereby placing the more useful documents at the top of the list. This is a particularly useful feature since it has been shown that users of boolean systems spend a considerable effort in reducing the size of the result set [Spink, 1993]. On the other hand, ranked output systems introduce a new problem: For a naive user, the logic behind the ranking of documents in response to a query is not as apparent and straightforward as a boolean system. The interface we have developed is aimed at alleviating this problem. It helps the user in understanding how the system computed the ranking of retrieved documents by visualizing the relationship between query terms and the results of the query.

The interface also aids the user in formulating complex structured queries by graphically manipulating objects on the screen. A simple mechanism of classifying any information on the screen into positive and negative instances lends itself to easy formulation of structured queries. The interface is built using Tcl/Tk [Ousterhout, 1994] on top of INQUERY [Callan *et al.*, 1992], a ranked output retrieval system based on Bayesian inference networks. The interface supports two types of feedback:

- feedback from the user to the system and

- feedback from the system to the user.

It is interesting to note that the term "feedback" in the field of Information Retrieval typically refers to user's feedback to the system, while in the field of Human Computer Interfaces, "feedback" usually refers to the system's feedback to the user. The user's feedback to the system and the different levels of granularity at which the feedback can be provided is discussed in section 3. The system's feedback to the user and the visualization technique is discussed in section 4.

2 RELATED WORK

Numerous studies on user interaction with online library access catalog systems with a boolean retrieval model have been conducted [Spink, 1993, Spink and Saracevic, 1992, Dalrymple, 1990, Fidel, 1991a, Fidel, 1991b, Fidel, 1991c]. Spink [Spink, 1993] studies the different forms of user feedback during a retrieval session. Of the total number of feedback actions by the user, 45% were aimed at adjusting the size of the retrieved set of documents, and about 40% were related to relevancy of documents. Fidel [Fidel, 1991a, Fidel, 1991b, Fidel, 1991c] discusses the issue of user interaction by studying the process of search term selection and searching styles in online library access catalogs. Dalrymple [Dalrymple, 1990] looks at the feedback process from a user-centered perspective. Bates [Bates, 1990] describes a boolean retrieval system which integrates an online thesaurus. None of the above studies involve a ranked output system supporting free-form textual queries. All of the systems deal with boolean retrieval model only. We believe that there is a significant difference in the way users interact with a boolean system and a ranked output system. The reader is referred to [Harman, 1992] and [Hancock-Beaulieu, 1992] for a comparative discussion of boolean systems and ranked output systems. While building our interface, we have borrowed valuable ideas from the studies mentioned above. In particular, the need to integrate an on-line thesaurus with the search interface in an easy-to-use fashion and a simple interaction scheme to include words from documents into the query have been influenced by the results of above-mentioned studies.

Walker and Beaulieu [Walker, 1987, Hancock-Beaulieu, 1992] describe their OKAPI system which is a ranked output retrieval system for library catalogs. Similarly, Fox [Fox *et al.*, 1993] describes their MARIAN sysem which is also a ranked output system for library catalogs based on the vector-space model. While OKAPI has facilities for relevance feedback and query expansion using a thesaurus, it largely lacks any means of providing system feedback to the user about how the ranking was computed. The interface we have developed integrates relevance feedback information from the user as well as feedback from the system illustrating the relationship between query results and query words.

A number of visualization schemes for information retrieval systems have also been proposed. The perspective wall [Card *et al.*, 1991] describes a visualization scheme which supports browsing of documents. While such a system will not handle qualitative doc-

ument classifications such as library subject catalogs, it is very useful for visualizing documents based on data which is linear in nature (like date of publication). Other visualization schemes such as [Korfhage, 1991, Spoerri, 1994, Hemmje *et al.*, 1994] have facilities for viewing a large document space. But visualizing the document space along more than 3 - 4 dimensions simultaneously becomes very cumbersome using the above systems. Also, most of them do not provide support for querying with relevance feedback and none of them provide support for query expansion using a thesaurus. The visualization scheme in our interface can gracefully handle much higher number of query word dimensions.

2.1 NOVELTY OF OUR APPROACH

The novelty of our system is in integrating a diverse set of interaction features in a seamless fashion into a single system thereby facilitating the interactive and iterative nature of the information seeking process. The following features are integrated in our system:

- Using a visualization scheme, the interface provides visual feedback to the user about how the query words influence the ranking of retrieved documents.
- By simple drag-and-drop operations of objects on the screen, the interface facilitates a naive end-user in constructing complex structured queries and in providing relevance feedback. This feedback is utilized by the system in a manner described later.
- The interface integrates an online thesaurus which provides words related to the query that can be used by the user to expand the original query.

Belkin and his group's work [Belkin *et al.*, 1993, Belkin *et al.*, 1991] on user interfaces for information retrieval systems [Henninger and Belkin, 1994] elucidates the issues in user interface and interaction techniques for full text retrieval systems. Belkin [Belkin *et al.*, 1991] mentions that

> This type of analysis led to another important conclusion, namely that information systems for end users must support a variety of goals and tasks, but through some common interface or seamless access mechanism to a variety of relevant information sources and system functionalities.

Our interface takes a step in that direction by integrating different pieces of information with a visualization scheme and simple interaction techniques.

3 INTERACTIVE CONSTRUCTION OF QUERIES

Searching a database for information is a highly interactive process with the user constantly refining the query after examining the results of previous iteration until he/she is

either satisfied with the results or is frustrated with the process and gives up. In existing information retrieval systems, the interaction proceeds by the user providing feedback on which of the retrieved documents are relevant to his/her information need. The system uses this information to modify the original query resulting in an improved ranking of retrieved documents. It has also been shown by Spink [Spink and Saracevic, 1992] that during iterative query reformulation, users tend to expand the query using search terms from various sources such as a thesaurus, previously retrieved documents and user's background knowledge. Expanding the query with terms from such sources can contribute to retrieval of more relevant documents in the next iteration.

Our interface encourages the interaction between the user and the system by providing the user with simple interaction technique to let him/her supply relevance feedback at different levels of granularity: whole documents, document portions, phrases and individual words. Almost any information appearing on the screen can be used for feedback. This is achieved by simple "drag-and-drop"ping the feedback object into either a "Positive Objects" window colored green or a "Negative Objects" window colored red. This scheme provides a simple abstraction to the user for classifying any type of information without having to worry about what action to take for what type of information. A typical user session along with the response of the interface for every user action is described below using an example (please refer to Figure 1). The database being queried contains a collection of titles, authors and abstracts of thousands of CACM articles.

- The user types in his free form textual query in the query window. In the example shown in figure 1, the query is "image audio and text data compression".
- As every query word is typed in, the system consults an on-line thesaurus and displays words and phrases related to the query word in an adjacent window.
- At any point during the session the user can drag-and-drop any of the related words/phrases into the positive and negative windows. Internally the system expands the query by treating the positive words/phrases as synonyms of the corresponding query word. The negative words/phrases are included in the query with a NOT operator. For example, if for a query word "bank", the phrase "financial institution" is classified as positive and "river bed" is classified as negative, the corresponding internal query would be "#SYNONYM(bank #2[1](financial institution)) #NOT(#2(river bed))". The end-result of this classification is a possible improvement in the precision measure since documents containing the phrase "river bed" will be weighted lower than other documents, and a possible improvement in the recall measure since documents containing the phrase "financial institution" are also retrieved. The interface facilitates construction of such structured queries by simple drag-and-drop operations. In the example in figure 1, three words related to the query word "compression", namely, "compaction, "shortening" and "condensation" have been classified as positive. Internally the systems treats these three words as synonyms of "compression".

[1]#2 is the proximity operator in INQUERY specifying that the words should appear within a distance of 2 within each other

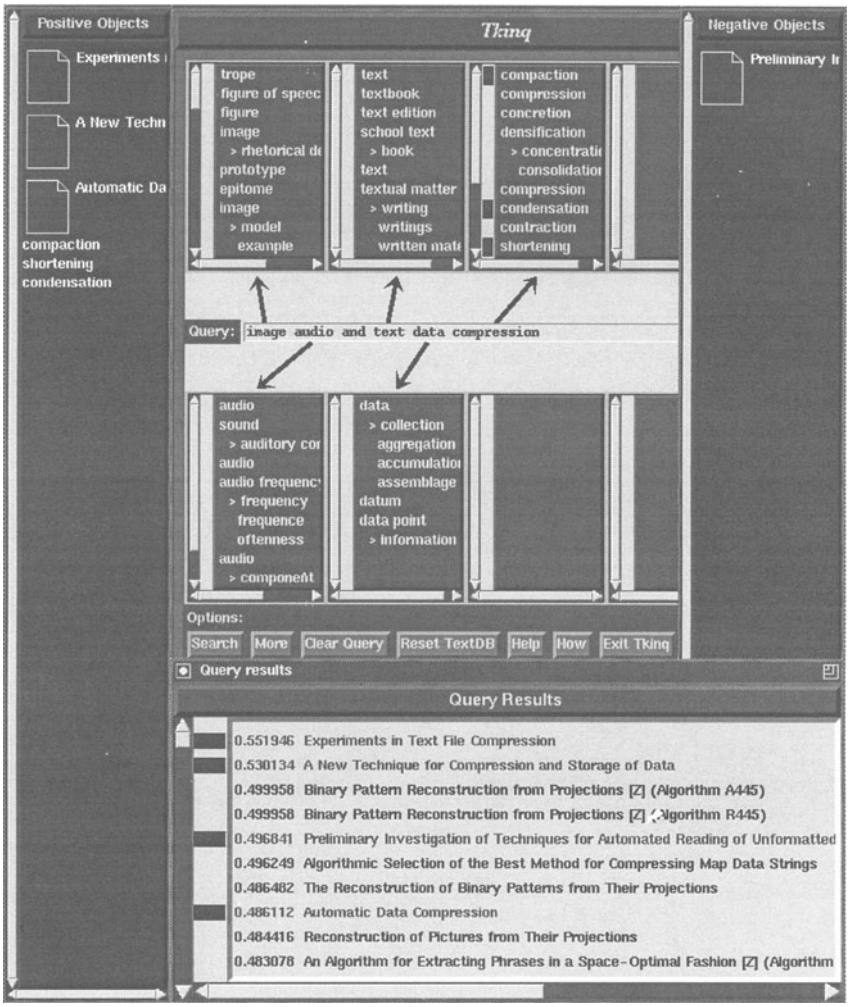

Figure 1: Sample querying session. The window titled "Positive object" is colored green and the window title "Negative Objects" is colored red. When a document is classified as positive/negative, the title of that document in the "Query results" window is also colored green/red.

- After the user types in the query, the system evaluates the query and displays the titles of top-ranked documents in the "Query Results" window.

- The user examines the query result. Double-clicking any title with the mouse will bring up the full document.

- The user can classify any document as being relevant or non-relevant by drag-and-drop'ing the document into positive and negative windows. In the example in figure 1, the user has classified three documents titled "Experiments in text file compression", "A new technique for compression and storage of data" and "Automatic data compression" as positive. The document titled "Preliminary investigation of techniques for automated reading of unformatted text" has been classified as negative. Internally, the systems extracts 4 - 6 high frequency words from the positive documents and adds it to the query thereby expanding the query. This results in the retrieval of documents similar to the positive documents.

- The user can also highlight a portion of a document and drag-and-drop it into the positive and negative windows. The words in the highlighted document portion are used to expand the query in the next iteration.

- During the next iteration, the reformulated query with the relevance feedback information is processed by the system resulting in an improved ranking of documents.

The positive and negative windows for feedback are aimed at mimicking the user's view that some information is in line with the information need and some not. After an object has been classified as positive (or negative), the system always colors the object green (or red) whenever the object is displayed, thereby reinforcing the user with the fact that the object is being used for relevance feedback. While arguing for the use of direct manipulation techniques for Information Retrieval, Mitev [Mitev, 1989] mentions that

> "Parts of document(s), individual word(s), sentences or groups of word(s) displayed could be used directly as something to be input for another search. This could be done, for example, by pointing and 'picking' them on the screen and carrying them across another area of the screen. The user would not have to input them again."

This is precisely what has been accomplished in our interface. In their retrieval system, Campbell [Campbell and Sanderson,] uses a cut-and-paste mechanism for relevance feedback by letting the user add portions of retrieved documents back into the query window.

This section dealt with the interaction technique to let the user provide relevance feedback information to the system. The next section deals with visual feedback from the system on how the query results were computed.

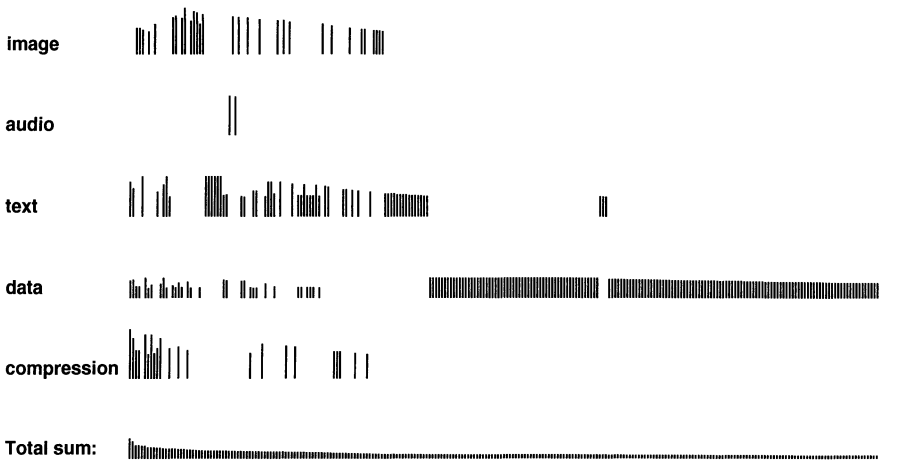

Figure 2: Visualization of results for the base query.

4 VISUALIZATION OF QUERY RESULTS

While systems with a boolean retrieval model retrieve an unordered set of documents in response to a query, ranked output information retrieval systems retrieve a ranked set of documents. While the reason for retrieving a document is fairly clear in the case of a boolean system, the reason why a document is assigned a specific rank is not apparent in the case of a ranked output system. Without knowing how the system computed the ranking of documents, the user will have to treat the retrieval mechanism as a black box. We stand to gain a lot by keeping the user more informed about the retrieval process of the system. If the user has more information about how the ranking was computed, he/she will be in a better position to reformulate the query for the next iteration. He/she can take into account the deficiencies of the system in adjusting his/her query. It will also help in reinforcing the right mental model.

In our interface, we keep the user informed about the retrieval mechanism by providing visual feedback about how the query results are related to the query words. This is done by a visualization scheme as shown in the figure 2. The visualization reveals the extent to which each query word was responsible for retrieving the set of documents. The visualization consists of a set of histograms, one for every query word (except stop words) typed in by the user, and one histogram for the total query (labeled "Total sum"). All the histograms are placed one below the other with the "Total sum" histogram appearing at the bottom and the query-word-histograms appearing in the order in which query words were typed in. Each histogram consists of a set of vertical bars, one bar for each retrieved document. For the top ranked document, a vertical bar is drawn in the leftmost position (i.e, lowest X coordinate position) in the "Total sum" histogram. The height of the bar is

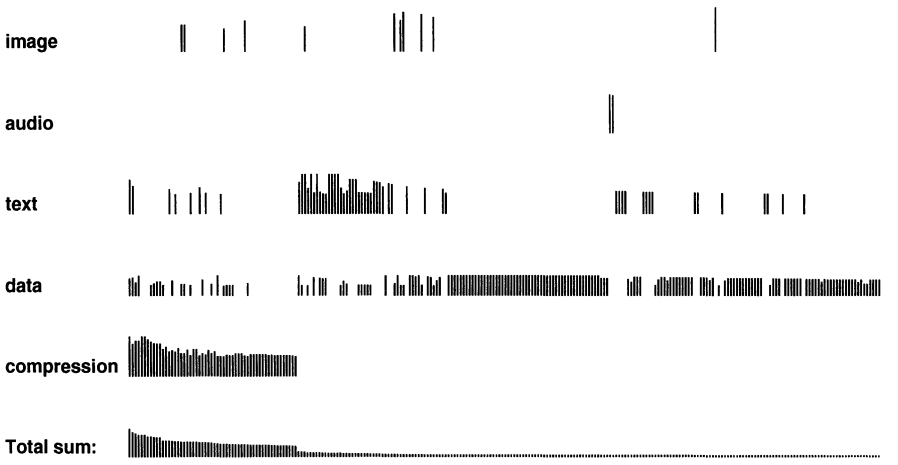

Figure 3: Visualization of results for query with feedback information.

proportional to the weight of the document. (Note that each document is given a weight. The higher the document weight, the more likely it is to be relevant to the query.) For the same document, vertical bars in the same X-coordinate position are also drawn in the query-word-histograms. The height of the vertical bar in any given query-word-histogram is proportional to the weight of the query word in that document. It represents the contribution of the query word in retrieving that document. If the query word does not appear in the document, thereby getting a weight of zero, a bar of zero height is drawn which shows up as an empty space in that X-coordinate position. The second ranked document occupies the next higher X-coordinate to the right and so on upto a maximum of top 200 documents.

The visualization shown in Figure 2 corresponds to the base query with no feedback information from the user. We can see that all but two of the top 200 documents have nothing to do with audio. Almost all of the second half of the 200 documents were retrieved because they contained the query word "data". More significantly, only about 10% of the documents have anything to do with compression – which is the crux of the query. This illustrates that the query should be expanded with more words related to "compression". In fact, the decision to classify the three synonyms of "compression" (as shown in figure 1) was made after examining the distribution of "compression" in the visualization. Figure 3 shows the distribution of query terms in the query result for the revised query in the second iteration with all the feedback information. We can see that almost all the documents about "compression" have been ranked at the very top. Also there are more documents retrieved due to "compression" because of the synonyms and the positively classified documents. Our experience with this visualization scheme has shown it to be very useful in identifying differents facets of the query.

5 CONCLUSION & FUTURE WORK

A prototype interface for a ranked output information retrieval system has been implemented. The interface facilitates the inherently interactive nature of the information seeking process. Drag-and-drop operations form the basis of interaction encouraging the user to provide feedback information to the system and helps in the dialog between the user and the system. Almost any information on the screen can be used by the user to provide feedback information. An online thesaurus, WordNet [Miller *et al.*, 1990], is integrated with the interface to form a single system.

The interface also supports a visualization scheme which illustrates how the query results are related to the query words. Visualizing the results of the query keeps the user more informed on how the system computed the ranking of documents. With this information, the user is better equipped to reformulate the query for the next iteration. It is our opinion that integrating all of the above features in a seamless interface leads to an interplay between different items that is much more beneficial than the sum of the individual items in isolation.

In demonstrating the system to the reference librarians at Georgia Tech and in observing casual users of the system, we believe that the features we have implemented in this system contributes to enhancing the end-user's interaction with the system. As a result, the system is better able to assess the user's need and the user has a better understanding of the system's inference. However we cannot categorically conclude the effectiveness and the utility of the interface without conducting formal user-studies.

In future, we plan to test the effectiveness of the interface by conducting two studies: One with users having real information needs searching a traditional library database and another with volunteers searching the TREC [TRE, 1994] document collection with supplied search statements. Since all the relevant documents for the supplied search statements in the TREC collection are known, recall and precision of searches performed with our interface can be compared against other systems.

6 ACKNOWLEDGMENTS

We are thankful to Dr. Bruce Croft for letting us use the INQUERY retrieval system. Many thanks to Dr. Marti Hearst whose Tcl/Tk code for the SMART system was helpful as a spring board for us to write the interface. Support in part by ARPA Grant No. F33615-93-1-1338 under the Intelligent Integration of Information Program is appreciated.

References

[Bates, 1990] Marcia J. Bates. Design for a subject search interface and online thesaurus for a very large records management database. In *Proceedings of the 53rd Annual*

Meeting of the American Society for Information Science, pages 20–8, 1990.

[Belkin and Croft, 1987] Nick Belkin and W.B. Croft. Retrieval techniques. In E. Martha, editor, *Annual Review of Information Science Technology*, pages 110–45. Elsevier Science Publishers, 1987.

[Belkin *et al.*, 1991] N.J. Belkin, P.G. Marchetti, M. Albrecht, L. Fusco, S. Skogvold, H. Stokke, and G. Troina. User interfaces for information systems. *Journal of Information Science*, 17:327–44, 1991.

[Belkin *et al.*, 1993] N.J. Belkin, P.G. Marchetti, and C. Cool. Braque: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–44, 1993.

[Callan *et al.*, 1992] J.P. Callan, W.B. Croft, and S.M. Harding. The inquery retrieval system. In *Third International Conference on Database and Expert Systems Applications*, September 1992.

[Campbell and Sanderson,] I. Campbell and M. Sanderson. Personal communication. University of Glasgow.

[Card *et al.*, 1991] S. Card, G. Robertson, and J. Mackinlay. The information visualizer, an information workspace. In *Proceedings of CHI 91 Human Factors in Computer Systems.*, 1991.

[Dalrymple, 1990] P.W. Dalrymple. Retrieval by reformulation in two library catalogs: toward a cognitive model of searching behaviour. *Journal of the American Society for Information Science*, 41(4):272–81, 1990.

[Fidel, 1991a] Raya Fidel. Searcher's selection of search keys: I. the selection routine. *Journal of the American Society for Information Science*, 42(7):490–500, 1991.

[Fidel, 1991b] Raya Fidel. Searcher's selection of search keys: II. controlled vocabulary or free-text searching. *Journal of the American Society for Information Science*, 42(7):501–14, 1991.

[Fidel, 1991c] Raya Fidel. Searcher's selection of search keys: III. searching styles. *Journal of the American Society for Information Science*, 42(7):515–27, 1991.

[Fox *et al.*, 1993] Edward A. Fox, Robert K. France, Eskinder Sahle, Amjad Daoud, and Ben E. Cline. Development of a modern OPAC: From REVTOLC to MARIAN. In Robert Khorfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of sixteenth ACM SIGIR conference*, pages 248–59. ACM SIGIR. June-July 1993.

[Hancock-Beaulieu, 1992] Micheline Hancock-Beaulieu. User friendliness and human-computer interaction in online library catalogues. *Program*, 26(1):29–37, January 1992.

[Harman, 1992] Donna Harman. User-friendly systems instead of user-friendly front-ends. *Journal of American Society for Information Science*, 43(2):164–74, 1992.

[Hemmje *et al.*, 1994] Matthias Hemmje, Clemens Kunkel, and Alexander Willet. Lyberworld – a visualization user interface supporting full text retrieval. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, pages 249–59, 1994.

[Henninger and Belkin, 1994] Scott Henninger and Nick Belkin. Tutorial on interface issues and interaction strategies for information retrieval systems. In *Human Factors in Computing Systems CHI 94 Conference Companion*, pages 387–8, 1994.

[Korfhage, 1991] Robert Korfhage. To see, or not to see – is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR conference on Research and Development in Information Retrieval*, pages 134–41, 1991.

[Miller *et al.*, 1990] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–44, 1990.

[Mitev, 1989] Nathalie N. Mitev. Ease of interaction and retrieval in online catalogues: contributions of human-computer interaction research. In Charles R. Hildreth, editor, *The online catalogue*, chapter 8, pages 142–76. Library Association Publishing, London, 1989.

[Ousterhout, 1994] John K. Ousterhout. *Tcl and the Tk Toolkit.* Addison-Wesley, 1994.

[Rijsbergen, 1979] Keith Van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[Salton and McGill, 1983] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, New York, 1983.

[Spink and Saracevic, 1992] Amanda Spink and Tefko Saracevic. Sources and use of search terms in online searching. In *Proceedings of the 55th Annual Meeting of the American Society for Information Science*, pages 249–55, 1992.

[Spink, 1993] Amanda Spink. Interaction with information retrieval systems: Reflections of feedback. In *Proceedings of the Annual Meeting of the American Society for Information Science*, pages 115–21, 1993.

[Spoerri, 1994] Anslem Spoerri. Infocrystal: A visual tool for information retrieval and management. In *Human Factors in Computing Systems CHI 94 Conference Companion*, pages 11–2, 1994.

[TRE, 1994] In D.K. Harman, editor, *The Second Text REtrieval Conference (TREC-2).* NIST Special Publication, March 1994.

[Walker, 1987] Stephen Walker. Okapi: Evaluating and enhancing an experimental online catalog. *Library Trends*, Spring:631–45, 1987.

7 BIOGRAPHY

Aravindan Veerasamy is a PhD student in the College of Computing at Georgia Tech. He earned his B.E. in computer science at College of Engineering, Guindy, India. For his PhD, he is investigating research issues related to user interface design for Information Retrieval systems. His research interests include human information seeking behaviour, document databases and user interfaces.

Scott Hudson is an associate professor in the College of Computing at Georgia Tech and a member of the Graphics, Visualization, and Usability Center. His research interests include a wide range of topics in user interface software. He was previously an Assistant Professor of Computer Science at the University of Arizona. He earned his Ph.D. in computer science at the University of Colorado in 1986. He has regularly served on program committees for the SIGCHI and UIST conferences, and served as Program Chair for UIST '90 and Symposium Chair for UIST '93. He currently serves as an Associate Editor for ACM Transactions on Computer Human Interaction.

Shamkant Navathe is a professor in the College of Computing, at Georgia Tech. His research interests include data modelling, database design and integration, intelligent information retrieval, engineering database applications and federated databases. He has co-authored (with R. Elmasri) a widely used database textbook "Fundamentals of Database Systems" and another book (with C. Batini and S. Ceri) entitled: "Conceptual Database Design: An Entity Relationship Approach." He is the general co-chair of the 1996 VLDB Conference at Bombay India.