SPECULATIONS CONCERNING INFORMATION RETRIEVAL

I. J. Good

IBM

RESEARCH CENTER

INTERNATIONAL BUSINESS MACHINES CORPORATION
YORKTOWN HEIGHTS, NEW YORK

SPECULATIONS CONCERNING INFORMATION RETRIEVAL

by

I. J. Good
Royal Naval Scientific Service

ABSTRACT: Various schemes of classification and measures of relevance are reviewed from the point of view of how they might fit into a general theory of information retrieval. Problems in such areas as syntax, semantics, and library economics that would have to be considered by theory are touched upon. Also discussed are contributions that might be expected from such related studies as are being made in mechanical translation, artificial intelligence, and information theory.

Research Report
RC-78
December 10, 1958

SPECULATIONS CONCERNING INFORMATION RETRIEVAL*

I.J. Good

INTRODUCTION

At the recent International Conference on Scientific Information (I.C.S.I.),† Area 6 was concerned with the possibility of a general information retrieval theory. Most people held that no adequate general theory existed yet. Perhaps a general theory could be defined as one that usefully unifies the field and requires mathematical training to be understood. What I propose to do is speculate on what the ingredients of such a general theory might be.

INFORMATION RETRIEVAL AS AN INTERACTION OF TWO INFORMATION-HANDLING SYSTEMS

The process of information retrieval is a process of communication between two information-handling systems, a man and a library. Although the library often contains men too, the two systems are very different. For one thing the memory in the library is largely verbation, whereas the man's is more semantic; also the man has a logically more complex indexing system. The man may ask vague questions and get exact answers, though not necessarily the most relevant ones. The communication between the two systems is in both directions, so that searches can profitably be sequential, provided that the delays are small. The two systems can educate each other.

A general theory of information retrieval should take into account the properties of both the man and the librarian. I am going to describe some gropings towards a general theory. My method will be to list some theories that may be relevant and then discuss them.

LIST OF POTENTIALLY RELEVANT THEORIES

The following headings are not in order of importance, but are arranged merely for convenience of discussion.

---

*Given at the Mohansic Laboratory, IBM Research Center, Yorktown Heights, New York, December 10, 1958.

†International Conference on Scientific Information, Washington, D.C., November 16-21, 1958.

Classification and indexing

Trees

Botanical classification

Lattice theory

The algebra of classes, and Boolean algebra (the algebra of logic)

Syntax

Partial ordering

Oriented linear graphs

Oriented linear graphs with impedance

Random branching theory

Random motion on a network

Clumps

Semantics

Mechanical translation

Information flow & information theory

Cerebral cortex & artifical intelligence

Theories of evolution

Rational behavior

Hierarchies of memories in a computer

I shall now elaborate superficially on each of these headings.

CLASSIFICATION AND INDEXING

It will be convenient to use the word "document" in a very general sense to include books, chapters of books, paragraphs, and perhaps other things. When you have a collection of more than, say, 1000 documents it becomes important to classify or index them so as to be able to retrieve any specified document. If the index or classification can be expressed linearly, then the documents can be placed in this linear order and can act as their own index. An obvious example of a linear ordering is an alphabetical author index (perhaps with titles entered chronologically under each author). Another form of linear ordering is the familiar decimal classification, of which U.D.C. (Universal Decimal Classification) is one example. The digits may of course be replaced by letters

without essential theoretical change, though such a change would be expensive in practice where several hundred thousand index cards had already been inscribed with digits. This is an example to show how mere size can lead to inflexibility. Even the cost per book due to a change of indexing system may be larger for large libraries since the complexity of the indexing system needs to be greater for larger libraries if it is to give the same service (i.e. the same number of documents per query). For a constant "service", the size of the indexing system is proportional to n log n, where n is the size of the library.

TREES

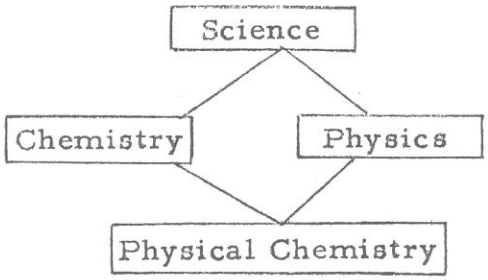
Both the U.D.C. and ordinary alphabetical indexing are examples of tree-like structures. Unfortunately knowledge does not have a tree structure in detail, so that U.D.C. does not achieve all it was intended to. The subject of information retrieval itself provides an example, since such U.D.C. headings as Telecommunication, Linguistics, Education, Psychology, Bibliography, Electronic computers, are all relevant to the list of theories given above. (This difficulty does not arise in a very serious way for alphabetical indexing.) In the filing of small typewritten documents the difficulty can in principle be overcome by using duplicate copies, but this system can easily get out of hand for large collections. With books it is even worse, but with possible future forms of storage such as micromicrofilm this duplication method may become practicable. Until that time librarians will have to make arbitrary decisions. For example, at Manchester, Kendall's books on advanced statistics were not even in the same building as the Science Library. Statistics was regarded as a part of economics, which was in the Arts Library.

BOTANICAL CLASSIFICATION

At the recent International Conference on Scientific Information, Mandelbrot suggested that some lessons may be learnt from botanical classification. One in particular is that we cannot expect equiprobable subdivisions to occur with many natural classifications.

LATTICE STRUCTURE

The structure of the subdivision of knowledge into fields in some respects corresponds more closely to a lattice than to a tree. For example, we have the following substructure:

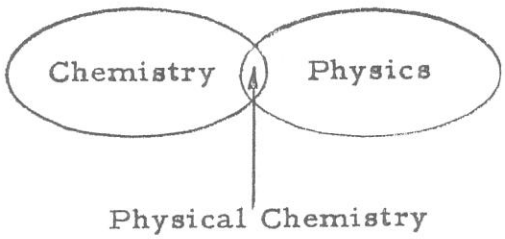


A lattice is a partially ordered system in which any two elements have a lowest common ancestor and a highest common descendant.

In order to try to preserve a tree-like classification, a librarian may decide arbitrarily that physical chemistry is to be regarded as a branch of chemistry. This decision is liable to be influenced by his customers. If most of those who request books on physical chemistry are in the Chemical Faculty, his decision may be justified. This shows how classification can depend on the properties of the customers and not just on logic.

THE ALGEBRA OF CLASSES, AND BOOLEAN ALGEBRA (THE ALGEBRA OF LOGIC)

We can get at the same idea by thinking of fields of knowledge as areas or sets, as in the following diagram.

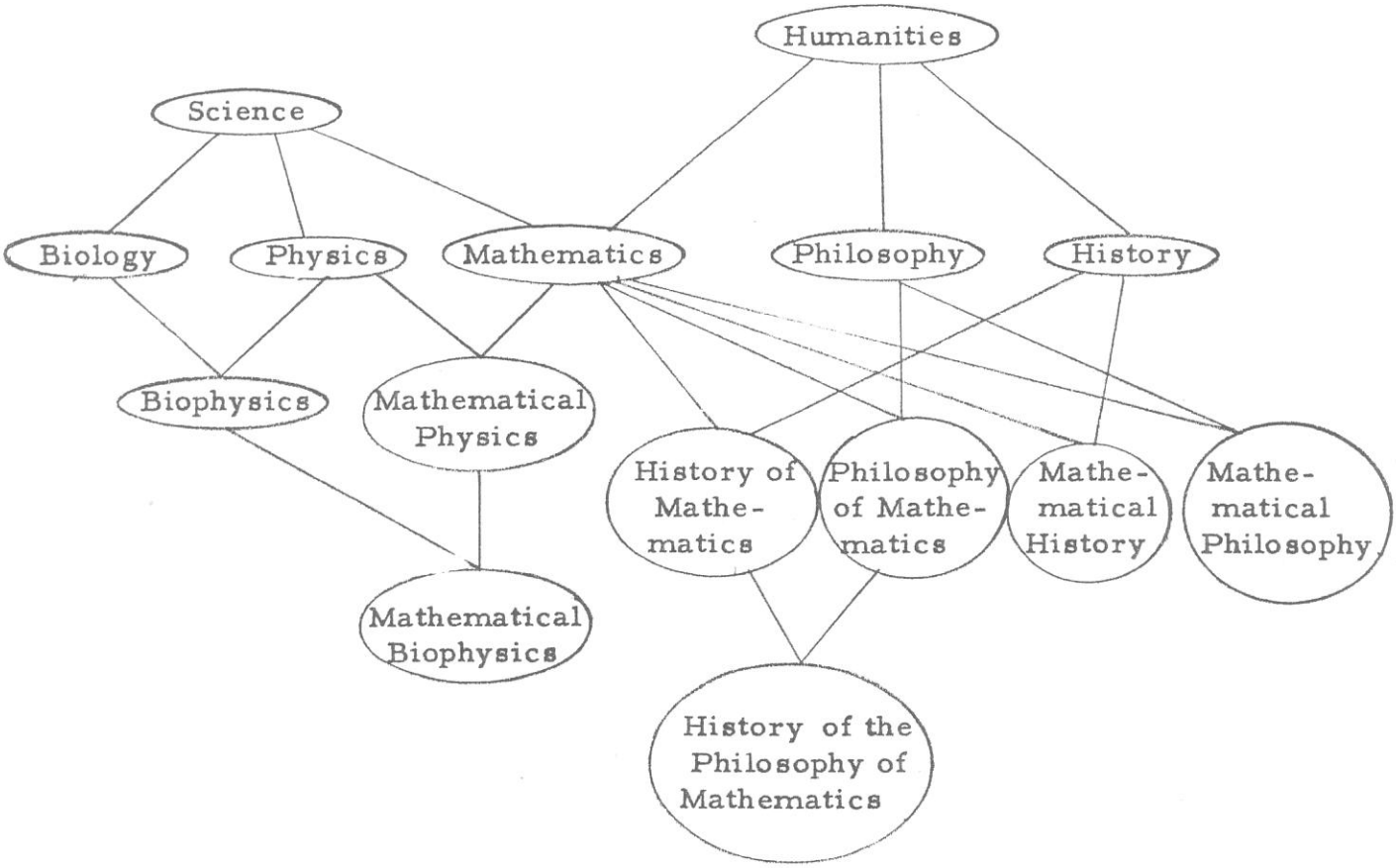


The highest common descendant of physics and chemistry is here represented by the intersection of the two classes. The union of chemistry, physics, biology, etc. is "science", but it would be impossible to specify all the subjects that make up what is now called "science" if only because this word does not have a clear-cut meaning. The difficulty of vague meanings is a permanent headache to the classifier, and he can resolve it only by making arbitrary conventions.

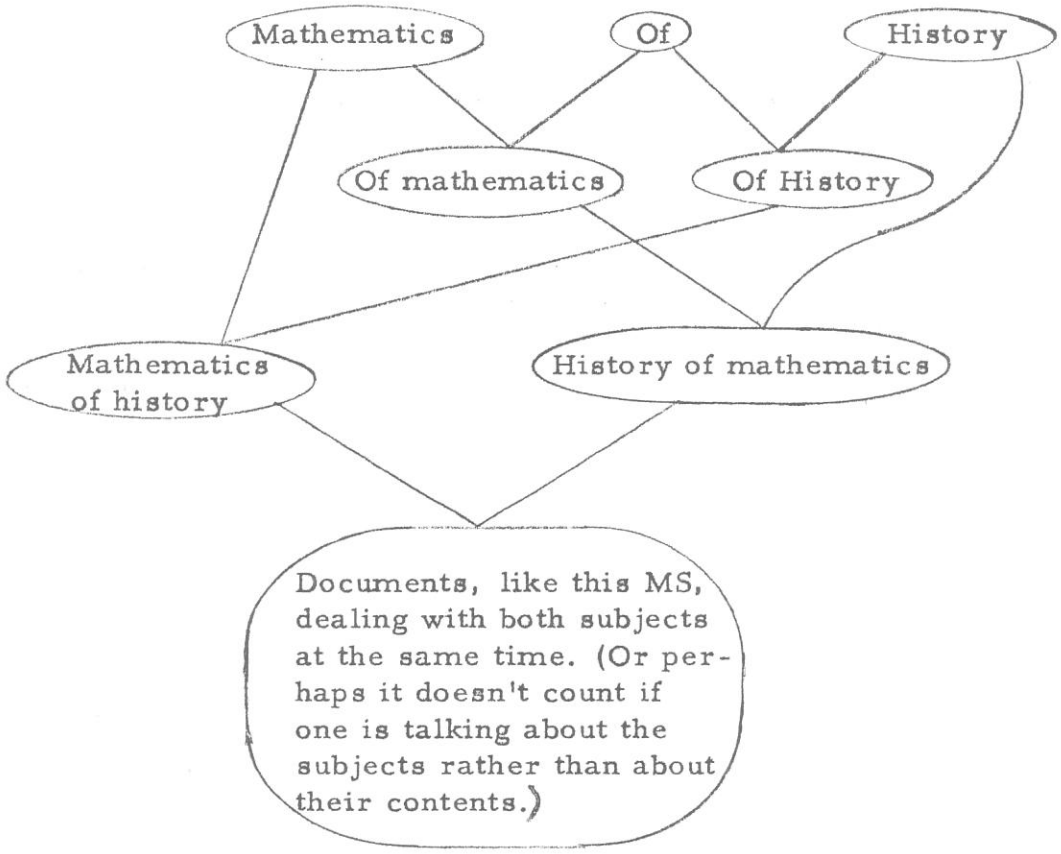
The algebra of logic comes to the same thing as the algebra of classes, but "negation" is not usually a useful operation in the present application. As Bar-Hillel mentioned at the ICSI it is more natural to use the terminology of classes than of logic if one is searching for classes of documents.

SYNTAX

An objection to the lattice model is that it does not easily deal, so to speak, with syntax. Consider the following structure.



An example of mathematical philosophy, as I am using the term here, is the definition of causality (see a forthcoming article in the British J. for the Philosophy of Science). The distinction between the expressions "philosophy of mathematics" and "mathematical philosophy" is given by the syntax, and syntax is not easily coped with in a classification system. (I understand that Forradane has done some work on this problem.) Notice that the lattice structure breaks down unless syntactical marks can somehow be introduced as elements of the lattice; thus:



Note the important question here: how can syntax be well introduced into classification systems? This question is being considered by the Cambridge University Language Research Unit.

PARTIAL ORDERING

Even though the lattice structure broke down to some extent in the above example, partial ordering was preserved. In a theory suggested by Calvin Mooers at the ICSI partial ordering was taken as basic. The relation of partial ordering is that of inclusion.

ORIENTED LINEAR GRAPHS

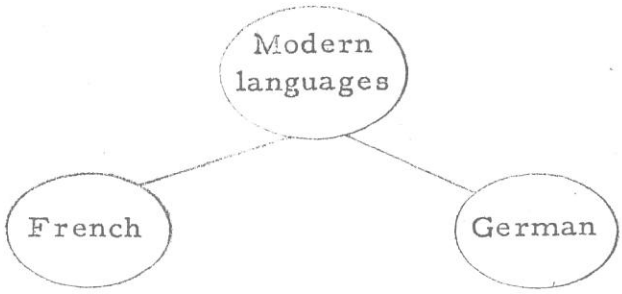
But it is not always clear what includes what. Is logic a branch of pure mathematics or conversely? Or is logic a special case of probability theory? Is operational research a branch of statistics? Is negative feedback a branch of cybernetics, cooking, or the behaviour of cows? These ambiguities may depend on how accurately words have been defined, but I suspect that some of them are fundamental. It may be possible to overcome the difficulty in the course of processing the lattice structure or perhaps by regarding the fields of knowledge as forming the nodes of an oriented linear graph.

ORIENTED LINEAR GRAPHS WITH IMPEDANCE

But I think a more interesting model involving oriented linear graphs is one in which the nodes represent something more general than just fields of knowledge. In fact, imagine that we have a node for every field of knowledge, document, proportion, phrase, word, or customer of the library. Each pair of nodes, A and B, is joined by a path with an arrow on it and an associated measure of the "impedance" for going from A to B, or the "relevance" of B to A. The relevance of B to A is not necessarily equal to that of A to B, so that we must have two paths, one in each direction. Denote the impedance from A to B by Z_{AB}. This may be a vector, since there are various kinds of relevance; but if it is a number, Z_{AB}^{-1} is a measure of the relevance of B to A. Now the $64,000 question is how to define Z_{AB}.

The following different types of relevance are all pertinent:

a. Distance apart (or, rather, its reciprocal) in a classification by partial ordering, or various measures of lattice distance. For example, in the following substructure, "French" and "German" are two links apart:



This measure is supposed to apply to fields of knowledge. It is a measure of logical relevance.

b. Frequency of use. If A and B are two documents, the probability that B is requested, given that A is requested, is a measure of the relevance of B to A. It may be called the "frequency relevance" of B to A. It was suggested by MacKay at the ICSI. It depends on a class of customers and has the further drawback that the number of samples from which the estimate is to be made is liable to be very small, especially when documents are recent. An immediate application of frequency relevance (suggested by Bigelow and MacKay) is that when document A is requested, document B may be offered to the customer unrequested. Association of words can also be measured statistically, i.e., by frequency of occurrence in the same context.

c. Reference relevance. If document A refers to document B, then we may say that B is relevant to A "by reference".

d. Citation relevance. If document B refers to document A, then we may say that document B is relevant to A "by citation".

e. Interest relevance. If customer A is known to be interested in a certain field of knowledge and a document, B, highly relevant to this field is acquired, then it is relevant to A "by interest" (and conversely).

f. Linguisitic relevance. If the vocabulary in documents A and B are similar, or correlated, then they are linguisitically relevant. It would be desirable here to allow for synonyms, and for other words known to be highly relevant to each other, but let us ignore this complication for the moment. Let documents A and B have N_A and N_B words respectively. Let W_1, W_2, W_3, ... be a sequence of words and phrases.

Let $n_A(W_1)$, $n_A(W_2)$, ... be the frequencies of occurrence of words (or phrases) W_1, W_2, ... in document A, and $n_B(W_1)$, $n_B(W_2)$, ... in document B. Here extra weight is to be assigned to words occurring in headings, summaries, abstract or index. Then define the linguistic relevance

$$R_{AB} = Z_{AB}^{-1} = F\left(\frac{n_A(W_1)}{N_A}, \frac{n_B(W_1)}{N_B}, \frac{n_A(W_2)}{N_A}, \ldots\right)$$

for some function F. A simple example of such a function, perhaps adequate for some purposes, especially if synonyms were lumped together, would be

$$R_{AB} = \frac{1}{Z_{AB}} = \sum_W \frac{n_A(W)}{p(W)N_A} \cdot \frac{n_B(W)}{p(W)N_B},$$

where p(W) is the probability of W in some wider field. (Likewise, in automatic abstracting, the importance of a word or phrase, W, can be reasonably measured by $n_A(W)/(p(W)N_A)$. Since the estimation of p (W) involves extra work, the present IBM and ITEK experiments use the less good measure $n_A(W)$, and still have some success.)

The relevance of one word, i, to another one, j, may be measured by the "association factor" $p_{ij} / (p_i p_j)$, where p_i is the probability of i, and p_{ij} that of i and j in the same context.

This whole subject of the measurement of relevance or impedance bristles with difficulties, but I believe that it will be of use long before it is anywhere near optimal. In fact, in a qualitative way, it has clearly been of great value for many centuries, otherwise the word "relevant" would not be in use.

RANDOM BRANCHING THEORY

If an information retrieval system were based on some measure or measures of relevance, then we would be interested in the distribution of the number of documents retrieved. Ignoring closed loops, we may get approximations by means of random branching theory.

RANDOM MOTION ON A NETWORK

Even more relevant perhaps would be the theory of random motion on a network, the probabilities of steps being dependent on the impedance between pairs of points. The limiting frequencies of being present at any node would give some measure of the time that should be spent at that node in the long run.

CLUMPS

There are various types of clumps. Clumps of documents correspond to fields of interest. Clumps of words correspond to possible thesaurus headings and are of interest to Parker-Rhodes and Needham. Clumps consisting of people,(or their names) and of documents correspond to lists of people to whom specialized bibliographies should be sent. Clumps of people, or of words, not associated with clumps of documents indicate a serious gap in the coverage of the library. Clumps consisting of people and words suggest the need for a specialized technical dictionary. Clumps of documents not associated with an English word would suggest the need for a new word. Conversely the existence of a word makes the corresponding clump of documents more easily noticed (whether automatically or otherwise). If our measure of impedance satisfied the triangle inequality, and was therefore a metric, then the nodes could be regarded as belonging to a metric space, and clusters of points could be defined as in the work by Neyman and Scott on clusters of galaxies. But the impedance may not be a metric. Definitions of clumps can nevertheless be provided, though I am not yet clear which definitions would be useful. If the Z_{AB} are positive numbers, a possible definition of the clumpiness of any subset of n nodes is

$$\min_{A} \sum_{B} Z_{AB}^{-k} \quad (k > 0),$$

or

$$\frac{1}{n(n-1)} \sum_{A,\,B} Z_{AB}^{-k} = \operatorname*{Av}_{A,B} Z_{AB}^{-k} .$$

A clump is a subset of large clumpiness.

Clumps can possibly be defined recursively as follows: First select as good candidates, A, for belonging to clumps, those nodes for which

$$\sum_{B} R_{AB} \quad \text{or} \quad \sum_{B,C} R_{AB}R_{BC} \quad \text{or} \quad \sum_{B,C,D} R_{AB}R_{BC}R_{CD}, \text{ etc.}$$

is large. (These definitions could be applied without too much arithmetic if most of the elements in the matrix $\{R_{AB}\}$ were zero. The limiting case of this suggestion can be obtained by taking large components of the main eigenvector of $\{R_{AB}\}$.) Let A be a good candidate. Now consider, as candidates for belongings to clumps containing A, those nodes B for which R_{AB} is not too small. For each candidate A this will define a subset of nodes B. Discard nodes B not "well connected" to the rest of the subset. This will give one or more clumps containing A. It can be separated out into individual clumps by a repetition of the same procedure. If it fails to separate out then the recursive procedure has terminated. This procedure could be applied especially easily if each R_{AB} was 0 or 1, as in linear graphs and lattices.

Having defined clumps we could go on to try to define clumps of clumps and so on. The hierarchy of clumps thus defined would provide a suggested tree-type classification.

Very often a document would belong to more than one tree; hence the requirement for the use of hierarchical classification combined with "class intersection" (such as is provided by a decimal classification combined with the colon notation).

SEMANTICS, OR THE MEANING OF MEANING.

There is a relationship between meaning and classification. A description of required information is analogous to the definition of a word. Consider, for example, the well-known "Wisdom's cow". A cow is supposed to be a four-legged animal that supplies milk, and so on. But it may have three legs, or even five, and it may not supply milk. In fact a cow may have various properties, and, according to Wisdom, no one of these is essential to its cowness. This definition is analogous to a suggestion made by Estrin at the ICSI. He suggested that in information retrieval we may insist on say k out of n index words (terms) before selecting a document.

But John Wisdom's definition of the definition of a word can be improved. For example, some properties may be essential. A newly discovered particle heavier than a proton would almost certainly not be called a "meson". A more general and more realistic definition of meaning can be given, and it has potential relevance to the problem of information retrieval. Let Q_1, Q_2, ... be a sequence of qualities. Let pi = 1 or 0 depending on whether a thing, T, has quality Q_i (or let pi equal an estimated probability that T has quality Q_i). Then the word "meson" corresponds to a function f such that f (p_1, p_2, ...) in the probability that T is a meson. In practice a word does not have a constant meaning, so that the function f depends on some other parameters, such as time, dialect, occasion and context. Note that the probabilities p_1, p_2, ... may be determined by other functions; i.e., the qualities themselves must be defined. (To understand that a cow usually has four legs we must define a leg.) Thus we could go on for ever trying to complete the definition iteratively. The process may converge. The complication of this definition of definitions should cause no surprise, since philosophers, such as G. E. Moore, have stressed the difficulty of supplying complete definitions.

An information retrieval system could be based directly on an analogy with this definition of definitions, or on various weakened forms of it. For example, a customer could say that he wants a document that certainly contains reference to either Relativity or to the Lorentz transformation and probably refers to electrons or protons, or if it refers to mercury will probably refer to perihelions, etc., and the retrieval system would return a list of documents with the probabilities of relevance attached. This list would constitute an ordered bibliography. (After making this remark at the ICSI, I was informed after the meeting that a somewhat similar idea is being tried out at Ramo-Wooldridge.)

MECHANICAL TRANSLATION

Since the library and the customer speak different languages, there is clearly some scope for processes like mechanical translation. Connections between information retrieval and mechanical translation have been pointed out by several writers, including Masterman, Needham, and Spark Jones at the ICSI. The previous remarks about syntax and semantics also show that there are connections.

INFORMATION FLOW AND INFORMATION THEORY

Any general theories about information flow should be relevant, including what is called information theory. Information theory is concerned mainly with channel capacity and the coding of stationary time series, and a more general theory of information flow may be possible.

CEREBRAL CORTEX AND ARTIFICIAL INTELLIGENCE

Since both brains and libraries are concerned with information storage and retrieval, there may be lessons to be obtained from theories or facts about the brain. At least they should teach us something about the customer. Any advances in work on artificial intelligence may also be relevant. Conversely, new techniques in information retrieval may be of use in the design of automatons.

THEORIES OF GROWTH

A library is a growing organism, so any general theory of growth should be relevant. An example of a principle in the theory of growth is the tendency towards inflexibility with increasing age. If this tendency were understood well enough, it might be possible to avoid it to some extent.

THEORIES OF EVOLUTION

Vocabulary seems to grow by a process of mutations and natural selection. The same is true of information retrieval and filing systems. A general theory ofevolution should perhaps be regarded as a part of a general theory of growth. In the process of evolution the punishment for inflexibility is especially severe.

RATIONAL BEHAVIOUR

So far I have largely ignored economic questions. In a complete theory, a guiding principle would be the principle of rational behaviour, of raising the expectation of value minus cost. This principle may seem too general to be useful, but I believe it could be used as a starting point. It leads, for example, to information theory (coding should be redundant enought to avoid costly mistakes) and to a realistic attitude

to the question of how much mechanization to do. It may also lead to an explanation of the above inflexibility principle. You become inflexible through attaching too little utility to the more remote future. On the other hand if an engineer, say, attaches too much utility to the remote future he would be so flexible that he would never complete the design of any machine.

HIERARCHIES OF MEMORIES IN A COMPUTER

In an electronic computer there are a number of memories of various sizes, costs per-element, and access times. The problem of optimal design is partly a question of determining the size of each memory so as to maximize their aggregate value for a given cost. A similar problem may arise in an information retrieval system. The different memories are, for example, cards, abstracts, books, books at other libraries, and photostats. It may be worthwhile to try to carry out a conscious optimization.

I have now finished dealing superficially with all the 20 headings on the list of potentially relevant theories. I have asked a lot of questions and have not answered many of them. Now perhaps other people would also like to ask a few questions that I won't be able to answer. [G. J. Rath mentioned, in the discussion, that much the same subjects were of interest in Systems Theory, and pointed out that the reason for the similarity is that the main unsolved problems of engineering today are problems of information handling.]