Scientific and Technical Report

Sponsored by
Advanced Research Projects Agency/ITO
and United States Patent and Trademark Office

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

ARPA Order No. D570

Issued by EXC/AXS under Contract #F19628-95-C-0235

Date Submitted: April 14, 1999

Period of Report: January 1, 1999 to March 31, 1999

Submitted by: Professor W. Bruce Croft, Principal Investigator
Computer Science Department
University of Massachusetts, Amherst

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Distribution Statement A: Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
04/13/99

3. REPORT TYPE AND DATES COVERED
Scientific/Tech 01/01/99 - 03/31/99

4. TITLE AND SUBTITLE
Browsing, Discovery, and Search in Large Distributed Databases of Complex and Scanned Documents

5. FUNDING NUMBERS
F19628-95-C-0235
ARPA Order No. D570

6. AUTHOR(S)
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
TR5281810499

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Mr. Harry Koch
ESC/AXS
Bldg 1704. Room 114
5 Eglin St.
Hanscom AFB, MA 01731-2116

Ms. Monique Dillon
Office of Naval Research
Boston Regional Office
495 Summer St., Room 103
Boston, MA 02210-2109

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project aims to integrate powerful, new techniques for interactive browsing, discovery, and retrieval in very large, distributed databases of complex and scanned documents. Emphasis is placed on going beyond full-text retrieval techniques developed in the DARPA TIPSTER program to support different types of access and non-textual content. These techniques should be particularly relevant to the patent domain where it is important to find relationships between documents and where the patent or trademark may be based on a visual design. The specific tasks identified involve studying representation techniques for long documents with complex structure, browsing and discovery techniques for large text databases, image retrieval and scanned document retrieval techniques, and architectures for large, distributed databases.

14. SUBJECT TERMS

Browsing	Query Processing	Indexing
Image Retrieval	Scanned Document Retrieval	Bayesian Network
Text Retrieval	Probabilistic Retrieval Model	Large Distributed Databases

15. NUMBER OF PAGES
9

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents

Technical and Scientific Report

Task 1: Representation Techniques for Complex Documents

Task Objectives

In this task, the goal is to extend the word-based representations that are common in retrieval systems in order to support summarization, browsing, and more effective retrieval. Specifically, we have been studying phrase-based representations and relationships between phrases in individual and groups of documents as the basis for our approach. Document structure will be used as part of the information that is used to "tag" the phrasal representation.

Technical Problems

The technical problems have to do with defining a "phrase", developing techniques for rapidly extracting them from text, comparing phrase contexts to identify significant relationships, producing summaries from these representations, extending the underlying retrieval model to be able to make effective use of phrasal representations, and using complex document structure in indexing and retrieval.

General Methodology

The general methodology for this task is to demonstrate effectiveness through user-based and collection-based experiments. As well as the PTO text databases, we will make extensive use of the TIPSTER document collection, which consists of a large number of text documents from a variety of sources, queries, and user relevance judgments for each query.

Technical Results

Some improvements to the automatic phrase extractor were made and further experiments were carried out. The results of these experiments have been used as the basis for a paper describing the technique. We have also improved the implementation of the phrase handling in the demonstration retrieval system.

Another set of experiments we have started are attempting to improve our understanding of how phrases contribute to retrieval. We have been analyzing the statistics of phrases in relevant documents samples collected from TREC in order to determine better weighting techniques and improve retrieval performance. We have also been doing experiments with a new retrieval approach based on language models that may also yield retrieval improvements for PTO data.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We will continue to improve the retrieval performance of the demonstration system through better use of phrases.

Task 2: Browsing and Classification Techniques for Document Collections

Task Objectives

The goals of this task are to develop techniques for summarizing and classifying collections of documents. These techniques will be designed to support interactive browsing and text classification in environments like the PTO.

Technical Problems

The technical problems involve producing an effective summary of a group of documents, such as a retrieved set or an entire database. Both document and phrase clusters could be used as part of this process. The classification task emphasizes the ability to accurately assign predefined categories (as in the PTO classification) to new documents (patents). An additional problem is to determine when existing classifications do not match well to new documents, such as when a PTO category covers too many patents and needs to be refined.

General Methodology

Evaluation of these techniques will be done using both the TREC corpus and PTO data. For the classification task in particular, we are designing evaluation criteria with substantial input from PTO staff.

Technical Results

In the classification work of the last 3 months, we have continued to do classification experiments based on the patent class hierarchy. This work is described in a new paper.

In the summarization/visualization area, we have implemented a prototype of a system that summarizes using a concept hierarchy. Experiments were carried out that tested whether the relationships found were meaningful. This work was reported in a paper written for the SIGIR conference.

Important Findings and Conclusions

None.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We will continue to improve the demonstration system and plan to carry out further classification experiments.

Task 3: Image Indexing and Retrieval

Task Objectives

The goal of this task is to develop similarity-based techniques for retrieving images such as trademarks, logos, and designs.

Technical Problems

The central issue is how images can be indexed to support efficient, content-based retrieval. The primary type of query in these environments is “find me things that look like this”. We are developing "appearance-based" retrieval of images as well as more straightforward features such as color and texture. Filter based and frequency domain based techniques offer some potential in this area, but significant work needs to be done on making this approach efficient enough to deal with hundreds of thousands of images.

General Methodology

The evaluation of these techniques will be done in a similar way to text by developing test collections of images. Specifically, we are working to obtain large collections of trademark and design images, both from the PTO and from general sources such as the web.

Technical Results

We have incorporated a relevance feedback technique into the trademark retrieval system. This technique allows the searcher to indicate which trademarks are examples of the types of images that are relevant. Based on these examples, the system updates the original query and produces a new ranking. We have begun experiments to test the effectiveness of this technique. We have also incorporated the ability to directly input a query image for the demonstration system. We continue to carry out experiments on appearance-based techniques that combine local and global features. We have also evaluated a technique that may be able to be used to separate text from design in mixed trademarks. The technique was evaluated by segmenting words in handwritten manuscripts.

Important Findings and Conclusions

Relevance feedback is a viable technique for trademark retrieval. Widely varying images of text can be accurately segmented.

Significant Hardware Development

None

Special Comments

None.

Implications for Further Research

We continue to focus on evaluating the accuracy of our techniques using trademark testbeds from Britain and, hopefully, from the U.S. PTO. We will also continue to improve the demonstration system.

Task 4: Distributed Retrieval Architecture

Task Objectives

The goals of this task are to scale up our current methods of automatically selecting collections and merging results, and to investigate architectures that can support efficient retrieval, browsing and relevance feedback in distributed environments with terabytes of information.

Technical Problems

The current INQUERY text retrieval system uses a client server architecture to support simultaneous retrieval from multiple collections distributed across one or more processors. A number of efficiency bottlenecks develop, however, when the size of the databases is very large. Deciding which subcollections to search can address part of the problem, but there are other problems associated with the fundamental efficiency of the processes involved and the use of distributed resources. Image indexing and retrieval tends to make all of these problems worse since the databases and indexes are considerably larger.

General Methodology

The architectures and algorithms produced in this task will be evaluated using a combination of standard performance (efficiency) measures and effectiveness measures. The efficiency tests will be done using TREC data and large PTO databases, including images, and the collection selection algorithms will be evaluated using the text subcollections of the patents.

Technical Results

A new technique for collection selection in distributed retrieval was evaluated and described in a paper. This technique is particularly appropriate for an environment like the PTO that has very large databases and control over how those databases are partitioned. The technique creates language or topic models through clusters and bases the partition on those models. The retrieval results show that it is even possible to outperform centralized retrieval using this technique. Partitioning the databases by patent classes is similar to this technique and we have begun discussing the implication of these results with Dataware.

We have also carried out experiments and described a technique for data replication for large distributed databases that results in performance improvements. We are currently considering how these two new results can be integrated into the PTO environment.

Important Findings and Conclusions

Distributed search can be more effective than centralized search if it is based on language models. Replication can significantly improve the performance (response time) of a large distributed system.

Significant Hardware Development

None.

Special Comments

None

Implications for Further Research

We will continue to evaluate performance of distributed architectures for scalable IR using the new demonstration system.