INTRODUCTION AND PERSPECTIVES
FOR THE 1971
ACM INFORMATION STORAGE AND RETRIEVAL SYMPOSIUM

Jack Minker and Sam Rosenfeld
University of Maryland, College Park, Maryland, and
National Aeronautics and Space Administration, Washington, D. C.

ABSTRACT

An introduction and some prospectives are provided for the 1971 ACM Information Storage and Retrieval Symposium held at the University of Maryland on April 1 and 2, 1971. The symposium, sponsored by the University of Maryland, the National Aeronautics and Space Administration and the Special Interest Group on Information Retrieval (SIGIR) of the ACM, focuses on advances in techniques in the computer oriented technology of information retrieval. Early developments and the status of recent efforts in document retrieval, question-answering and data management systems are reviewed briefly.

KEY WORDS AND PHRASES

document retrieval, question-answering, data management, automatic indexing, natural-like languages, syntactic analysis, semantic analysis

I. Developments in Information Retrieval

The field of information storage and retrieval may be partitioned into three areas: document retrieval systems, question-answering systems, and generalized data management systems. Although the techniques required in each of these three areas are quite similar, work in each area has been performed in isolation of the others. It is towards bringing together the common underlying methods in these areas that this symposium is directed.

To better understand our objectives with respect to the symposium, it will be well to review the developments in each of the three areas.

The modern era of information storage and retrieval is generally considered to have started with the landmark article written by Vannevar Bush (3) that appeared in 1945 in the Atlantic Monthly. The concept of a library with individuals having visual displays connected remotely to a large store by terminals was foreseen by Bush. Indeed, the concept of statistical association techniques was described implicitly in the article.

Ten to fifteen years elapsed before some of the techniques described by Bush came into being. The use of computers for information retrieval started in earnest in the late 1950's. It was during this time that the Key Word in Context Index (KWIC) was developed by H. Peter Luhn (10), and independently, and perhaps earlier, by a group at Rocketdyne Corporation (4). Work in statistical analysis of text was prominent in that time period, and culminated in 1964 with an important conference sponsored by the National Bureau of Standards (18) and devoted entirely to statistical techniques. The 1964 conference represents, perhaps, the high point of work in statistical analysis for document retrieval systems. The State-of-the-Art Report written by Mary Elizabeth Stevens (17) in 1965 surveys the major techniques in automatic indexing. The Stevens' report which was updated by adding a new section and reissued several years later, shows no great progress in the intervening years.

It was also during the late 1950's that work in question-answering systems started. A system, called BASEBALL, developed by Green et al. (8) was implemented. Syntactic analysis of English-like query statements was used in BASEBALL. The development in BASEBALL led to the recognition of the importance of natural-like language inputs to a computer for query systems.

The first conference sponsored by the ACM in information storage and retrieval was held in 1961 in Princeton, New Jersey (7), with Jack Minker and Mandalay Grems as co-chairmen. Several important papers appeared at that conference. Among them were the string manipulation language, COMIT, developed by Yngve (19), and the Cheatham and Warshall paper (5) on techniques to translate retrieval requests couched in a semi-formal English-like language.

The technology of generalized data management systems also started in the late 1950's at a number of places (11). Some of the early work was done in the government at the David Taylor Model Basin, and in industry at the General Electric Company and RCA . The major techniques in this technology were developed during that time, although they were not made available to other researchers who subsequently had to reinvent the technology themselves.

The 1960's, in contrast to the flurry of developments in the late 1950's was somewhat disappointing as far as the technology of information storage and retrieval was concerned. One would have hoped for the achievement of operational document retrieval systems used by a wide variety of individuals. However, only a handful of large government organizations, or government sponsored

:ganizations, have implemented document retrieval /stems. Notable among these is the National ibrary of Medicine's MEDLARS effort (2) that ecame operational in the second half of the)60's.

The work by Salton and his colleagues on the IART system (14) attempted to provide information ; to the effectiveness of alternative methods iat had been proposed for automatically indexing ext. The results are summarized in Salton's ook (14). However, many questions still remain iresolved since the text samples that were sed were rather small relative to practical :oblems.

A conference on the Intrex Project (13), large automated library research effort at the assachusetts Institute of Technology, was held in ie mid 1960's. The Intrex project has adopted iny of the objectives originally set forth by annevar Bush. However, few technical papers appear i the literature on Project Intrex. A progress eport on Project Intrex was presented in a series E papers at the 1969 Spring Joint Computer Con- erence (1). The effort still remains in the esearch stage, and no dramatic advances have een evidenced by the work.

The technology of question-answering systems, urveyed so optimistically by Simmons in 1965 (15) id again in 1970 (16), still remains in the re- earch stage. No operational question-answering ystems have been developed, nor are any expected o be developed within the near future. In anoth- r survey of the technology, Minker and Sable 12), note that the reasons for this are not ecause of fundamental limitations in computer ardware or software technology, but becuase of undamental gaps in our intellectual knowledge f syntactic and semantic analysis of natural anguage, and in effective search procedures or performing inferences by mechanical means.

One possibly bright area in the 1960's was he development of a large number of generalized ata management systems (11). The CODASYL ommittee surveyed several such systems in 1969 6). And yet, the techniques employed by the ost advanced systems were no more advanced than hose described by the early systems. It took lmost ten years to recognize the importance of he technology.

Perhaps one of the most glaring disappoint- ents in the 1960's is the lack of effective ystems for library automation, and the paucity f such systems. One of the world's largest ibraries, the Library of Congress, although aving sponsored a major study on automation of ts functions (9), lies mainly untouched by omputers. The call for networks of libraries eems to be currently rampant although the develop- ent of economical systems for small libraries as yet to be attained. Fundamental processing techniques for work in information storage and retrieval either do not yet exist, or are not generally available. Indeed, advanced textbooks have not been written for either the area of question-answering systems or generalized data management systems. Salton's textbook (14) is a pleasant contrast, in that it makes available many of the major techniques currently applicabl in document retrieval.

II. The 1971 Information Storage and Retrieval Symposium

In this symposium, we focus on advances in techniques in the computer oriented technology of information retrieval. It is our contention that unless the fundamental processing technique are developed, and made available to workers in the field, information retrieval will continue to flounder, as we believe it did in the 1960's. Hence, although many papers were received for this symposium concerning overviews of interesti systems, only those papers in which detailed algorithms, procedures, heuristics, evaluative methods or theoretical concepts were described have been accepted.

Two issues are perhaps common threads of thi symposium: specifying natural-like languages for the user to be able to formulate questions, and designing effective and efficient informatic systems. Of course, these issues provide, in some sense, different perspectives of the same problem. The specification of a user language provides an outside view of an information syste whereas the design provides the inside view of such a system. The technologies meet and overla Neither the language designer nor the question-answering designer can ignore data management and file organizations. Nor can the file organi ignore the nature of the query language. In bot cases, however, general techniques are required handle classes of problems rather than be re- stricted to only one application. The tradeoff problems in file organization techniques are explored in the session "An Approach to Researcl in File Organization," chaired by Dr. Michael Senko. Some optimization methods that may be used to save space on peripherical devices are developed in the session chaired by Dr. Michael Lesk, "Optimizing Methods." Techniques develope for answering questions both for document retrie and question-answering systems are covered in tl sessions chaired respectively by Dr. Harold Borl "Natural Language in Document Retrieval Systems, and Dr. Jack Minker, "Natural Language Processi and Query Systems." Theoretical issues that ha arisen in question-answering systems, are explo in the session, "Theoretical Concepts," chaired by Dr. H. P. Edmundson. Finally, developments : evaluating and developing generalized data mana; ment systems are covered in the session, "Data Management Systems," chaired by Mr. John Gosden

We hope that future conferences sponsored b:

cial Interest Group on Information Retrieval
GIR) of the ACM will continue to stress
oretical results; optimization methods; and
orithmic, heuristic, and computational methods
information storage and retrieval as we have
ed to do in this symposium. It is our hope
t through such efforts, greater progress
be achieved during the 1970's than has been
ieved in the 1960's. Perhaps then we will be
e to come closer to achieving Vannevar Bush's
am.

REFERENCES

AFIPS CONFERENCE PROCEEDINGS, vol. 34; 1969 Spring Joint Computer Conference, Boston, Massachusetts, May 1969, p. 457-490.

AUSTIN, C. J. The medlars system. Datamation, 10:12 (December 1964) p. 28-31.

BUSH, VANNEVAR. As we may think. The Atlantic Monthly 176, (July 1945) p. 101-108.

CARLSEN, R. D.; GERNER, N. H.; MARSHALL, H. S. Information control. Industrial Engineering Department 564, Ref. 11, Rocketdyne Div., North American Aviation, Canoga Park, California. (August 1958).

CHEATHAM, T.; WARSHALL, J. Translation of retrieval requests couched in a "semiformal" English-like language. Communications of the ACM, 5:1 (January 1962) p. 36-38.

CODASYL SYSTEMS COMMITTEE TECHNICAL REPORT. A survey of generalized data base management systems. Association for Computing Machinery, May, 1969.

COMMUNICATIONS OF THE ACM, 5:1 (January 1962).

GREEN, B. F., JR.; WOLF, A. K.; CHOMSKY, C.; LAUCHERY, K. BASEBALL: An automatic question-answerer. In: Proceedings of the Western Joint Computer Conference, 1961, vol. 19, p. 219-224.

KING, GILBERT; et al. Automation and the Library of Congress. Library of Congress, Washington, D. C., December, 1963.

) LUHN, HANS PETER. Key-Word-In-Context idea for technical literature (KWIC Index). Presented at American Chemical Society, Division of Chemical Literature at Atlantic City, N. J., 14 September 1959. Report No. RC127, International Business Machines Corp., Yorktown Heights, N. Y., 30 September 1959.

) MINKER, JACK. Generalized data management systems - some perspectives. University of Maryland, Computer Science Center, TR69-101, December 1969.

(12) MINKER, J.; SABLE, J. D. Relational data system study. RADC TR-70-180, Final Technical Report, Rome Air Development Center, September 1970.

(13) PLANNING CONFERENCE ON INFORMATION TRANSFER EXPERIMENTS (INTREX). Woods Hole, Massachusetts, 2 August - 3 September 1965. Report: Overhage, Carl F. J.; and Harman, R. Joyce (eds.). MIT Press, Cambridge, Mass. (1965) 276

(14) SALTON, G. Automatic information organization and retrieval. McGraw-Hill, Inc., New York, 1968.

(15) SIMMONS, R. F. Answering English questions by computer: a survey. Communications of the ACM, (January 1965) p. 53-70.

(16) SIMMONS, R. F. Natural language question-answering systems. Communications of the ACM, 13:1 (January 1970) p. 15-30.

(17) STEVENS, MARY ELIZABETH. Automatic indexing: a state-of-the-art report (revised ed.). U. S. Department of Commerce, National Bureau of Standards, NBS Monograph 91, February 1970, 290 p.

(18) STEVENS, M. E.; GIULIANO, V. E.; HEILPRIN, L. B. Statistical Association methods for mechanized documentation. United States Department of Commerce, National Bureau of Standards, Miscellaneous Publication 269, Washington, D. C. 1964.

(19) YNGVE, V. H. COMIT as an IR language. Communications of the ACM, 5:1 (January 1962) p. 19-28.