USE OF COM AND KWIC TO SUPPORT
THE SPECIAL LIBRARY CATALOG

Emil H. Levine *

INTRODUCTION

This paper discusses the development of a Computer Output Microform (COM) system to support a personnel data base and the methods used to index the data base which resulted in retrieval being superior to that on a large scale computer. Applications of this methodology to the special library catalog are then developed, specifically, the use of Key-Word-in-Context (KWIC) indexing to provide searching against non-key fields. The power of KWIC indexing, as compared with Machine Aided Indexing (MAI), will shown based on experiments on recall and relevance. Advantages of KWIC and COM as a catalog in a small (up to 20,000 titles) special library, including Chinese language, using an automated microfiche retrieval device, will be discussed.

DESCRIPTION OF THE DATA BASE

In the mid 1970’s, a major U.S. agency decided to utilize COM as a backup to a primary computerized personnel file. The decision to use COM was based on several criteria.

* Mr. Levine is the Specialiser of the Planning Staff, Office of Information Systems, Drug Enforcement Administration, US Department of Justice, Washington, DC.
Online Backup

The computerized data base supported queries of a worldwide nature. At that time, the computer was only available 9AM to midnight, Monday through Friday. The nature of the information required 24 hour, 7 day a week access. In addition, the down time of the computer and the difference in user time zones caused problems. While it was not practical to reproduce the entire data base on COM, it was possible to extract basic information and make that available. For example, the full record consisted of approximately 2,000 characters of data. However, it was possible to describe the subject of the record with such information as name, address and computer number using 132 characters, or the length of one line on a COM page at 42X. The COM data base contained about 840,000 records, which were held on 84 microfiche. Each fiche held about 1,400,000 characters. The primary offices using the data base were equipped with the Bruning (now Consolidated Micrographics, Incorporated) Model 95 automated microfiche readers. The reader cassettes hold 30 microfiche (2,240 frames at 42X). One bank of keys on the reader is used to retrieve the proper fiche from the cassette and a second bank of keys is used to select the proper row and column on the retrieved fiche. This provides a retrieval of any frame in about 10 seconds for an experienced user. The 30 fiche cassette held about 43,000,000 characters (counting spaces) and provided for random access retrieval in about 10 seconds. Significantly, the commercial state of the art computer disks at this time held about 40,000,000 characters and provided random access in about the same time. However, the Bruning cassette sold for $12.50.

Overseas Access

The agency involved had personnel assigned around the world. The COM project made information available to these
users at a fraction of the cost of telecommunications. While the
data base was out of date in a short period (it was updated about
every three months), the percentage of new or changed records
was small enough to not impact on the overall use of the data
base for most users. In addition to the extracted data base des-
cribed above, some foreign offices were given the total data base
that related to their area. That is, while the basic data base
consisted of only the 132 character extract from the 2,000
character online file, overseas office were also given the full
records that could be associated with their geographic area. In
addition to the basic record, extensive cross references were pro-
vided to those specific records. The significance of this and the
methods used to create these indexes will be discussed further.
The transfer of this methodology to the special library will be
discussed.

24 Hour Access

The use of COM provided the capability to search the data
base 24 hours a day, seven days a week. The Bruning readers were
placed in the communication centers of the agency, which were
manned around the clock.

Access to Other Data Bases

A number of other agencies maintained data bases of interest.
However, direct access was not available. COM provided two
methods of access. The agencies of interest made their computer
tapes available and these were converted to COM. More simply,
these agencies provided the data on COM when it was available to
them in that form.

Cost Savings
Use of COM and KWIC

Conversion of the data base to COM resulted in considerable savings over the cost of conventional computer generated products. In the year prior to the adoption of this system, one such product consisted of 14,000 pages of computer printout. The office which received the listing was not able to effectively use it due to its bulk. Adequate storage for this much paper was lacking. The realization that this data could have been provided on 70 microfiche was one of the major factors in adoption of the COM system.

Access to Non-Key Fields

A major advantage in the COM was realized in the indexing of the full records provided to overseas offices, previously mentioned. The computer system did not allow Boolean retrieval. Retrieval was limited to a single name field on the record. Data elements which were present in the records, and which were highly relevant to retrieval, were not keyed. It was possible to conduct such retrieval by writing off-line programs with Boolean and/or string searches, but this was time consuming on the part of the programmer, used excessive computer resources and often was non-responsive to the time requirements of the user.

The solution to this problem was creation of alphanumerical indexes to each of the fields of interest. For example, the data in the “city” field was listed in alpha order. By searching this field on the microfiche, the user could determine if any records existed in the data base related to a specific city of interest. Since these records were already extracted from the entire data base by geographic region, the subset was in fact the result of a precoordinated Boolean query. For example, a data base prepared for South America could be searched for records which contained city=La Paz. This was the equivalent of an online Boolean query where area=South America and city=La Paz.

The most significant use of non-key fields resulted when
KWIC indexing was utilized. This was particularly important when applied to the address field. The diverse ways in which an address can be expressed in various countries makes conventional computerized indexing and retrieval difficult. For example, in some countries apartment number or street number comes after the street name. The various methods of reporting such information and the various ways of searching leave open the potential for serious problems of recall. The use of KWIC on the address field provided a solution to this problem. In addition, it provided a method of searching both matronymic and patronymic of Hispanic names, using the "—" as a delimiter (see Figure 1). Another application was in searching the occupation field. Terms such as "pilot" were often buried within leading terms such as "senior pilot." An alphabetical listing would not retrieve such a record. The KWIC index did.

The accession number associated with each term in both the alphanumerics and KWIC indexes related to the cassette, fiche, row and column of the full record on which the indexed term appeared. The user could search for an address by street name or number, and if a hit occurred, he had only to enter that accession number on the keyboard in order to view the full record within 4–10 seconds. The full 2,000 character records was displayed. This was faster than computerized retrieval at 300 baud.

Maximizing Recall and Relational/Associative Retrieval

The "pulling together" of all related data through the use of the alphanumerics and KWIC indexing proved to be a significant aid in retrieval. A user searching for information about a specific address might find several other records related to that exact or a nearby address. Thus, this method of indexing provided the capability of rapidly showing relationships between records. Because many of these relationships were revealed through non-key fields, they would not have been obvious when searches were
conducted through the online computer. In fact, the use of these indexes allowed the searcher to maximize recall and retrieve related records in a relational and associative sense. Figures 1 and 2 show examples of the KWIC index and the relation of the indexes to the main record.

APPLICATIONS TO SPECIAL LIBRARIES

In the COM/KWIC combination can be used to support

With conventional storage and retrieval of personnel records, it would seem that it could support retrieval of items in a special library. The COM indexes related to computerized personnel records; these often served primarily to assist the user in selecting a physical personnel file. In the same way, the catalog often refers to a main entry which leads to an item in the collection. In recent years, many libraries have adopted the microfiche format to replace printed card catalogs. A typical example in the United States is the Prince Georges County (Maryland) Public Library, which catalogs its collection on about 120 microfiche. Each fiche contains 269 frames (15 row and 18 columns minus one index frame) and each frame contains 30 entries. This provides for 968,400 entries on the 120 microfiche. Subject, author and titles are merged alphabetically. Children from the age of 10 are noted using this catalog without difficulty. Manual readers are located throughout the libraries and the fiche are held in looseleaf binders that facilitate both retrieval and replacement of the fiche in the proper slot. This conventional microfiche system works well in the public library. However, it is anticipated that it would not work as well in the special library where retrieval (user) time is expensive and the smaller-size of the collection favors total automation. In this setting, it is considered that COM, when combined with KWIC indexing
and an automated microfiche reader such as the Bruning (Consolidated Micrographics, Inc.) Model 95 can exceed or perform as well as an automated catalog.

KWIC OF TITLES

KWIC indexing was one of the first applications of data processing to information retrieval. In 1979, the U.S. Defense Logistics Agency, Defense Documentation Center (now the Defense Technical Information Center) performed an experiment to compare retrieval using a Key-Word-Out-of-Context (KWOC) index of titles with retrieval using terms assigned by Machine-Aided Indexing (MAI). The experiment involved 635 summaries which had been index with MAI. Twenty searches were made on each data base.

Relevance and Recall Implications of KWOC

Relevance and recall studies were then made to evaluate the two retrieval methods. Relevance is expressed as the ratio of the number of documents retrieved which were considered relevant to the total number of documents retrieved. Recall is expressed as the ratio of the number of relevant documents retrieved by the system to the total numbers of relevant documents in the system. The results of this study showed that average recall was .73 for MAI and .79 for KWOC. The relevance measurements were similarly close, .62 for MAI and .60 for KWOC. This indicates that both systems performed equally well. However, the KWOC generated index is significantly cheaper to produce.

In another study, the author has anecdotal evidence that certain types of researchers prefer use of KWOC or COM indexes over the more complex computerized Boolean searching, which would provide greater relevance but less recall. This was more
recently confirmed, anecdotally, in a special library setting. In both cases, the users showed a preference to the Bruming Model 95 automated microfiche reader containing an index to the collection over an online terminal located at the same work station. This will be discussed in more detail.

CATALOGING/INDEXING IMPLICATIONS

The use of microform to replace conventional catalogs is now universally accepted. However, library traditions, MARC and other standards, heavily influence the format of these catalogs. The impact of tradition is most apparent in the formatting of some COM catalogs, in which the (totally non-standard) COM indexes are place at the top of each row on a fiche, rather than being located at the (standard) lower right frame. (This is considered an attempt to emulate the card catalog drawer!) A monograph cataloged by author, title and two subject headings would use 20 lines. These conventional systems fails to make use of the value of words in a title.

A PROPOSED SPECIAL LIBRARY CATALOG USING KWIC/COM

The following microform system is proposed for a special library of about 10,000 items which has a machine readable catalog that could be used to produce COM. The purpose of the microform system is not unlike that of the personnel system described previously:

A. Backup to the computer;
B. Increased distribution of the catalog;
C. Enhanced searching;
D. Cost savings in production, shipping and storage;
E. Integration of other library catalogs.

The system assumes that users are interested in maximum recall. That is, the researcher normally wishes to see all items related to a subject. Further, the system assumes that the time of the researcher is valuable and that search time should be support by automation. Given these assumptions, it is proposed that the collection utilize KWIC indexing against the titles, subject headings, author(s), corporate author (partial), and accession number. A typical entry might appear as follows:

CUBIC CRYSTAL PRODUCTION FOR ELECTRONIC COMPONENTS/MANUFACTURING/WRIGHT, E.; JOHNSON, A.; BELL TELEPHONE LABS, NEW JERSEY; 12345; 10/21/B/7.

This would appear as seven lines in the KWIC index, under:

CUBIC
CRYSTAL
MANUFACTURING (subject heading)
WRIGHT
JOHNSON
BELL
12345 (document number)

PRODUCTION, ELECTRONIC and COMPONENT would not be indexed in this data base due to their high frequency. They would be eliminated with other common stop words. The number 10/21/B/7 relates to the cassette, fiche, row and column on which the main record is found, and to the buttons to be pushed on the automated reader for instantaneous retrieval when the proper cassette is inserted. It will often be possible to hold both the index and the main entry COM within the same cassette.
This provides a significant capability to both browse and retrieve at the maximum rate of cognition. The elements of information provided in the index shown above will often support a user decision about retrieval of the monograph. The main entry, which may be rarely searched by the user, could contain additional information felt necessary by the library staff (such as publisher, date, city of publication and other MARC specified data elements).

The size of the microform catalog using this model is estimated as follows for 10,000 items. Eight lines of KWIC index for each item (assuming two authors, one corporate author, document number and four indexing terms from the title or subject heading). Five lines would be used for the main entry. Thus, the collection of 10,000 items could be indexed by 130,000 lines of information, or less than 15 fiche at either 42X or 48X. This is half a cassette for the Bruning Model 95. A larger collection (up to about 20,000 items) could be held in one cassette or more terms could be used in each index. The retrieval time using such a system would be significantly less than that obtained through most computer systems and requires virtually no training.

ANECDOITAL OBSERVATIONS-OPERATIONAL SYSTEMS

The author has made two observations relating to the efficacy of this approach. In both cases, the collection index was available online. In one instance a KWOC index was available in both hard copy and on microfiche. In the second, the automated catalog was backed up on microfiche using conventional subject, author and title headings (not KWIC or KWOC). In both instances the Bruning Model 95 reader was used to hold the microfiche. Users of both collections consistently chose the microfiche over the computer assisted retrieval. The reasons for this are attributed
to:

A. Ease of use of the microfiche over the computer system;
B. Speed of retrieval;
C. Matching of human cognition to the display (an entire screen of possible hits could be displayed and viewed at one time rather than scrolling);
D. No requirement for training or assistance with the microfiche;
E. Requirement for maximum recall.

In both cases, the use of the automated microfiche reader was considered essential to the success of the system. The class of users involved would not have tolerated manual readers, and required a browsing capability.

**IMPLICATIONS FOR CHINESE LANGUAGE MATERIAL**

At least two companies manufacture COM equipment that can generate Chinese characters, Information International, Inc. (III) and 3M. Chinese titles which have been converted into the Standard Telegraphic Code (STC) can be sorted by this code. Therefore, it would be possible to produce a permuted KWIC or KWOC index on the STC. (The bottom part of Figure 2 shows a similar sort on telephone numbers, permuting on area code, exchange and last four digits of the phone number.) While the author has been unable to find any specific instances of a combined KWIC/KWOC index using Chinese characters, it would appear to be within the state of the art to convert a permuted STC title index into the Chinese characters using either of the COM devices previously mentioned.
CONCLUSION

Microform is now being accepted as the medium of choice for catalogs in public libraries. Through use of specialized indexing techniques, such as KWIC or KWOC, the COM generated product can be used to enhance or supplement other cataloging medium, including the online computer. When used with automated retrieval devices, it appears to be the retrieval method of choice over both hard copy and online systems. This has particular application to special library collections and special classes of users. The relatively high cost of automated readers has not seen this technology tested in the public library. These costs are not expected to be reduced. It appears within the state of the art to produce a KWIC/KWOC COM catalog of Chinese character titles by first permuting titles which have been converted to STCs, then converting back to the Chinese characters.

References


7. Author conversation with Mr. J. Mathias, Executive Director, The Chinese-English Translation Assistance (CETA) Group, Kensington, Maryland.

Bibliography


FIGURE 1

EXAMPLE OF KWIC INDEX
USED ON A PERSONNEL FILE

1 SEARCH ON STREET NUMBER
2 SEARCH BY STREET NAME
3 SEARCH BY CITY
4 SEARCH BY TELEPHONE AREA CODE
5 SEARCH BY TELEPHONE EXCHANGE
6 SEARCH BY TELEPHONE LAST FOUR DIGITS
7 SEARCH BY NAME (INCLUDING MATRONYMIC AND PATRONYNMIC FOR HISPANIC NAMES)
8 POSSIBLE MATCH OF PARTIAL DATA

(Note that examples 4, 5 and 6 resemble STC codes)