SAN JOSE RESEARCH LABORATORY

January 27, 1958
Reprinted March 16, 1960

FILE MEMORANDUM: FBW-16.2

SUBJECT: Information-Retrieval-Problems of the Distant Future

Abstract

This report represents the views of one member of the Information Retrieval Committee. It concentrates on the far distant future for the purpose of providing a perspective. It is assumed that other members of the committee are reporting on the more immediate problems that can be profitably attacked in the next few years. This report recognizes the historical phenomena of the rise and fall of great civilizations. It is the author's opinion that mankind has reached a new level of understanding of itself and nature such that our civiilzation is on the threshold of solving the international problems of the foreseeable future. The nature of future astronomical and geological crises which may confront humanity are discussed to give physical limits in a future perspective. The opinion is stated that the problems of "social responsibility" and "economic gain" are really the same when integrated over a long time, provided care is taken to establish a reasonable future perspective. The difficulty in realizing the economic gain from the areas of largest potential displaceable cost is the tremendous capital required to sustain research over the long time periods, unless intermediate overlapping problems are developed.

The far distant and most complex problems are discussed first, followed by problems closer and closer to the present as follows:

Future time unknown, earth crust computer system to predict future era of mountain-making; twenty years from now, world-wide economic homeostat; ten years from now, phoneme-associator for real time inter-lingual communication; five years from now, form associator. To continue in toward the present, refer to the collected reports of the other committee members.

F. B. Wood

F. B. Wood

Information Retrieval-Problems of the Future

Table of Contents

Title	Page
Introduction	3
Past and Future Partial Perspective (Glacial and Warm Ages)	3
Computation of Earth-Crust Activity to Predict Era of Mountain-Making	4
Alternative Paths for the Future Development of Civilization	5
Information Retrieval and Processing Problems of Twenty Years from Now (World-Wide Economic Homeostat)	6
Information Retrieval Problem of Ten Years from Now (Phoneme Associator)	7
Information Retrieval Problems of Five Years from Now (Form Associator).	8
Conclusions	10

Introduction

Many civilizations have risen to great heights of achievement and then have deteriorated. The great civilizations of the past have been analyzed by historians such as Toynbee, and sociologists such as Sorokin. These analyses tend to emphasize the cyclical rise and fall of different civilizations. I feel that the present crises in our civilization have one important difference from the crises that marked the decline of previous civilizations -- namely, we have the understanding and tools with which to solve the problems of our era. In spite of having two catastrophic wars so far in this century, we have a much better insight into how we human beings behave and into the nature of the feedback loops in our social, political, and economic systems.

The problems of communication, information retrieval, and simulation of social systems on computers offer keys to solution of the crucial problems of our civilization. The engineering problems of designing communication systems, information retrieval systems, and computers, are generally looked up as to their potential financial return and sometimes as to the aspect of social responsibility. It is my opinion that the problems of the greatest "social responsibility" will be found, in the long run, to have the largest financial return in terms of displaceable cost. This means we have to look for the intersection (or overlapping) of the areas of "social responsibility" and "optimum financial return".

Since these areas of maximum financial return probably would not pay off for many years, requiring an exhaustive drain on capital funds, a series of graded steps or stages must be found which proceed, in the direction of the optimum payoff problems, but which can realize a financial return in the intermediate stages.

Past and Future Partial Perspective (Glacial and Warm Ages)

It would be difficult to deal adequately with the sociological problems of our civilization, but there are certain limit points of significance which can be calculated from astronomical data. Through the evolutionary process of the development of the human race a series of glacial ages changed conditions over large parts of the earth such that in some parts of the world only the more intelligent variations were successful in adapting to changing weather conditions. The last glaciation was concluded at about 20,000 B. C. E. and the last warm period or melting of glaciers of a magnitude to cause great floods occurred about 7000 B. C. E. From calculations of the occurrence of the combination of suitable conditions of the eccentricity of the earth's orbit, the inclination of the axis of rotation of the earth, and the summer position of the earth relative to the sun's rays, it is possible to predict future warm periods and glacial periods. The next glacial age will be preceded by a period during which the earth will become much warmer, reaching a peak at about 20,000 A. D., when tropical forests will reach the Canadian Border.

We will have new glacial ages in about 50,000 A.D. and 90,000 A.D. in which much of North America and Europe will be covered with ice.

Now I propose to examine these long range problems briefly and work down to the more immediate problems to see if approaching the problems from this type of perspective can give us greater insight into our present day problems.* The coming warm period and its changes in climate and vegetation plus possible changes in sea level can make valueless many of our major cities causing either tremendous migrations of people on the earth's surface or requiring major engineering works to protect present population centers from these geological changes. Mankind will have to successfully solve the problems of international cooperation before these temperature changes procede too far, if the problem of relocation of vast populations is to be resolved in a humanitarian way.

Computation of Earth-Crust Activity to Predict Era of Mountain-Making

There is another threat to the human race which is not so easily predicted. The process of mountain formation due to the crust of the earth trembling and crumbling under accumulated stresses from the cooling of the earth. Forty million years ago, a gigantic outburst of volcanic and mountain-making activity resulted in the formation of the Himalayas, the Rockies, and the Andes. A second outburst about twenty million years ago elevated the Alps and the Cascade Range. It is probable that another cycle of tectonic activity (i.e. mountain forming) will occur. We do not have the tools nor the information to predict this future catastrophe. George Gamow**back in 1941 estimated it would take thousands of years to calculate the future behavior of the earth's crust if we had enough data. The trend of development of electronic computers may bring the computation time down to a reasonable value, provided ways of obtaining the geological data are developed. The human race may sometime have to set up a geological computing bureau to predict when and where the future mountain-making will occur and the effects upon sea level and land stability over the earth's surface. Will there be special places on the earth's surface that will be safe for human life during the future tectonic epoch, or will man have to move to the moon or to artificial moons during the peak of the earthquakes and mountain-making? I feel it is premature to accurately estimate the storage logic and data gathering system for a problem of this magnitude. However, in this new epoch of the biogeochemical history of the earth which some have named the "Noösphere", mankind has the potential skill to make geologically significant changes on the earth by which we can survive future geological catastrophes.

* To answer this question requires further study of this report, together with the reports of the other committee members.

** Biography of the Earth, p. 181

The consideration of problems as remote from the present as this one permits us to more easily evaluate the classification of projects as ones of "social responsibility" or of "economic value". A developing concept of the difference is that projects of social responsibility have large economic value after a long time delay, but that the time delay requires excessive capital. For an engineering organization to prepare for realizing a future economic gain from long-run projects which are now in the social responsibility area, a series of stages must be conceived where each stage in reaching the long-run goal has an area of economic gain by itself.

For an approximation to the memory size involved in the analysis of earthly crust activity, consider the following trial parameters:

> Consider a time span of 60×10^6 years in 100 year intervals or 6×10^5 times. For position use r, ϕ, Θ as follows: r in 2-mile intervals in binary notation gives $2^{12} = 4096$ miles or 2048 radial position,
> ϕ in 180/8 degree sectors $2^3 = 8$ ϕ position
> Θ in 360/64 degree sectors $2^6 = 64$ Θ position

The number of storage slots is then

$$8 \times 6 \times 10^5 \times 2048 \times 64 = 6.3 \times 10^{11} \text{ slots}$$

The word length for each slot is derived as follows:

10 bits significant figures
1 bit sign — 18 binary digits per parameter
7 bits exponent

Eight words per slot from one index word plus seven variables such as T, dt/dx, dT/dt, t, v, p, etc. Then allowing three trials in active storage gives $3 \times 18 \times 8 \times 6.3 \times 10^{11} = 2.7 \times 10^{14}$ bits of storage. When the above problem becomes more urgent to the human race, it will probably be simplified. This problem is not strictly an information retrieval problem, but is more of an ordered scientific computation. It is described in order to illustrate the magnitude of future problems that the human race may face in the distant future.

Alternative Paths for the Future Development of Civilization

The result of past cycles of glacial periods and the intervening warm periods of the Pleistocene glacial epoch is that after each glacial period there evolved a more highly developed race of men. This is probably due to the survival of only the most skilled remnants of the tribes whose homelands became cold from the advancing glaciers. After Gunz's Glacial Period (600, 000 B. C. E.)

the Piltdown, Peking, and Java Man of 985 cc brain size appeared; after Mindel's Glacial Period (450,000 B.C.E.) the Heidelberg Man appeared. After Riss's Glacial Period (200,000 B.C.E.) the Neandertal Man appeared; and after the Würms Glacial Period (120,000 B.C.E. to 20,000 B.C.E.) the Cro-Magnon man appeared, followed by Modern Man with a brain of 1400 cc.

The development of electronic computers and modern communication systems enables man to collect data and make calculations with the data faster than ever before, so that man is not dependent upon evolving a larger brain over centuries of evolutionary development. The data processing equipment frees man's brain for dealing with abstract representations of the problems of the world.

When crises develop such as glaciers, floods or tremendous earthquakes, destroying the usefulness of large populated areas, the people of the world can either fight it out to determine who gets the remaining useful land area, or they can set up an international agency of the United Nations to arrange for an equitable sharing of the inhabitable land. It is probable that a sharing of the available land area would result in a net gain to the people of the world in that their energy devoted to military preparation could be diverted to more productive work. This leads me to consider moving attention to a shorter time into the future to see if there is an area of large displaceable cost toward which information retrieval research could be directed. Thus, we shall look at some information retrieval problems in reverse order* for twenty years from now, ten years from now, and five years from now.

Information Retrieval and Processing Problems of Twenty Years from Now (Economic Homeostatic System)

Twenty years from now we can expect that a higher level of international cooperation will be developed through the agencies of the United Nations. The advance in international cooperation would be dependent upon two major factors; (1) a better understanding of how to meet mankind's economic needs in a stable way, and (2) a higher level of understanding of man's psychological needs.

The first item offers the possibility of a world-wide economic information processing system with a retrieval system for historical data on economic cycles and for data/economic potentialities of different countries.

on

If we develop a world-wide information processing system which works under the United Nations as a vast store of world economic conditions

*See arrangement of book Out of Revolution by Eugen Rosenstock-Huessy.

on
and information/the needs of the peoples of the different countries around the world, it would be possible to solve a problem that is both financially remunerative and fulfilling a social responsibility.

In 1956, the Defense Department and Mutual Assistance Budget of the United States was $42 billion out of a total Federal Budget of $66 billion. If the world economic condition could be stabilized so that the nations of the world felt more secure and consequently the military budget could be cut perhaps ten per cent or $4.2 billion per year, it would be conceivable that a one-billion dollar per year rental could be available to pay for the United States share of the computer retrieval system, and communication link of the UNESCO world-wide homeostat. This would leave a $3.2 billion saving in the Federal budget per year.

The idea of a homeostat or ultra-stable system has been developed by W. Ross Ashby. The basic idea is that the computer system would compute contour planes in systems of n-variables (economic) for different values of the (n + 1) st variable. These (n + 1) st variable representing step-function arising from possible decisions of industries, countries, and other institutions to change interest rates, foreign exchange rates, tariffs, close-down or open factories or agricultural units, etc. A simple example of a step-function and an n-3 system is illustrated in Reference* page 87. The computing system would show stability points in respect to proposed decisions. A slightly different but more general description of the possible process is given in Reference **. This problem is a combination of information retrieval and scientific computing. A rough estimate of the memory required is about 3×10^9 records of 100 characters or 2×10^{12} bits. The system could probably be broken up into tree-like subsections to materially reduce the total storage.

Information Retrieval Problem of Ten Years from Now (Phoneme Associator)

The possibility of real time voice-to-voice language translation leads to the consideration of translation by phoneme.

The most general analysis of the phonemes or distinctive sounds of language is reduced to twelve binary attributes or $2^{12} = 4096$ possible phonemes. The phoneme pattern of English can be reduced to nine attributes*** Only 40 of

* W. Ross Ashby, Design for a Brain (1952)

** W. Ross Ashby, "Design for an Intelligence-Amplifier" Automata Studies

*** Colin Cherry, On Human Communication (1957) p. 95

the possible 512 phonemes are used in English. Most languages use only a few dozen of the possible phonemes.

The proposed system would sample the voice input in a manner similar to the six channel system built by Melpar, Inc.* A logic system would then sample the six analog signals and code them into the nine attributes. The next logic stage would test for the significant combinations of attributes, i.e., vocalic/non-vocalic, consonantal/non-consonantal, compact/diffuse, etc., and code in a binary representation of the phonemes. The international phonetic symbols would be used as phoneme designators.

The dictionary for the translation process would be alphabetical by the phonetic alphabet so this problem becomes more of a table look-up system, rather than a retrieval system. This system might have two variations: (1) a literral translation mode: word for word by table look-up operating in real time and (2) a precise translation mode operating on a time delay basis from recorded voice input. The second mode of operation would have a more complicated feedback logic system for checking which of multiple meanings applied to the particular speech.

Table of Dictionaries Required

No. of Language	No. of Dictionaries
2	2
3	6
4	12
5	20

A four-language system would require: 100,000 words/language x 4 transl. wds. /word x 12 (dictionaries) x 6 phoneme/word x 9 binary bit/phonen = 2.4 x 10^8 bits (not including grammer rules).

Information Retrieval Problems of Five Years from Now (Form Associator)

Consider three levels of mechanization of the information retrieval process as shown in Figure 1. Stage (A) represents the present steps in a retrieval system. Stage (B) assumes that 9-character recognition system is developed to replace the human reading and translation to binary code. Stage (C) represents a possible system requiring a form associator (logic) and a larger memory by/order of magnitude larger than/binary memory system.
an a

*S. J. Campanella and T. E. Bayston "A Continuous Analysis Speech Bandwidth Compression System", Abstracts of Technical Papers IRE, Third Aero-Com Symposium, 1957 pp. 10-12 (S.J. Research Library. PAM 2198)

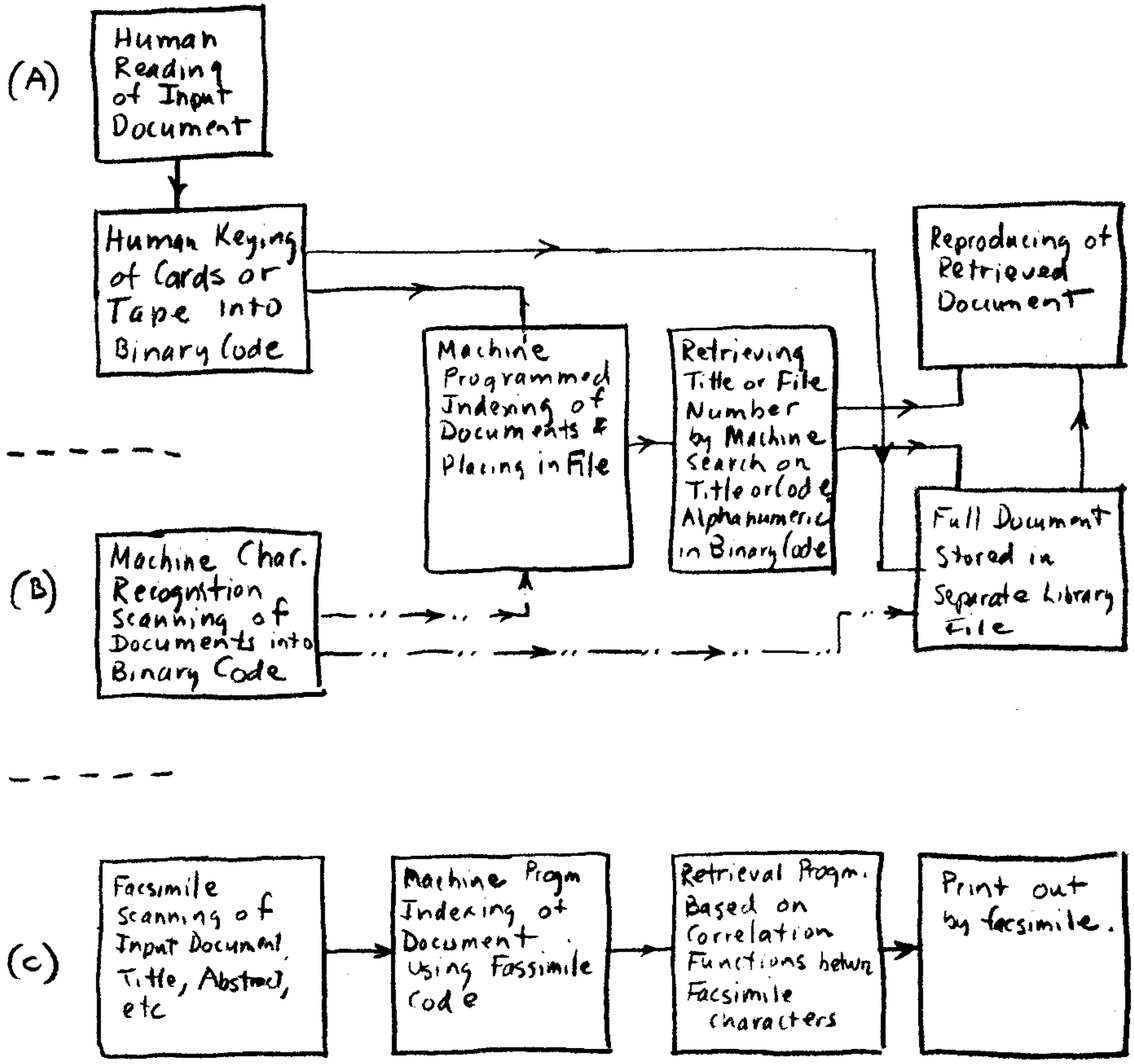
(A)
Human Reading of Input Document
Human Keying of Cards or Tape into Binary Code
Machine Programmed Indexing of Documents & Placing in File
Retrieving Title or File Number by Machine Search on Title or Code Alphanumeric in Binary Code
Reproducing of Retrieved Document
Full Document Stored in Separate Library File
(B)
Machine Char. Recognition Scanning of Documents into Binary Code
(c)
Facsimile Scanning of Input Document, Title, Abstract, etc
Machine Progm Indexing of Document Using Fassimile Code
Retrieval Progm. Based on Correlation Functions betwn Facsimile characters
Print out by facsimile.

Fig. 1 - Comparision of Stages in Retrieval System, for Human Keying (A); Machine Character Recognition (B); and Form Associator (C).

In this system facsimile scanning is used to read printed or typed characters as a way to bypass the character recognition problem at the expense of larger memory and more logic in the retrieval process.

The function of such a facsimile (or form) associator is two fold:

(1) To permit searching for forms such as chemical bond diagrams, special symbols, circuit elements, Chinese, etc. It could be expanded to larger matrices for direct fingerprint correlation.

(2) To bypass the character recognition problem on reading the input documents.

The logical form of this Form Associator would be groups of 10, 20, etc. tracks in parallel which would use optical disks, magnetic disks, etc., with ganged heads.

Compared with the present RAMAC of 3.5×10^7 bit storage an equivalent disk file for the form associator could require 3.5×10^8 bits storage. The increase in the logic and accumulator elements for performed the correlation functions compared to present RAMAC logic is approximately:

$$\frac{7 \times 10 \text{ (matrix bits)}}{7 \text{ (character bits)}} \times 4 \text{ (x-registration)} \times 4 \text{ (y-registration)} = 160.$$

The four unit variation in both the x and y direction for registration is proposed to allow for a range of registration scale out of line.

This form associator may lead to a preliminary test of phoneme association direction from sonograms.* This type of form associator would require a $5000/300 = 16$ (ΔF) by 12 (Δt) matrix of points per character. This form would be approaching an analog recording system. Binary representation of alphanumeric data takes six or seven bits per character. The facsimile representation take about 70 bits per character, while the sonogram type approaches 200 bits per character.

Conclusions

The brief analyses of future information retrieval problems are initial approaches which require review by consultants in the appropriate fields of geology, economics, and linguistics** before they can be accepted as reliable. Even though this material has not yet been verified, it has a value is establishing a perspective to guide the planning of more immediate engineering research.

* Colin Cherry, On Human Communication, p. 145

** Including people in established IBM projects, such as the speech recognition work in the Information Research Group at Yorktown.