absorbed by IFIPS. Also, the British Standards Institution recently circulated to many individuals in the U.S. a "Draft British Standard Glossary of Terms Used in Data Processing" for comment.

All these groups are working toward the same goal. They hope to clarify and establish the meaning of certain technical terms used in the computing literature so that these words convey the same meaning to the readers.

Theory of Files

Numerical Analysis Research Project, University of California at Los Angeles, Calif.

Reported by: Lionello Lombardi (May 1961)

Descriptors: **theory of files, flow control expressions, non-arithmetic data processing, data flow, algebraic business language, von Neumann languages, sorting, Boolean algebra**

In hardware design, the theory of files is used as a tool to analyze the features of data flow of the systems for which the equipment is designed. As a result, it is proved possible to formulate a punched card data processing system considerably less expensive than any presently available. The logical features of this system, based on a novel organization of data flow, is determined by synthesis.

Currently, the theory of files is being applied to the study of the flow of information through random-access fixed-plus-erasable memory systems in order to select a compact set of parameters which characterize the flow involved in any specific application, and then to formulate relations between these system parameters, the minimum storage capacity requirements of the systems, and the optimum storage capacity distribution among the components.

Development and applications of the theory of files are carried out under the sponsorship of the Office of Naval Research. The following results have been obtained:

(a) Determination of a common pattern to which the coordinated data flow conforms of any non-arithmetic data processing procedure involving files, such as machine accounting, access to stored information, dictionary analysis, etc. In particular, it is shown that the data flow configuration of any procedure involving n files can be fully represented at any time by means of only $5n+4$ boolean variables (the *indicators*).

(b) Development of a pattern of system language (the Algebraic Business Language) which allows for the use of logico-mathematical techniques to describe and control the data flow by means of specially designed boolean expressions (the *Flow Control Expressions*).

(c) Description of processes in rigorous mathematical terms where the relevance of arithmetic operations is low and the main problem is logical input-output coordination. Such procedures, where a small amount of processing is performed on relatively large amounts of information, include most business data handling and document processing procedures. Now it is possible to represent such processes as sets of operations defined in boolean algebras whose elements are files.

For example, a complete k-way v. Neumann sorting procedure is represented by the formula,

$$F_i^{(m+1)} = \sum_{k*(i+1)j}^{k*(i+1)} F_j^{(m)}, \quad (i = 0, 1, 2 \cdots ; \ m = 0, 1, 2 \cdots)$$

where $F_s^{(t)}$ represents the sth ordered sequence of records available after t runs.

(d) Representation of procedures as sets of equations relating the input data to the output results.

(e) Development of other languages, e.g., the Algebraic Business Language. The v. Neumann-type data processing languages (i.e., languages in which procedures are represented as images of flow charts) do not satisfy the needs of most non-arithmetic procedures.

References:

Goldstine, H.; Neumann, J. v. Planning and coding for an electronic digital computer. The Institute for Advanced Study Publications, 1947.

Lombardi, L. Theory of files. Proc. 1960 Eastern Joint Comput. Conf., Paper 3.3.

——, System handling of functional operators. *J. ACM 8*, (1961), 168.

——, Mathematical structure of non-arithmetic data processing procedures. Submitted for publication.

——, Inexpensive punched card equipment. In press, *Journal of Machine Accounting*.

——, Logic of automation of system communications. In preparation.

——, Coding in storage and searching systems. In P. Garvin (Editor), *Natural Language and the Computer*, ch. 14; in press, McGraw-Hill Book Co.

A Class of Search-Models for Machine Retrieval

Information Retrieval Project, IBM Research Center, Yorktown Heights, N. Y.

Reported by: Eugene Wong (May 1961)

Descriptors: **search, search-strategy, optimization, access to mechanical storage, information retrieval, stochastic models**

A class of theoretical models for studying optimum search-strategies in one dimension, has been constructed with a view towards the eventual application of these strategies to machine retrieval. The basic assumption underlying these models can be briefly stated as follows:

(1) The probability that the object of search lies in the interval $(x, x+dx)$ is $p(x)dx$, where the density function $p(x)$ is known a priori.

(2) The search is conducted with constant speed v.

(3) It is possible to skip, i.e., to move from one point to another without searching, with constant speed s. In general it can be assumed that $s > v$.

The major feature of this model is a balance between the advantage of always searching where the probability of success is the highest against the disadvantage of frequent skipping that this may incur. This feature exists in many practical storage devices, e.g., magnetic tape.

Without loss of generality any search procedure can be expressed as a sequence of alternately searching and skipping operations. Using the letters v for searching s for skipping, the search-sequence can be expressed as

$$x_0 \xrightarrow{v} x_1 \xrightarrow{s} x_2 \xrightarrow{v} x_3 \xrightarrow{s} x_4 \xrightarrow{v} \text{etc.}$$

The object of the analysis is to find the sequence $\{x_n\}$ which optimizes the search in some sense.

The model, that has been described, with $p(x) = (\alpha/2)e^{-\alpha|x}$ and the criterion of minimum mean search-time T has been studied intensively with considerable success. Both parametric optimization, e.g., let $x_n = (-1)^n(x_0+n\delta)$ and find x_0 and δ which minimize T, and non-paremetric optimization have been achieved.

Analysis with other forms of $p(x)$ and other criteria for optimization are being undertaken. Modifications of the basic model to adapt more closely to practical storage arrangements are also being considered.