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: January 14, 1999

Period of Report: October 1, 1998 to December 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
	01/14/99	Scientific/Tech 10/01/98 - 12/31/98

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
TR5281810199

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

In the last three month period, we have worked on implementing a phrase discovery module called Phrase Works for the Patent retrieval demonstration. This enhancement was suggested by the PTO and discussed at the last meeting.

Phrase Works helps patent searchers to add phrases to a query. The steps done to accomplish this are as follows:

1. Create a phrase dictionary for each patent class (400 classes).
No global dictionary was used because we were concerned about running time and memory space when handling 60GB of data. To extract phrases from the text, we used the heuristic method developed earlier in this project combined with a post-repair stage. The post-repair stage

uses WordNet to get the part-of-speech tag for every word in a phrase and invokes a set of disambiguating rules to decide if a word should be removed or kept in the phrase.

2. Generate a phrase file for each class.
A phrase lookup program lists all phrases found in the dictionary and also those single terms occurring in the phrases, so that the resulting file looks like:
<docid> <field name>
<phrases>+
<single term>+

3. Simple stemming of phrase files.
The steps involved here are:
a. Combine upper and lower case words.
b. Combine hyphenated and non-hyphenated forms (keep the hyphen if the text has only the hyphenated form).e.g. both 'cross-shaped' and 'cross shaped' are found in the class then change 'cross-shaped' to 'cross shaped'.
c. Combine singular and plural forms (keep the plural if there no singular form exists in the class). e.g. both 'device' and 'devices' are found in the class then change 'devices' to 'device'.

4. Generate a phrase collection for each class.
We generated a document for each term or phrase in the stemmed phrase file, with the term or phrase in the title field, the co-occurrence data (the list of co-occurrence terms or phrases with their statistics) in the text body, and the frequency of each term or phrase in the source field. These documents look like:
<DOC>
<TITLE> <term or phrase> </TITLE>
<SRC> <frequency> </SRC>
<TEXT> <co-occurence data>* </TEXT>
</DOC>
Note: the text body maybe empty if there is no co-occurrence data for a term or phrase.

5. Build the phrase database for each class.
This was done using the INBUILD indexing module from U.Mass.

6. Create an API function to access modified and co-occurring phrases.
To find modified phrases a query (a term or phrase) is processed for the phrase collection. These phrases will contain all query term(s). e.g. if the query phrase is 'vision system', the API will return:
night vision system
aviators night vision imaging system
night vision imaging system
Co-occurring phrases can be found in a similar way.

The API function takes six optional parameters for listing modified or co-occurring phrases,

showing frequencies or statistics, sorting the returned list, and cutoff thresholds. The API function also does simple stemming and will replace query words with stemmed words when the returned list is empty.

Important Findings and Conclusions

The phrase extractor based on statistical techniques continues to be improved. Its performance is now equal to our best heuristic approaches.

Significant Hardware Development

None

Special Comments

None.

Implication for Further Research

We will continue to improve the demonstration system and the statistical phrase extractor.

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 focused on experiments testing the performance of automatic classification that exploits the patent class hierarchy. We have focussed on the hierarchy of subclasses under 2.09, speech signal processing, which provides a very difficult testbed for automatic classification. Most of the recent work has been on using the hierarchical relations among subclasses to improve the performance of Bayesian classifiers for the subclasses.

We have found small but consistent improvements in classification performance when we take advantage of the hierarchical relations. Performance is better when classification takes into account the results of classifying a document into parent subclasses of the subclass of interest. We are currently working on a paper to describe these experiments.

We have also returned to the experimenting with ways to take into account the relative frequency of occurrence of a patent subclass in the k-nearest algorithm. Currently, the Bayesian classifiers take this into account, but the nearest neigbor classification algorithm does not.

We have devoted considerable effort to including classification in the new Patent search demonstration (described later).

Visualization research is continuing and we expect to produce new papers for the next report.

Important Findings and Conclusions

We have shown that classification accuracy can be improved using the class hierarchy. We have also integrated the classification techniques into the new demonstration system.

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 spent much of our time developing a new trademark retrieval system that incorporates image search against many trademarks with text search against more than 1 million trademarks. The demonstration has a new interface and a number of other features.

The private URL for the demonstration is http://darwin.cs.umass.edu/tmk.

We are also continuing to improve the image search and are incorporating these changes into the demonstration system. We have acquired data from the British Trademark Office and Professor John Eakins of the University of Northumbria that will be used to start evaluating the effectiveness of trademark image search.

Important Findings and Conclusions

The new trademark retrieval demonstration appears to be a useful tool for verifying research in this area.

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.

Implications for Further Research

We will place more emphasis 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.

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 developed and evaluated new techniques for distributed search based on language models and query probing. We are working on papers describing the results.

We have devoted considerable effort to the new Patent search system, which uses the distributed search architecture and is described below. Figure 1 shows the basic architecture of the retrieval system.

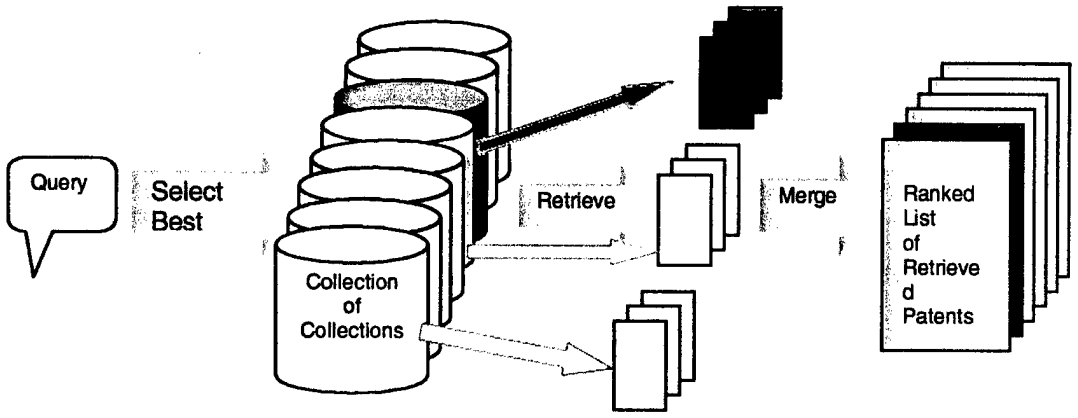

Figure 1: Distributed Retrieval Architecture

TEXT SEARCH AND CLASSIFICATION DEMONSTRATION

Private URL is http://darwin.cs.umass.edu/pto

Major new features:

1. New interface
Improved look and feel.
Includes easy input of fielded queries - user does not need to use INQUERY syntax to get fields and Boolean operators. Can enter search terms into a field on screen, and choose "desire", "require" or "reject" for the content of that field.
Provides three alternative input screens:
Standard Query
Advanced Query
Patent Number
Phrase help: Helps users find good phrases to add to their query

2. Phrase help

User can enter a phrase and ask for additional phrases containing or associated with their phrase.

The interface makes it easy for the user to select any of the returned phrases and add them to their main query. This is done on a class-specific basis. User can either type in the class, or let the system choose the best class for their phrase. (the choice is made via distributed collection selection, using the phrase as query).

We have prebuilt an INQUERY database for each class that is used to find these phrases. The title of each doc is a phrase or word extracted from the text of the patents in that class. Phrases were found using WordNet. Each doc includes several fields, including frequency information, a list of phrases associated with the title phrase, and co-occurrence scores for the phrases. The same phrases database is used to find containing phrases and co-occurring phrases. Figure 2 shows the query expansion process.

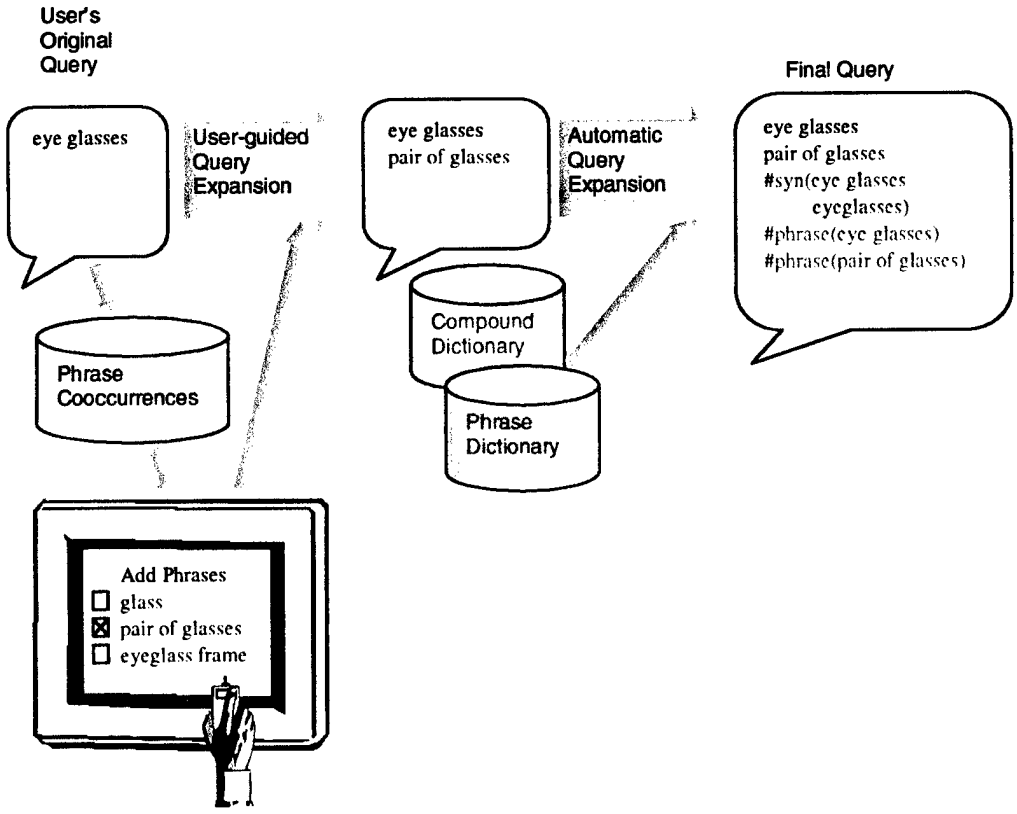

Figure 2: Overview of query expansion process.

3. Collection Selection.

The size of the collection has been increased from about 9 gig of raw data from 2 years (1995 and 1996) to about 55 gig of raw data covering utility and plant patents from 1980 through 1996.

It accesses multiple databases (400, one per class), and uses distributed collection selection to choose the best 10 collections for a query.

We are in the process of adding fields to the collection selection database so that we can do a better job of choosing good collections for queries involving with fields.

Our experience with this collection of collections has given us a better understanding of the differences between collections like this that are divided up by topic, and the more usual situation where collections are divided up by source. We are doing research on the merging part of collection selection to better handle this case of collections divided up by area.

4. Classification

The demo system uses a k-nearest neighbor algorithm to classify text.

5. Other enhancements

Previously, you could type in a query and search for patents, or select a patent and try to classify the text in it.

Now you can enter text in any of three ways:

1. Type in a query
2. Take text from a file (with browsing for the file)
3. Get text from a patent in the collection

and submit the text as a query, or classify it.

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.