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: October 8, 1998

Period of Report: July 1, 1998 to September 30, 1998

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	10/08/98	Scientific/Tech

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
TR5281811098

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
12

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 will be 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

We have significantly improved the recall of the statistical phrase recognizer. Some minor problems remain but this approach is now looking more promising.This approach uses an HMM phrase "segementer" that is trained on large collections of phrases obtained through simply counting co-occurrences. We have observed three kinds of problems in our previous experiments:

1. incomplete phrases, e.g.

James H. Rosenfield ==> James H.

Seagram Classics Wine ==> Classics Wine
executive vice president ==> executive vice

2. missed high frequency phrases, e.g.

law firm
half price
common stock
life insurance

3. bad phrases, e.g.

plan involving
plan designed
John Blair represents
officials of
& Estate
of sales

To address these problems, we first fixed the basic tokenization. As a result, the performance of the segmenter improved, with recall of 65% and precision of 94%. Further analysis revealed the following remaining problems:

1. problems caused by the training, e.g.

James H. Rosenfield ==> James H.
there is no entry for 'James H. Rosenfield' in the
training dictionary, because of its low frequency (2);

New York law firm ==> New York
the missed 'law firm' was caused by a low
transition probability for 'law' occurring
at third and 'firm' occuring at fourth, or
'law' occurring at first and 'firm' occuring
at second,

life insurance ==>
the missed 'life insurance' was caused by a low initial
probability to start a phrase with 'life' as the first word;

2. problems caused by smoothing, e.g.

'officials of' and '& Estate'

since the smoothing routine used an average
probability for the zero probabilities,
a word never occurred at a certain position
of a phrase, thus, the 'of' and '&' that
never occurred at the beginning or end of any phrase
were smoothed with the average probability of this occurring.

To solve these problems, we took the following steps:

1. Full-matching training
Using a new algorithm, full matching, for the training.
the full matching will count every phrase found from a
tokenized text segment (used to count the longest one),
e.g. for the tokenized segment,
"New York County grand jury indicted" we used to
count only "New York County" and "grand jury",
but the full matching training will count
==> New York
==> New York County
==> York County
==> grand jury

thus, those phrases like 'law firm' and 'grand jury'
are included in the training more often. Using retrained
tables with a small collection of news articles, the segmenter
performed much better, with recall 76%, and precision 93%.

2. Ending-state
Adding the ending states to the training, so that the
trained word table will have a ending state for every word
and the transition table will have 14 more transitions.
Using such trained tables the segmenter
performed much better, the recall became 86% while the
precision became 93%. The segmenter cannot use the
ending-state to terminate a phrase because there are
many words which have a zero ending-state, so that
bad phrases, like 'officials of', are still possible.

3. Heuristic fixing
In order to fix incomplete phrases, some heuristics were
Added to deal with proper names and low frequency combinations.
With these heuristics, the recall became 89% and the precision 97%.

4. Smoothing parameter and transition biases
   Using a command line parameter to reduce the smoothing value (average probability), by the experiments the smaller value showed better performance.

   There are another two command line parameters, a bias for the initial probability to start a phrase which can help to get more phrases (but decreasing precision), and a bias for the transition probabilities which can help to get longer phrases. After adjusting the parameters experimentally, the recall was 91% and precision 98%.

There are still some problems from the training dictionary, so improving the quality of this dictionary should further improve the segmenter's performance.

As a result of a visit to the PTO in May, we have begun work on a demonstration system for patent search and classification. The aim of this system is to integrate many of the research results we have obtained in these areas. In terms of representation, we are making use of phrases and fields in the patents to improve performance. We have also begun to develop a more efficient, class-specific phrase display tool for this demonstration.. The idea of this tool is to allow examiners to look at phrases from specific classes that are related to query words and phrases.

Important Findings and Conclusions

The statistical phrase recognition approach is getting much improved results. Both recall and precision are now better than 90%, whereas previously recall was around 50% or less.

The demonstration system for patent search and classification will be an important tool for studying the impact of the new representation techniques.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

With regard to the phrase segmenter/recognizer, the following steps remain:

1. regenerating the phrase dictionary
2. training the table with larger data set
3. using the binary format for tables to speed up loading
4. make another testbed out of the training data
5. more experiments with bias, try to figure out their relationship with the training data

We plan to put most of our other effort into the demonstration system.

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

More classification experiments were performed on the new data from the speech patent classes. This data contains all the documents from 1985-1997 in the roughly 100 subclasses under the "speech signal processing" node in the PTO hierarchy. We have been training classifiers to place documents in the 30 or so subclasses fall under the

"speech recognition" node, using the others as closely related negative examples. The years 1985-1995 has been used as training data and 1996-1997 as test data.

The focus of this work has been to study if the hierarchical structure of the patent classes could be used to improve the document placement accuracy. After considerable experimentation, a technique was discovered for improved classification accuracy using the hierarchical structure of the patent classes. These results are described in:

(1998), Larkey, L., "Some Issues in the Automatic Classification of US Patents," Papers from the 1998 Workshop, AAAI Press, Technical Report WS-98-05, pp 87-90.

The techniques that have been developed for classification are now being incorporated into the demonstration system.

A paper on the visualization of search results was presented.

(1998), Leouski, A. and Allan, J., "Evaluating a Visual Navigation System for a Digital Library," Proceedings of the Second European Conference on Research and Technology for Digital Libraries, 21-23 September, Heraklion, Crete, Greece.

Important Findings and Conclusions

We have shown that classification accuracy can be improved using the class hierarchy.

We are integrating the classification techniques into the new demonstration system.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We have begun collaboration with another group to test alternative techniques for hierarchical classification. We are now devoting most of our time to producing the new demonstration.

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 are now focusing on integrating our latest image matching techniques into a new demonstration system for multimodal (text and image) searching of the 650,000 trademark database. This involves significant effort in developing the indexing techniques for large databases. We are also talking to trademark searchers to determine how to more effectively support this type of search.

We continue to refine our color matching techniques for flower images. The project has been hampered by a lack of images from the PTO.

Important Findings and Conclusions

The new trademark retrieval demonstration will be a valuable testbed for understanding how to effectively support trademark searching.

Significant Hardware Development

None

Special Comments

The progress of this part of the project depends on data from the PTO. Specifically, we still need more flower patents.

Implication for Further Research

We will continue to evaluate the accuracy of trademark and flower retrieval, with a focus on delivering a new 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

We have continued to develop distributed search techniques and have recently shown that an approach based on language models can considerably improve performance. To demonstrate this approach, we are developing a system with 60 gigabytes of patent data split into their patent classes. This creates about 400 subcollections. The demonstration will show how a small subset of these collections can be selected for a given query and still give excellent retrieval performance.

Important Findings and Conclusions

The demonstration system will be used as a testbed to evaluate algorithms for distributed search.

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.

UNIVERSITY OF MASSACHUSETTS
AMHERST

Lederle Graduate Research Center
Box 34610
Amherst, MA 01003-4610
(413) 545-2744

Computer Science

DATE: October 8, 1998

TO: Defense Technical Information Center (DTIC)

FROM: W. Bruce Croft, Principal Investigator

SUBJECT: Quarterly Scientific and Technical Report for F19628-95-C-0235

Enclosed is your required number of copies of the quarterly R&D Status Report and Scientific and Technical Report for ARPA order number number D570 (note: changed from old AO #D468) issued by ESC/ENS under contract number F19628-95-C-0235. The title of the project is "Browsing, Discovery and Search in Large Distributed Databases of Complex and Scanned Documents." These reports are being distributed in the appropriate amounts to ESC/AXS, ESC/ENK, ARPA/ITO, DTIC, and ARPA/Technical Library.

I have also enclosed a copy of the slides from the December meeting.

If you have any questions, I can be reached by email at croft@cs.umass.edu.