DOCUMENT RESUME

ED 075 988 EM 010 866

AUTHOR Olson, Tom; And Others
TITLE WISE-ONE System Description.
INSTITUTION Wisconsin Univ., Madison. Dept. of Educational Administration.
PUB DATE 14 Mar 73
NOTE 6p.; See also EM 010 865

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Computer Programs; Electronic Data Processing; *Information Retrieval; Information Services; *Information Systems; *Relevance (Information Retrieval); *Search Strategies
IDENTIFIERS ERIC; University of Wisconsin; *WISE ONE

ABSTRACT

WISE-ONE is a bibliographic information retrieval system which is designed to perform keyword searches of such data-bases as the ERIC RESUMAST and the ERIC CIJEMAST. Produced as a result of the search are the ERIC citation numbers, titles, authors, and, in the case of the journal file, the journal citation. Because WISE-ONE allows for nesting of the search formula to a depth of fifteen parenthetic levels, it gives the user a great deal of power in finding entries of interest. The heart of the system is the hash coding scheme which is incorporated into the data-base structure. A hash coding scheme is a method of telling the computer the storage location of a record based on the search key contained within the record. WISE-ONE is currently running on the Univac 1108 computer at the computing center at the University of Wisconsin at Madision. Detailed explanation of how the system works is provided in this paper. (JK)

WISCONSIN INFORMATION SYSTEMS FOR EDUCATION (WISE)
WISE-ONE SYSTEM DESCRIPTION
WISE-ONE*

Tom Olson
with
Donald N. McIsaac
Dennis W. Spuck

WISE-ONE is a bibliographic information retrieval system which is designed to perform keyword searches of such data-bases as the ERIC RESUMAST and the ERIC CIJEMAST. The system operates in both batch and interactive modes and takes requests on the data-base in the form of search formulas. At Each step in the search, the number of references to the last keyword and the number of references in the search queue at that point are reported. Produced as a result of the search are the ERIC citation numbers, titles, authors and in the case of the journal file the journal citation.

WISE-ONE allows the user to develop complex search formulas through the unlimited use of the logical operators AND OR and NAND. It also allows for nesting of the search formula to a depth of fifteen parenthetic levels. In the interactive mode the user enters his search formula dynamically, giving him a great deal of power in finding entries of interest.

The heart of the system is the hash coding scheme which is incorporated into the data-base structure. A hash coding scheme is a method of computing the storage location of a record based on the search

---

*The ERIC search program - WISE-ONE - was funded by the School of Education, Department of Educational Administration, Wisconsin Information Systems for Education (WISE). Mr. S. C. Yang and Professor Venezky contributed to the development of the hashing scheme. The program was also a class project in Computer Science - CS 638 taught by Professor Travis. These contributions are acknowledged and appreciated.

key contained within the record. The algorithm WISE-ONE employs generates two numbers: the hash address and the virtual key or residue. These two numbers correspond to the remainder and quotient of the division of the keyword bit pattern by the size of the base table; this process is called hashing.

The structure of the data-base is a linked table scheme and is an adaptation of a direct chaining hashing scheme which uses a linked list structure. There are three types of tables employed in this structure; a base-table, collision tables and citation tables. The role the hashing scheme and the table structure play in the structure of the data-base is best explained by tracing the search process. See Figure I.

When a search key is entered, it is hashed as described above. The hash address is used to point into a base table. The base table is a core resident list which contains the address of all collision tables which reside on secondary storage. The collision table contains the residue of all keys which hashed to the same base-table address. The collision table is searched for a residue which matches that of the hashed key. Associated with the matched residue is the address of a citation table. This table contains a list of the citation numbers of articles which reference the keyword. This list is then used by the search logic routine to abstract the citation numbers which fit the search formula and to place these numbers in the search queue.

When the constructed search queue is complete, the citation numbers are hashed in the same manner as the keywords. The hash address is used as a pointer to another base-table to obtain a link to a collision table on secondary storage. The collision table is then searched for the matching virtual key or residue and the associated link is then followed

to obtain the title, author and journal citation of the ERIC number. This process is repeated for each citation number in the search queue until all citations have been printed.

The creation and update of the data-base follow a different line of development than the search process. The keywords in the form of descriptors, identifiers and authors last names are abstracted from the ERIC tapes along with the title, author and date of the citation. Each keyword is hashed and the hash address residue and keyword are written into a file along with the ERIC citation number. The title, author and citation numbers are written into another file. The keyword file is then sorted on citation number within residue within hash address. This file is then merged with the existing master file to create a new master file. The master file contains all the information in the proper order for easy generation of the table structure. The data-base search files are then generated from master file and the title and author file. (Figure II).

There are a number of advantages to this method of storage and retrieval. The most notable being its extremely fast search time. The CPU time on the 1108 per keyword is in the order of hundreths of seconds. The overall search time is less than a tenth of a second per keyword.

Another important point feature of this search method is that search time will not increase significantly as the data-base grows in size. This is due to the fact that the number of probes to the disk to search for any keyword is two, one to read the collision table and one to read the citation table. The only portions of search-time that will increase are the time required to search the collision table for the residue and the time the logic section of the program needs to process the longer lists.

WISE-ONE is currently running on the Univac 1108 at the computing center on the University of Wisconsin-Madison campus. It is written in 1108 assembler and Fortran V. It uses about 31K 36 bit words of core storage and about 1500K words of disk storage for each file. The nature of the hashing scheme forces the code to be machine dependent and it would take considerable reprogramming effort to bring this system up on computers other than a UNIVAC 1100 series machine.

FIGURE I

Hash Coding Scheme

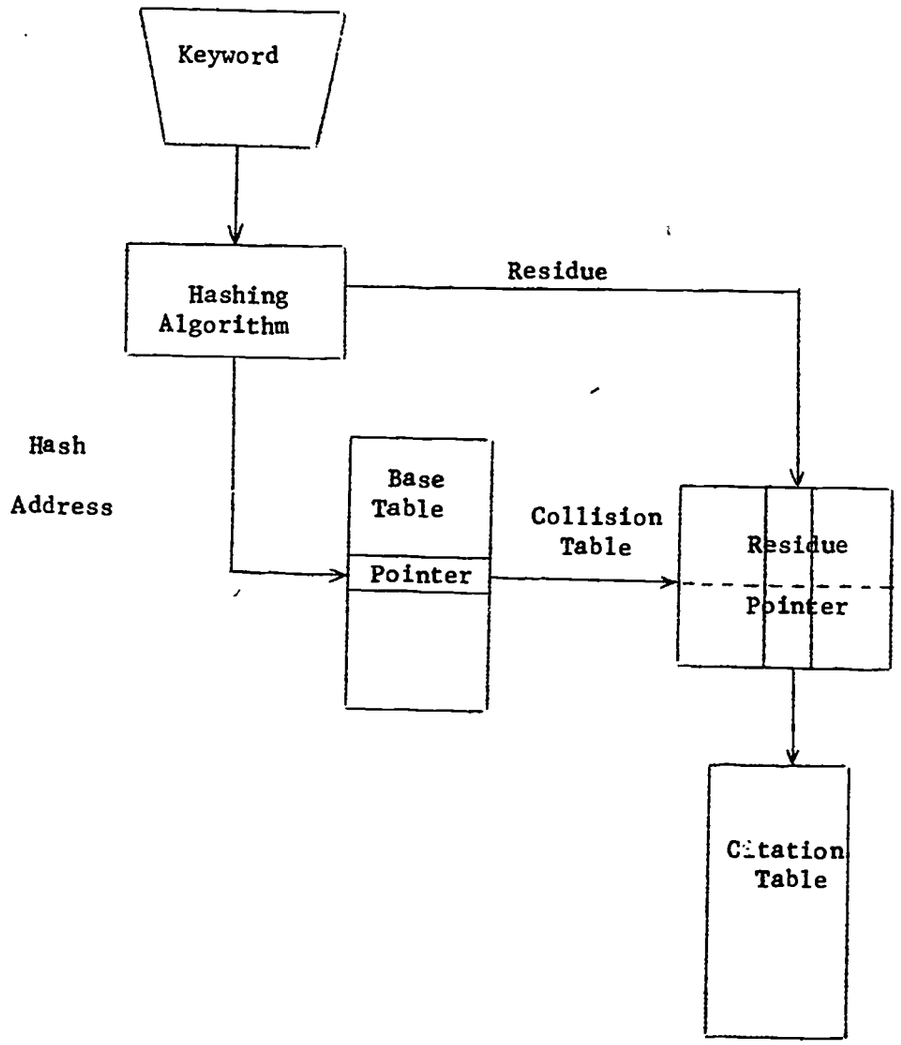

FIGURE II

File Generation and Update Procedures