N PAGE

Form Approved
OMB No. 0704-0188

Public rep... gathering... collection... Davis Highway, ... hour per response, including the time for reviewing instructions, searching existing data sources, ...ection of information. Send comments regarding this burden estimate or any other aspect of this ...ngton Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson ...ment and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

**1. AGENCY USE ONLY** *(Leave blank)*

**2. REPORT DATE**
October 1993

**3. REPORT TYPE AND DATES COVERED**
Final Technical 15 Aug 91 - 14 Aug 93

**4. TITLE AND SUBTITLE**
Retrieval Using Plausible Inference (u)

**5. FUNDING NUMBERS**
G: AFOSR-91-0324
PR: 2304
TA: A7

2

**6. AUTHOR(S)**
Professor W. Bruce Croft

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
University of Massachusetts, Amherst
Box 36010, OGCA, Munson Hall
Amherst, MA 01003-6010

**8. PERFORMING ORGANIZATION REPORT NUMBER**
FTR-528996

AFOSR-TR-

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Dr. Abraham Waksman
AFOSR/NM
Building 410
Bolling AFB DC 20332-6448

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
AFOSR-91-0324

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**
Approved for public release;
distribution unlimited.

**12b. DISTRIBUTION CODE**
UL

**13. ABSTRACT** *(Maximum 200 words)*

The goal of this project was basic research and development in the area of text-based information systems. A new probabilistic model of text retrieval using Bayesian inference nets was proposed and extended. An efficient implementation of this model for large full-text databases resulted in the INQUERY indexing and retrieval engine. Retrieval experiments carried out using INQUERY an a variety of databases have shown that it produces high-quality results compared to other approaches. We have also made significant progress, both conceptually and with prototype implementations, in showing how text retrieval can be integrated with hypertext and database management systems in a single framework. The success of the INQUERY system in ARPA-sponsored evaluations has led to a number of technology transfer projects.

**14. SUBJECT TERMS**
text retrieval, probabilistic retrieval model, Bayesian network, text-based information system, indexing

**15. NUMBER OF PAGES**
7

**16. PRICE CODE**

**17. SECURITY CLASSIFICATION OF REPORT**
UNCLASSIFIED

**18. SECURITY CLASSIFICATION OF THIS PAGE**
UNCLASSIFIED

**19. SECURITY CLASSIFICATION OF ABSTRACT**
UNCLASSIFIED

**20. LIMITATION OF ABSTRACT**
SAR

# 1 Overview of Accomplishments

The proposal for this project described basic research and development in the area of text-based information systems. Specifically, a new probabilistic model of text retrieval using Bayesian inference nets was discussed, and previous work using the model was summarized.

Four new areas of research were proposed:

1. Development of extensions to the inference net retrieval model.
2. Development of efficient algorithms for building and searching the networks.
3. Experiments comparing the performance of different forms of the network.
4. Experiments with improved text representations produced by natural language processing.

In the two years of the project, considerable progress has been made in each of these areas, resulting in a number of publications and the development of an effective and efficient text retrieval engine called INQUERY. INQUERY has been distributed to universities, research laboratories, and the government, and is also the basis of ongoing technology transfer projects. Experiments on a wide range of full-text databases up to 3 GBytes have been carried out. In order to describe the research that has been done, the following pages contain a collection of abstracts from the major papers produced during the period of the contract.

- H. Turtle and W.B. Croft, "Evaluation of an Inference Network-Based Retrieval Model", *ACM Transactions on Information Systems*, 9(3), 187-222, (1991).

  The use of inference networks to support document retrieval is introduced. A network-based retrieval model is described and compared to conventional probabilistic and Boolean models. The performance of a retrieval system based on the inference network model is evaluated and compared to performance with conventional retrieval models.

- H. Turtle and W.B. Croft, "A Comparison of Retrieval Models", *Computer Journal*, 35(3), 279-290, (1992).

  Many retrieval models have been proposed as the basis of text retrieval systems. The three main classes that have been investigated are the exact-match, vector space, and probabilistic models. The retrieval effectiveness of strategies based on these models has been evaluated experimentally, but there has been little in the way of comparison in terms of their formal properties. In this paper we introduce a recent form of the probabilistic model based on inference networks, and show how the vector space and exact-match models can be described in this framework. Differences between these models can be explained as differences in the estimation of probabilities, both in the initial search and during relevance feedback.

- W.B. Croft and H. Turtle, "Text Retrieval and Inference", in *Text-Based Intelligent Systems*, Paul Jacobs (ed.), Lawrence Erlbaum, New Jersey, 127-156, (1992).

  The basic processes in a text retrieval system are text representation, representation of a user's information need, and comparison of these two representations. These processes are complementary, and improving the effectiveness of text retrieval will involve improving them all. Retrieval models provide the theoretical frameworks for integrating research in these areas. In this paper, we give an overview of the basic text retrieval models and then describe a recent model that is based on probabilistic inference. This model has been tested successfully in a variety of retrieval environments and can potentially make effective use of complex text representations produced by natural language processing techniques.

- R. Krovetz and W.B. Croft, "Lexical Ambiguity and Information Retrieval", *ACM Transactions on Information Systems*, 10(2), 115-141, (1992).

  Lexical ambiguity is a pervasive problem in natural language processing. However, little quantitative information is available about the extent of the problem or about the impact that it has on information retrieval systems. We report on an analysis of lexical ambiguity in information retrieval test collections and on experiments to determine the utility of word meanings for separating relevant from nonrelevant documents. The experiments show that there is considerable ambiguity even in a specialized database. Word senses provide a significant separation between relevant and nonrelevant documents, but several factors contribute to determining whether disambiguation will make an improvement in performance.

- N.J. Belkin and W.B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?", *Communications of the ACM*, 35(12), 29-38, (1992).

  Overview of approaches to information retrieval and filtering. Describes the essential similarities between these tasks and indicates some potential differences.

- W.B. Croft, "Knowledge-Based and Statistical Approaches to Text Retrieval", *IEEE Expert*, 8(2), 8-12, (1993).

  Overview of major approaches to text retrieval. Summary of the ways in which knowledge-based approaches may improve retrieval performance.

- W.B. Croft and H. Turtle, "Retrieval Strategies for Hypertext", *Information Processing and Management*, 29(3), 313-324 (1993).

  The links in a hypertext database distinguish it from a typical document database used in information retrieval applications. We show how these links can be incorporated into a probabilistic retrieval model based on inference nets. Retrieval experiments using a strategy based on this model show that it produces results that are at least as good as a spreading activation strategy. The results also show that citations can be a better source of links for a hypertext database than nearest neighbors, even though the latter appear to connect related documents. For this test collection, the overall improvement obtained by including hypertext links in the retrieval model was low. The hypertext model may, however, have additional benefits in multimedia environments.

- H.Turtle and W.B. Croft, "Efficient probabilistic inference for text retrieval", Proceedings of RIAO 3, 644-661, (1991).

  Probabilistic inference techniques have been shown to significantly improve retrieval performance when compared to conventional retrieval models, but their use can be prohibitively expensive for large collections. We give an overview of a retrieval model that is based on probabilistic inference networks and describe simplifications that allow us to build and evaluate networks efficiently, even with very large collections.

- W.B. Croft, H. Turtle, D. Lewis, "The Use of Phrases and Structured Queries in Information Retrieval", Proceedings of SIGIR 91, 32-45, (1991).

Both phrases and Boolean queries have a long history in information retrieval, particularly in commercial systems. In previous work, Boolean queries have been used as a source of phrases for a statistical retrieval model. This work, like the majority of research on phrases, resulted in little improvement in retrieval effectiveness. In this paper, we describe an approach where phrases identified in natural language queries are used to build structured queries for a probabilistic retrieval model. Our results show that using phrases in this way can improve performance, and that phrases that are automatically extracted from a natural language query perform nearly as well as manually selected phrases.

- W.B. Croft, H. Turtle, "Retrieval of Complex Objects", Proceedings of EDBT 92, 217-229, (1992).

  Many databases consist of large collections of "objects" that have complex structure and contain a wide variety of data such as text, numbers, and images. Current database systems can represent objects such as these only with difficulty, and often restrict the type of data that can be stored. In response to these shortcomings, object-oriented database systems have been designed specifically to represent complex objects and accommodate user-defined extensions such as new data types. One of the most important functions that a database system provides is to help users find data with particular characteristics. In object-oriented database systems, this content-based retrieval capability is typically limited to selection from a set of objects based on Boolean combinations of simple predicates. In this paper, we describe a retrieval model based on probabilistic inference that appears to provide the basis of a general retrieval model for complex objects. In particular, it can describe how the meanings of objects are related, including objects in composite object hierarchies, objects referred to by other objects, and multimedia objects.

- W.B. Croft, L. Smith, H. Turtle, "A Loosely Coupled Integration of a Text Retrieval System and an Object-Oriented Database System", Proceedings of SIGIR 92, 223-232, (1992).

  Document management systems are needed for many business applications. This type of system would combine the functionality of a database system, (for describing, storing and maintaining documents with complex structure and relationships) with a text retrieval system (for effective retrieval based on full text). The retrieval model for a document management system is complicated by the variety and complexity of the objects that are represented. In this paper, we describe an approach to complex object retrieval using a probabilistic inference net model, and an implementation of this approach using a loose coupling of an object-oriented database system (IRIS) and a text retrieval system based on inference nets (INQUERY). The resulting system is used to store long, structured documents and can retrieve document components (sections, figures, etc.) based on their text contents or the contents of related components. The lessons learned from the implementation are discussed.

- J.P. Callan, W.B. Croft, S.M. Harding, "The INQUERY Retrieval System", Proceedings of the 3rd International Conference on Database and Expert Systems Applications, 78-83, (1992).

  As larger and more heterogeneous text databases become available, information retrieval research will depend on the development of powerful, efficient and flexible retrieval engines. In this paper, we describe a retrieval system (INQUERY) that is based on a probabilistic retrieval model and provides support for sophisticated indexing and complex query formulation. INQUERY has been used successfully with databases of over 1,000,000 documents.

- E. Brown, J. Callan, W.B. Croft and E. Moss, "Supporting Full-Text Information Retrieval with a Persistent Object Store", EDBT 94, (to appear).

  Full-text information retrieval systems have unusual and challenging data management requirements. Attempts have been made to satisfy these requirements using traditional (e.g. relational) database management systems. Those attempts, however, have produced rather discouraging results. Instead, information retrieval systems typically use custom data management facilities that require significant development effort and usually do not provide all the services available from a standard database mangement system. Advanced data management systems, such as object-oriented database management systems amd persistent object stores, offer a reasonable alternative. We have taken an existing information retrieval system (INQUERY) and substituted a persistent object store (Mneme) for the portion of the custom data management system that manages an inverted file index. The result is an improvement in performance. We describe our implementation, present performance results on a variety of document collections, and discuss the advantages of using a persistent object store to support information retrieval.

## 2 Personnel

- W. Bruce Croft, Principal Investigator.
- Eric Brown, Research Assistant.
- Robert Krovetz, Research Assistant.

## 3 Advisory Activities

- NASA Lewis, April 1993, Advising on architecture for an image retrieval system.
- DoD Medical, April 1993, Washington D.C., Advising on applications of text analysis and retrieval in medical systems.
- Department of Commerce, May 1993, Advising on text retrieval techniques for National Trade Database.

## 4 Other Information

The research supported by this project was the catalyst for two other significant efforts at the University of Massachusetts. One is the participation in the ARPA TIPSTER program, which is the largest project ever undertaken to evaluate text retrieval and routing algorithms. U.Mass. was one of three sites chosen for this work and we have had considerable interaction with the intelligence agancies as a result. The other new effort is the NSF Center for Intelligent Information Retrieval, which started in September 1992. This center, which conducts basic research and technology transfer, is a collaboration between the Federal Government, the State of Massachusetts, industrial partners, and the university.