Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

APPLICATION OF THE BATCH ORIENTED RETRIEVAL SYSTEM MODOC 4*

WALTER KOCH

A special feature of the program system MODOC 4 ("Modular Documentation") is its application to information retrieval. This system enables the creation as well as the evaluation of data bases. The main principle of MODOC is the use of a standardized record format that allows fast access to all data elements and the use of one program system for various applications. The query language provides the formulation of complicated search requests and supports also free text searching. Two search systems can be used: one based on sequential files, the other based on inverted files that include word fragments of equal length. An output formatter enables free arrangements of data elements when printing search results.

Introduction. MODOC 4 is a software package that consists of several computer programs. These programs — the "modules" of the whole program system support many activities of documentation and information centres or libraries. The name "MODOC" is a translation of the German expression "Modulare Dokumentation" (MODOK). The system is used to create data bases (bibliographical files or statistical files) to evaluate such data bases and to process bibliographic records in different ways. Literature data bases can be used for information retrieval and for printing various catalogues.

Creation of Data Bases. All data bases have records with a unique structure that facilitates the application of all programs to all files. This structure is generated either by the data gathering program that also verifies new data or by a data transformation program which transforms other data structures as MARC to this unique structure. That makes it possible for documentation centres to run SDI-services (SDI: selective dissemination of information) based on magnetic tape services offered by information centres as Chemical Abstracts Service (USA) or by the Institution of Electrical Engineers (England). Libraries can use this system to merge different bibliographic files for printing one union catalog.

The record structure provides storage of numerical data or strings of characters. Each data element of a record is identified by a tag which provides access to small units of items. These tags are used by the data gathering system, by the query language of the retrieval processor and by the command language to create edit formats. MODOC 4 uses one unique tagging scheme to describe the different data elements. That means that beside the restructuring of other data as mentioned above it is necessary to transform tagging schemes to MODOC standards. The internal representation of a tag is a number which is the result of a mapping of the tags by polynomial transformation.

---

* Delivered at the Conference on Systems for Information Servicing of Professionally Linked Computer Users. May 23-29, 1977, Varna.

SERDICA Bulgaricae mathematicae publicationes. Vol. 10, 1984, p. 132—135.

This way a user of the system has the possibility to develop his own tagging scheme suited to his problems.

If for example in a bibliographic file "10 *A*" identifies the data field which contains data for the first author the mapping into a number can be done the following way.

$f(c_1c_2c_3)=(i_1 \cdot b_2+i_2) \cdot b_3+i_3$: mapping "*f*" for a 3-digit tag "$c_1c_2c_3$".

i_k : index-1 of the character c_k within the character set S_k which can be used for the k-th position in a tag.

b_k : number of elements of S_k.

$f(10\,A)=(1 \cdot 10+0) \cdot 26+0=260$; $S_1=S_2=(0, 1, 2, \ldots, 9)$; $S_3=(A, B, C, \ldots, Z)$.

$b_1=b_2=10$; $b_3=26$. (e. g.: for letter "*A*" i_3 has the value $1-1=0$).

Using this method each record consists of a numerical part including tag- pointer- and length specifications and a separate part which includes all textfields. This is important when the system is implemented on a word machine that makes it necessary to compress strings for saving space for core and mass storage.

Data bases consist of records that are put into sequential order according to their unique accession numbers that are assigned to each item in a file.

The Retrieval Processor. As the system primarily is used for literature information retrieval this feature of MODOC shall be outlined in some detail.

The query language is based on boolean logic (or, and, andnot, not) but it is also possible to use weighted terms for document retrieval. Logical expressions that identify search requests consist of "search codes" (operands), logical operators and an unlimited number of parentheses. Search codes are assigned to each search term and indicate the data element to which the matching process is restricted. Therefore all codes are split up into two parts: a tag and a number which allows an unique identification of all search terms within a special data element.

Searching for an author (authors shall be contained in data fields tagged "10 *A*") whose name is *MULLER* can be accomplished as follows:

10 *A*1 = *MULLER* "=": separation mark: 10 *A*: tag of the search field

1 : identification of the search term within field 10 *A*

MULLER : search term.

One has to define all search terms before the definition of a logical expression. A request "documents written by *MULLER* and dealing with *INFORMATION RETRIEVAL*" using special data elements (tagged 60 *A*) that contain descriptors can be formulated this way:

10 *A*1 = *MULLER*

60 *A*1 = *INFORMATION RETRIEVAL*

LX = 10 *A**60 *A*1

"*LX*" identifies a logical expression while "*" is the symbol which is used for the logical operator "and". Identification codes like "*LX*" can also be used in other — following — expressions. Double defining of an expression deletes the previous definition. An important feature of the search system is the possibility of truncating search terms. MODOC provides every kind of truncation, a fact which is of great interest in free text searching.

Performing batch retrieval processes one has to define all search terms and logical expressions. The system saves these data in a special file. Input is done by punched card and verification of new data is supported by symbol

and syntax checkers. If at least the input data for one profile are correct the actual search can follow immediately. MODOC provides two systems for the execution of search requests. The sequential file search system is relatively simple and based on character matching methods. Records are searched in sequential order and there is no restriction for searching only special data elements. If there are many search terms and profiles to be searched in one batch process the execution could last very long because all search items must be checked against every item in a search file. For this reason a search run can be interrupted by operator intervention and can be restarted because all working files are saved on magnetic tapes. To avoid long search runs a new system that uses inverted files was developed. Since the inversion of a search file takes some amount of computer time one has to decide if it is better to use the sequential system and at what time it is advisable to turn over to inverted files. This depends on the number of search items and on the kind of data bases. For a current awareness service based on Chemical Abstracts Condensate tapes for example it is justified to process up to 70 profiles before changing the search systems.

The inverted file system executes searches in two stages. First a preselection is performed. At this stage only a subset of all searchable data elements of a file can be searched. For this reason the query formulations are restricted to some selected data elements as descriptors, titles or authors. To provide possibilities during the preselection phase the inverted file consists of word fragments of equal length (three characters). These word fragments can easily be hash-coded and allow fast access to the connected sets of documents. Since every search term also consists of a combination of these word fragments a search item is retrieved by selecting its wordfragments and combining the corresponding sets according to the frequency of the fragments. For example the search term "*RETRIEVAL*" is split up to: *RET, ETR, TRI, RIE, IEV, EVA, VAL*. An advantage of this method is the small number of entries in the inverted file and the independence of data base languages. A disadvantage is the size of the sets (pointerlists to documents) connected to each fragment. After the preselection phase the selected documents are searched in sequential mode.

The Output Formatter. To print search results a modification of MODOC's catalogue printing system is used. The "print processor" uses a command language which operates on the tags of those data elements that are to be printed. The user of the system can define the length of print lines, head lines, the number of hits printed on one page and so on. Tabsetting facilities enable the user to arrange data elements on a printout as he likes it. It is also possible to put character strings between data elements to make an output more legible.

Technical Aspects. MODOC 4 was designed and developed at the Computer Centre Graz for a wide range of applications using various types of computer systems. The programs are written in FORTRAN IV. That made it possible to run the system on different computers as IBM, CDC, UNIVAC. Since the system primarily and very frequently is used at the Computer Centre Graz where the Institute for Mechanized Documentation is located it was necessary to optimize parts of MODOC in consideration of runtime. Therefore some programs especially for input and output handling were rewritten in UNIVAC-Spurt Assembler. The system is also usable in an on line mode but

this is not very convenient because of the complexity of the query language which was designed for batch processing.

Applications. At first the system was used to print catalogues and to build up data bases. The regional governments of Styria and Salzburg are using the system for evaluating data on road conditions and for printing periodically catalogues of the state of regional roads.

Because of its flexibility the system is very well suited to run SDI current awareness services on bibliographic data bases. The Institut für Maschinelle Dokumentation uses the system for various SDI-services. This Institute offers services based on CAC, INSPEC, and ERIC-tapes for industry and universities. For small data bases as IRRD (International Road Research Documentation) the system is also used to execute retrospective searches. The IRRD data base contains now approximately 30.000 items.

Institut für Maschinelle Dokumentation
Steyrergasse 17, A-8010 Graz, Austria

Received 11. 9. 1977