A SOPHISTICATED DOCUMENT RETRIEVAL SYSTEM

W. Bruce Croft
Lynn Ruggles

COINS Technical Report 81-30

December 1981

This research was supported in part by NSF grant IST-8011605.

Abstract

The significant advances made in theoretical and experimental research in information retrieval have many implications for system design. One possible design for a document retrieval system based on these advances is presented. A major part of this system design has been implemented as a bibliography filing and retrieval system for the Computer Science department at the University of Massachusetts. The implementation issues considered here are functionality, user interface and file organization. The main point of this implementation was to demonstrate that an efficient, effective and flexible system can be constructed using modern techniques.

1.0 INTRODUCTION

The techniques used in commercially available document retrieval systems bear no relation to the current thinking of researchers in this area. A typical system, such as Dialog [1] uses Boolean queries and searches for documents whose index terms (assigned manually) exactly match the query. Research has demonstrated that automatically indexed documents can, when used with search strategies based on probabilistic models [2,3], give good performance and a high degree of flexibility. The reasons for the differences between research and commercial reality have been stated elsewhere [4], but a major one is the investment in enormous databases. However, even if large commercial services remain firmly entrenched in the past, there is a major thrust for effective and flexible retrieval systems coming from other areas such as office information systems [5] and systems providing tools for professionals. These areas are developing rapidly due to advances in computer and communication technology and it is important to design filing and retrieval systems which are capable of meeting the challenges posed by these new applications. In this paper, we shall summarize the major results of recent work in information retrieval and discuss their implications for system design. A system based largely on the design recommendations will then be described in some detail. This system is used to file, retrieve and display bibliographic references in the Department of Computer and Information science at the University of Massachusetts.

2.0 RESEARCH DEVELOPMENTS AND THEIR DESIGN IMPLICATIONS

Research into the design of text filing and retrieval systems can be divided into two main areas - indexing and retrieval. Indexing refers to the process of describing the content of documents using keywords or index terms. Retrieval is the process of locating relevant documents in response to users' queries. We shall discuss these two areas separately.

2.1 Indexing

Indexing in commercial systems is done manually by professional indexers. However, in the design of systems which will be able to be used in a number of environments (such as business offices), the major concern is with automatic indexing techniques. A large number of studies on indexing were carried out by Salton [6]. The results of this work and the more recent work of Sparck Jones [7] suggest that the simplest techniques are the most effective. For example, the use of phrases in addition to simple keywords does not improve the performance of the system. There is some evidence to support the use of statistically generated phrases and thesaurus classes formed using discrimination values [8], but it appears that if the retrieval techniques mentioned in the next section are used, these more complicated indexing techniques are simply not needed. Harter [9] has proposed a sophisticated model of indexing based on a model of word occurrence in text, but it appears that there may be too many parameters to estimate for this to be reliable. The indexing process then, will look like the following

1. Identify words in text.
2. Remove common words using stopword dictionary.
3. Stem words using suffix removal.
4. Replace stems with numbers from a stem dictionary and count stem frequencies in documents.
5. Remove some low frequency stems from documents (Sparck Jones advises caution with this step, but it is an excellent way to reduce the number of index terms for large documents).

It should be noted that index term weights such as the inverse document frequency (IDF) [7] are specifically excluded from the indexing process. We regard these weights as part of the retrieval process rather than indexing. The advantage of this approach is that the indexing process described above can be done for each document individually without knowing the term occurrence characteristics over the whole collection. Weights such as the IDF or the discrimination value change as the collection changes and therefore should not be stored in the document representative.

2.2 Retrieval

The research on probabilistic models of retrieval has produced many interesting theoretical and experimental results. Sparck Jones and Robertson's work on this topic [2] established its effectiveness and pointed out the importance of relevance feedback for estimating the parameters of the model. In a system using a search strategy based on this probabilistic model, the user would identify relevant documents from an initial ranking of the documents in the collection. Van Rijsbergen and Harper's work

[3,10] extended the probabilistic model to include dependencies between index terms. This formalized much of the early work on query expansion through term clustering. However, the experimental results from this model have been ambiguous and it appears that similar performance gains can be made by expanding the initial query using terms from the identified relevant documents [11,12]. Croft and Harper [13] considered retrieval before relevance feedback and showed that the probabilistic model indicates that a modified form of the inverse document frequency weight should be used to do the initial ranking. Yu, Lam and Salton [14] extended this approach to include more sophisticated estimation methods but their results do not indicate significant performance gains. Finally, Croft [15] has recently extended the probabilistic model to include within-document frequency information from the indexing process. The experimental results indicate that this extension also provides performance benefits.

To summarize then, it would appear that the most effective retrieval strategy, which also happens to be easy to implement is

1. Do an initial ranking of the documents using the extended independence model proposed in [15]. This is similar to using a combination of inverse document frequency weights and within-document frequency weights.

2. Obtain relevance judgements for the top 10 ranked documents.

3. Expand the query using a subset of the terms from the relevant documents.

4. Rerank the documents using the extended independence model with parameters estimated from the set of relevant documents.

3.0 THE UNIVERSITY OF MASSACHUSETTS BIBLIOGRAPHY SYSTEM

In this section, we describe a system implemented using many of the design recommendations made above. The major task of the system is to provide for the filing, retrieval and display of sets of references to documents. These sets include personal bibliographies and departmental collections such as reference works and technical reports. The system records the standard bibliographic information about documents such as title, author(s), date of publication, journal, volume and page numbers as well as an abstract. The major functions of the system are as follows

File - record information for new references, including abstract.

- different types of documents have different formats.
- entry of information is by system prompt with extensive help facilities.
- automatic indexing is carried out based on the title, abstract and any keywords entered by the user.
- provision is also made for making the document a member of a particular set (e.g. one of John Smith's documents).

Retrieve - retrieve documents according to criteria specified by the user.

- the query can be specified in a number of ways

  a. natural language statement of interest.

  b. Boolean specification of fields such as document type, author, date and keywords.

  c. example documents.

  d. certain combinations of the above.

- the query is specified using a formal language designed for ease of use (for example, simple spelling correction is done and command retyping is minimized).

Display - print lists of references in a specified format.

- various standard formats provided (e.g. ACM, IEEE)
- users can specify their own formats.
- sets of references to be printed can be established using the retrieve function.

The system was designed for speed of access, ease of updating and reasonable storage overhead. As mentioned, it incorporates the indexing process in section 2.1 and uses the probabilistic model with relevance feedback. Extensions are planned to include the expansion of the query using relevant document terms. Figure 1 shows the main components of the file organization. Index terms are found in the hash table (1) where pointers to the inverted lists in table (2) are stored together with the lengths of the lists and the number of postings for the terms. The variable length inverted lists are stored in fixed length random access blocks linked by pointers. This permits efficient updating of these lists. The inverted list contains pointers to the fixed length bibliographic information stored in (3). This information can also be accessed directly using the document identifier. The bibliographic record contains a pointer to the variable length abstract stored in a number of fixed length random access blocks. Updating these files is also straightforward as bibliographic information and abstracts for new documents can be added to the end of the files.

The system is implemented on a VAX 11/780 in Pascal. It currently contains bibliographic information and abstracts for 3200 CACM documents and it is designed to handle tens of thousands of such documents. The response times are on the order of 1-5 seconds, depending on system loading. The programs take up approximately 30 kbytes of storage, but there has been no real effort to reduce their size. Figure 2 shows an example of an interaction with the system.

4.0 CONCLUSION

Text filing and retrieval systems can be designed using techniques derived from recent research in information retrieval. These systems are very effective in locating relevant documents and can be made efficient for large collections of documents. It is this type of system that will be used in the rapidly developing area of profession-based systems.

Acknowledgements

The authors wish to acknowledge the help of Diane Falconer and James Rochfort in building the system.

References

1. Lockheed Information Systems. "A brief guide to DIALOG searching", Palo Alto, California (1976).

2. Robertson, S.E.; Sparck Jones, K. "Relevance weighting of search terms", JASIS, 27:129-146; 1976.

3. Van Rijsbergen, C. J. "A theoretical basis for the use of co-occurrence data in information retrieval", J.Doc., 33:106-119; 1977.

4. Jamieson, S. H. "The economic implementation of experimental retrieval techniques on a very large scale using an intelligent terminal", Proceedings of the 2nd ACM SIGIR Conference, SIGIR Forum, 14:45-52; 1979.

5. Croft, W.B. "An overview of information systems", Information Technology : Research and Development, 1: 73-96; 1982.

6. Salton,G. Automatic Information Organization and Retrieval, McGraw-Hill, New York (1968).

7. Sparck Jones, K. ; Bates, R. G. "Research on automatic indexing 1974-1976", British Library Research Report, University of Cambridge, England (1977).

8. Salton, G. "A blueprint for automatic indexing", SIGIR Forum, 16: 22-38; 1981.

9. Harter, S.P. "A probabilistic approach to automatic keyword indexing", JASIS, 26: 197-206,280-289; 1975.

10. Harper, D.J.; Van Rijsbergen, C.J. "An evaluation of feedback in document retrieval using co-occurrence data", J.Doc, 34:189-216; 1978.

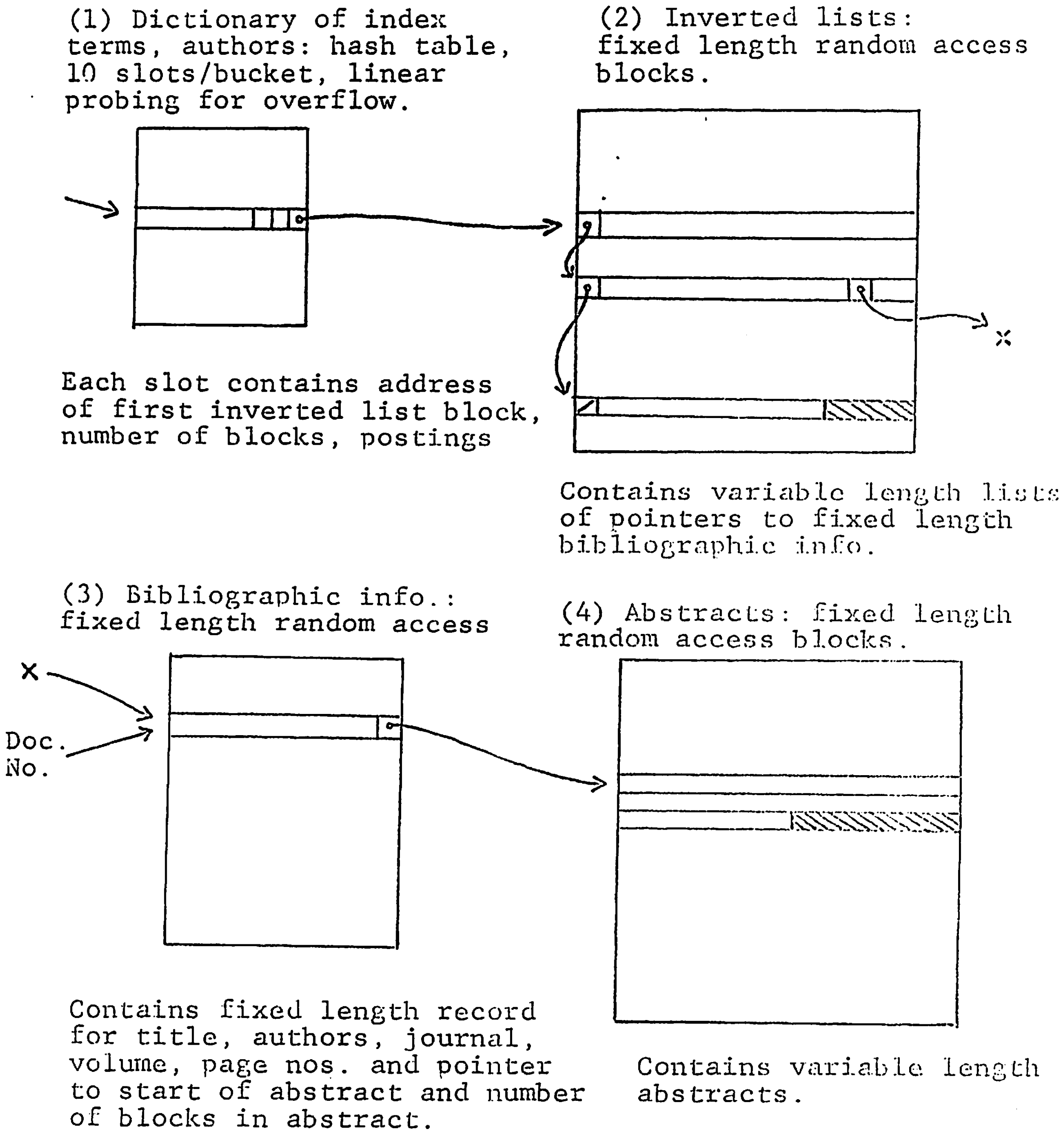
11. Sparck Jones, K; Webster, C.A., "Research on relevance weighting 1976-1979", British Library Research Report 5553, University of Cambridge (1980).

12. Harper, D.J. "Relevance feedback in document retrieval systems : An evaluation of probabilistic strategies", Ph.D. Thesis, University of Cambridge (1980).

13. Croft, W.B.; Harper, D.J. "Using probabilistic models of document retrieval without relevance information", J.Doc., 35:285-295; 1979.

14. Yu, C.T.; Lam, K.; Salton, G. "Term weighting in information retrieval using the term precision model", JACM, 29: 152-170; 1982.

15. Croft, W.B., "Document representation in probabilistic models of information retrieval", JASIS, 32: 451-457; 1981.



(1) Dictionary of index terms, authors: hash table, 10 slots/bucket, linear probing for overflow.

Each slot contains address of first inverted list block, number of blocks, postings

(2) Inverted lists: fixed length random access blocks.

Contains variable length lists of pointers to fixed length bibliographic info.

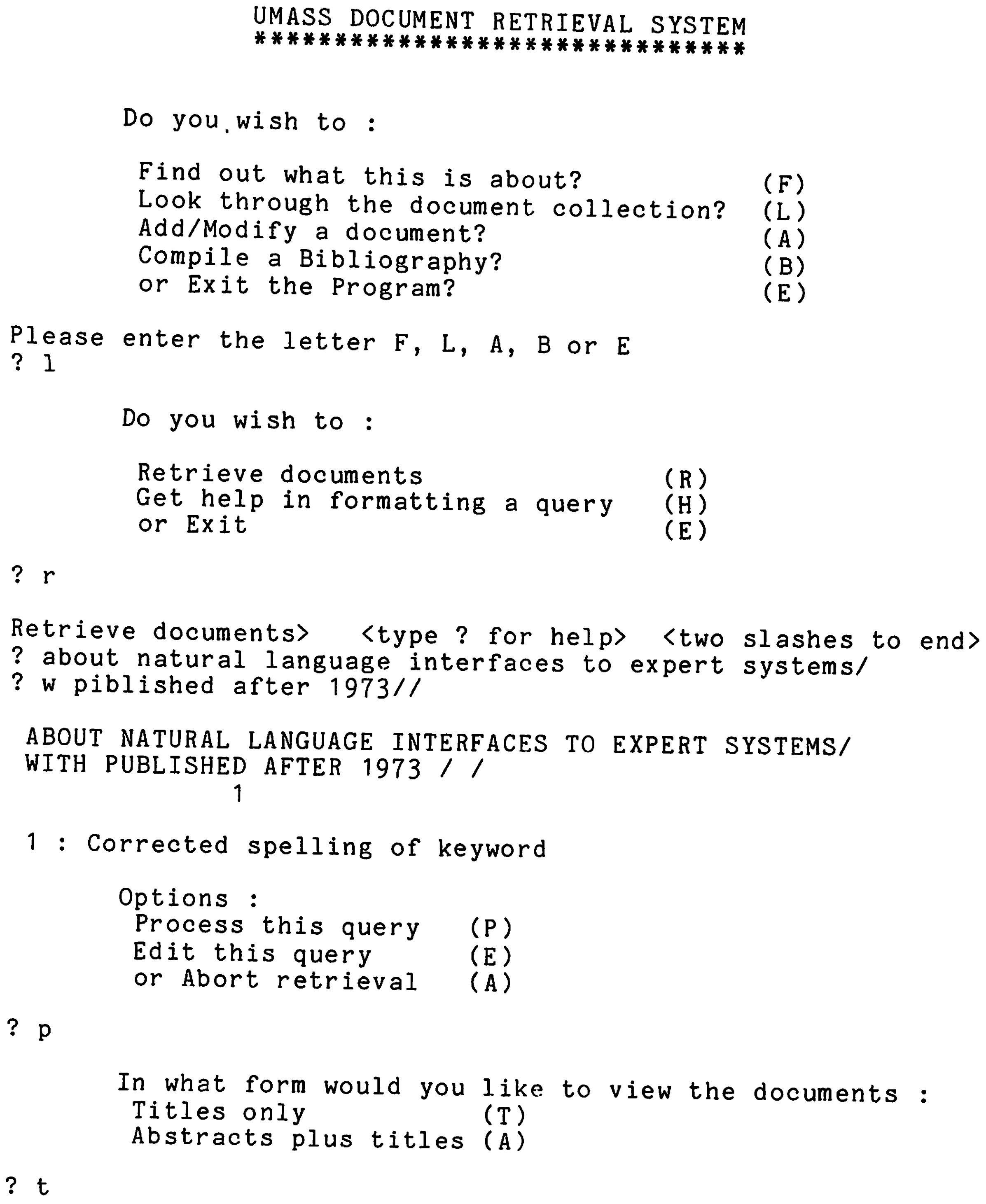
(3) Bibliographic info.: fixed length random access

Contains fixed length record for title, authors, journal, volume, page nos. and pointer to start of abstract and number of blocks in abstract.

(4) Abstracts: fixed length random access blocks.

Contains variable length abstracts.

Figure 1 : Overview of the File Organization



```
                 UMASS DOCUMENT RETRIEVAL SYSTEM
                 *******************************


        Do you wish to :

         Find out what this is about?                (F)
         Look through the document collection?       (L)
         Add/Modify a document?                      (A)
         Compile a Bibliography?                     (B)
         or Exit the Program?                        (E)

Please enter the letter F, L, A, B or E
? l

        Do you wish to :

         Retrieve documents                    (R)
         Get help in formatting a query        (H)
         or Exit                               (E)

? r

Retrieve documents>   <type ? for help>  <two slashes to end>
? about natural language interfaces to expert systems/
? w piblished after 1973//

 ABOUT NATURAL LANGUAGE INTERFACES TO EXPERT SYSTEMS/
 WITH PUBLISHED AFTER 1973 / /
             1

 1 : Corrected spelling of keyword

        Options :
         Process this query   (P)
         Edit this query      (E)
         or Abort retrieval   (A)

? p

        In what form would you like to view the documents :
         Titles only          (T)
         Abstracts plus titles (A)

? t
```

Figure 2(a).

```
   Document number: 1549
   Authors: Boguraev, B.K., Sparck Jones, K.
   Title: A Natural Language Analyser for Database Access
   Journal: Information Technology: Research and Development
   Date: 1982  Volume: 1  Pages: 23-40

       Do you want to see the abstract?
        Yes (Y)
        No  (N)

? n

       Is this document relevant?
        Yes (Y)
        No  (N)

? y

       Do you want the document
        Printed              (P)
        Saved                (S)
        Neither of the above (N)

? p
          .
          .       {User looks at top ten documents}
          .

 12 additional documents were retrieved.  {Relevance feedback}
 How many do you want to see?

? 5
          .
          .       {System displays five documents, user decides
          .        whether to print}
```

Figure 2(b) : An example of interaction with the system.