


# **Artificial Intelligence in Text-Based Information Systems**

Dik L. Lee, Ohio State University

W. Bruce Croft, University of Massachusetts, Amherst



**A WINDY DAY  
WHEAT FROM CHAFF:  
YOU HAVE TO SORT  
THROUGH A LOT OF WHAT  
YOU DON'T WANT TO FIND  
WHAT YOU NEED. THIS  
SPECIAL SERIES EXAMINES  
AI TECHNIQUES BEING  
APPLIED TO INFORMATION  
RETRIEVAL PROBLEMS.**

Information retrieval has been investigated for several decades by researchers in information science, library science, and computer science. The field is becoming more and more important because of the increasing amount of electronic information available and the advances in technologies such as multimedia systems, storage media, and computer networks.

In the past, text databases were mainly used in centralized environments: Information centers such as CompuServe and the On-line Computer Library Center maintained the databases, and users accessed them by connecting (mostly by dialing) to the center. Recently we have seen a trend toward replacing paper with electronic media, such as CD-ROM. For example, many computer vendors are sending their technical manuals to customers on CD-ROM instead of paper. Consequently, people are increasingly using textual databases directly without the help of intermediaries such as librarians.

Text-based applications need effective retrieval methods that can handle ad hoc and ill-defined queries, but the retrieval facilities available to users have been rather primitive. Commercial systems are almost exclusively based on the Boolean retrieval model: The user specifies a Boolean combination of keywords, and documents are retrieved based on their satisfying the condition. More recently, commercial systems based on statistical models have appeared, in which documents are ranked according to some similarity function defined by the system.

The functionality required of a text-based information system is also increasing. Users are no longer concerned only with retrieval but also with other aspects of text management, including document imaging, routing, dissemination, extraction, and categorization. Traditional information retrieval techniques cannot satisfy all these new demands. AI techniques, on the other hand, are considered by many as having promise for text-based applications. Unfortunately, the results reported thus far on the application of AI to information retrieval have been inconclusive, and many claims remain unsupported by large experiments or real application experience.

*IEEE Expert's* special series will examine recent projects applying AI techniques to traditional information retrieval problems. We hope it will bring researchers

in AI and information retrieval closer together, raise their awareness of work done in both fields, and increase their synergy, which will benefit both fields.

First, Bruce Croft's introduction to the major research issues in information retrieval describes recent developments in knowledge-based approaches. Croft argues that although a fair amount of work has been done, the effectiveness of this approach has yet to be demonstrated. He suggests that statistical techniques and knowledge-based approaches should be viewed as complementary rather than competitive.

The second article, by Paul Jacobs, describes a system that categorizes news stories by integrating statistical and knowledge-based techniques. Categorization of news articles facilitates dissemination, retrieval, and browsing. Jacobs' system uses a statistical technique to obtain simple lexicosemantic patterns from a large set of training data, while complex patterns are developed manually.

The third article, by Hsinchun Chen, Kevin Lynch, Koushik Basu, and Tobun Dorbin Ng, reports on the cooperative use of several thesauri for concept-based retrieval. Like Jacobs' system, this approach uses statistical techniques, in this case to generate concept relations for one of the thesauri. The thesauri are tied together with a blackboard architecture and controlled by a neural-net-based activation module that identifies concepts in the thesauri related to the user's search criteria.

To some extent, these articles show the need for statistical techniques in knowledge-based environments and for multiple domain knowledge sources to overcome information retrieval problems. Articles reporting on other techniques will be featured in subsequent issues.

## Acknowledgments

We are indebted to Chid Apte, member of the *IEEE Expert* editorial board, who oversees the development of this special series, and to B. Chandrasekaran, editor-in-chief of *IEEE Expert*, for his support and guidance. Special thanks go to the authors who submitted manuscripts and to the reviewers whose timely and professional reviews are vital to the series' success.



**Dik L. Lee** is associate professor of computer and information science at Ohio State University. His current research is in document retrieval and management, knowledge-based systems, and object-oriented and heterogeneous database systems. A

member of the IEEE Computer Society, IEEE, and ACM, Lee was the guest editor of *IEEE Data Engineering Bulletin's* special issue on document processing in 1990. He is an ACM lecturer and a member of the program committee for the International Conference on Data Engineering. He received his BSc in electronics from the Chinese University of Hong Kong in 1979 and his MSc and PhD in computer science from the University of Toronto in 1982 and 1985, respectively.

Readers can reach Lee at the Department of Computer and Information Science, Ohio State University, 2036 Neil Ave., Columbus, Ohio 43210-1277; email, dlee@cis.ohio-state.edu



**W. Bruce Croft** is professor of computer science at the University of Massachusetts, Amherst. His research interests are in formal models of retrieval for complex, text-based objects, text representation techniques, the design and implementation of

text retrieval systems, and user interfaces. Croft chaired the ACM Special Interest Group on Information Retrieval from 1987 to 1991. He is an associate editor of the *ACM Transactions on Information Systems* and *Information Processing and Management* and an editorial-board member of *Knowledge Acquisition*. He has served on numerous program committees and helped organize many workshops and conferences. He received his BSc and MSc from Monash University in Melbourne, Australia, in 1973 and 1974, respectively, and his PhD in computer science from the University of Cambridge, England, in 1979.

Readers can reach Croft at the Department of Computer Science, University of Massachusetts, Lederle Graduate Research Center, Amherst, MA 01003; email, croft@cs.umass.edu