International Federation of Library Associations and Institutions
UNIVERSAL DATAFLOW AND TELECOMMUNICATIONS
CORE PROGRAMME

OCCASIONAL PAPER

5

THE SEVEN AGES OF INFORMATION RETRIEVAL

Michael Lesk
Bellcore

IFLA

March, 1996

International Federation of Library Associations and Institutions
UNIVERSAL DATAFLOW AND TELECOMMUNICATIONS
CORE PROGRAMME

The IFLA Core Programme on Universal Dataflow and Telecommunications (UDT) seeks to facilitate the international and national exchange of electronic data by providing the library community with pragmatic approaches to resource sharing. The programme monitors and promotes the use of relevant standards, promotes the use of relevant technologies and monitors relevant policy issues in an effort to overcome barriers to the electronic transfer of data in library fields.

CONTACT INFORMATION

Mailing Address:

IFLA International Office for UDT
c/o National Library of Canada
395 Wellington Street
Ottawa, CANADA
K1A 0N4

UDT Staff Contacts:

Leigh Swain, Director
Email: leigh.swain@udt.ifla.org
Phone: (819) 994-6833

or

Louise Lantaigne, Administration Officer
Email: UDT@udt.ifla.org
Phone: (819) 994-6963

Fax: (819) 994-6835

Email: UDT@udt.ifla.org

URL: http://www.ifla.org/udt/

Occasional papers are available electronically at: http://www.ifla.org/udt/op/

UDT Occasional Papers # 5

Universal Dataflow and Telecommunications Core Programme
International Federation of Library Associations and Institutions

The Seven Ages of Information Retrieval

Michael Lesk
Bellcore
lesk@bellcore.com

March, 1996

ABSTRACT

Vannevar Bush's 1945 article set a goal of fast access to the contents of the world's libraries which looks like it will be achieved by 2010, sixty-five years later. Thus, its history is comparable to that of a person. Information retrieval had its schoolboy phase of research in the 1950s and early 1960s; it then struggled for adoption in the 1970s but has, in the 1980s and 1990s, reached acceptance as free-text search systems are used routinely. The tension between statistical and intellectual content analysis seemed to be moving towards purelyg statistical methods; now, on the Web, manual linking is coming back. As we have learned how to handle text, information retrieval is moving on, to projects in sound and image retrieval, along with electronic provision of much of what is now in libraries. We can look forward to completion of Bush's dream, within a single lifespan.

INTRODUCTION.

Shakespeare described seven ages of man, [Shakespeare 1599] starting from infancy and leading to senility. The history of information retrieval parallels such a life. The popularization of the idea of information retrieval started in 1945, with Vannevar Bush's article (still cited 96 times in the 1988-1995 Science Citation Index). [Bush 1945] And, given the current rate of progress, it looks like it will finish by 2015 or so, the standard life-span for someone born in 1945. By that time, most research tasks will be performed on a screen, not on paper.

There has, however, been a tension throughout the life of information retrieval between simple statistical methods and sophisticated information analysis. This dates to a memo by Warren Weaver in 1949 [Weaver 1955] thinking about the success of computers in cryptography during the war, and suggesting that they could translate languages. He wrote “it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the ‘Chinese code.” If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?” Just as Vannevar Bush's paper began hypertext, Warren Weaver's paper began research in machine translation.

In some ways, information retrieval has two parents. Bush reacted to the success of physicists in projects such as microwave radio (and of course the atomic bomb, which he could not yet mention publicly). He talked of intellectual analysis, both by people and by machines. Weaver reacted to the success of mathematicians in cryptography. And he thought in terms of simple exhaustive processing, not in terms of high-powered intellectual analysis. This is the same dichotomy familiar in computer chess between the “all-legal-moves” style and the “predicting good moves” researchers; artificial intelligence against statistics. Throughout the history of information retrieval, we have seen both techniques. The analytical process, the Bush approach, can either use manual indexing or try for artificial intelligence programs that will achieve the same accuracy of information identification. The accumulation of statistical detail, Weaver's approach, can be done entirely mechanically with probabilistic retrieval techniques. Of course, manual indexing pre-dates computers and information retrieval technology.

CHILDHOOD (1945-1955)

Rereading the original Bush paper, and looking at it from today's standpoint, the hardware seems mostly out of date, but the software goals have not been achieved. Bush, of course, did not realize the progress that would be made in digital technology, or in semiconductors, and wrote about bar coded microfilm and instant photography. We can do better today than what he described; we have digital recording and Polaroid film. On the software side, by contrast, he expected speech recognition; we don't have it yet.

Figure 1 shows some of the predictions Bush made in his paper, and where we now stand.

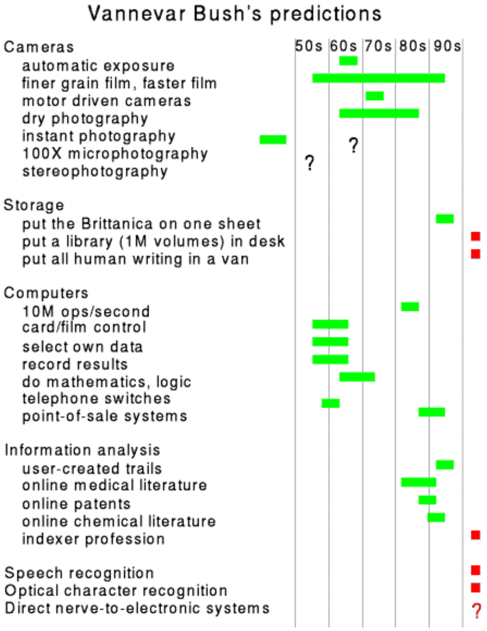


Figure 1

Note that we have all the photographic inventions, but they have been less important than he expected for information processing. For example, he thought that ultramicrofiche (100X reduction) would be particularly important. In practice, if a library can save 95% or move of its shelf space costs by going to 20X reduction, and that doesn't justify microfilming, saving a few more percent isn't going to help. For example, the physical cost of printing a typical book is about 15% of its cover price. Even if it could be reduced to zero, that would not permit a library to double its purchases. And the difference between reducing the printing cost to 1% of the book cover price and half a percent of the cover price is insignificant.

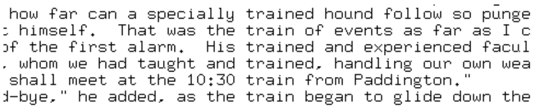
Bush also talked about automatic typing from dictation and OCR. Neither of those has quite been achieved yet. By contrast, most of his specific predictions for computers were achieved in the 1960s: computers selecting their own data, performing complex arithmetic, recording their results for later use, symbolic mathematics, and computerized telephone switching systems. Surprisingly, all these organizational goals were achieved before his specific performance goal, machines doing 10M operations/second.

His storage predictions are not quite met yet. He wrote of a user inserting 5000 pages per day into a personal repository and it taking hundreds of years to fill it up. At 30 Kbytes/page (he was talking photography, after all), with 250 working days per year, in 200 years the user would have accumulated 7.5 terabytes. As I write, storage in the software research area at Bellcore is about 2 GB per person; a particularly profligate user (such as me) has 20 GB, but still well short of a terabyte. Even if OCR worked, and the 30Kbytes per page were reduced to 3 Kbytes, we're still not there in terms of storage capacity per person.

Most important is the design of the user interface that Bush predicted. He emphasized individual interfaces, with particular trails through information space that would be personalized to the user, and traded among people. Until recently, the entire development of information retrieval was aimed at uniform interfaces, that reacted the same way to each user, and which did not depend on an individual's organization of personal files. This has meant, for example, that most people have much better searching tools available for the printed literature than for their own personal notes. By contrast, when most scientists or engineers want information, they look first in the closest sources. They consult personal notes, ask the nearest colleague, ask their supervisors or other local authorities, and only much less frequently do they look in journals. Asking a librarian is a desperate measure. [Schuchman 1981]. Bush envisaged a system that first of all, would provide the individual notes and memories of the scientists. Only later did he suggest that people would buy conventional publications and insert them into the Memex. This would have been a better match to what people need than the direction information services actually went, but it was so difficult before the 1970s to get information into computers that the kind of system Bush wanted was not practical.

In the 1950s, however, the Soviet Union sent up the first artificial Earth satellite. The success of Sputnik I prompted widespread fears that the United States was falling behind in science, and a realization that there was relatively little knowledge of Russian science in the United States. In addition to funding Russian language studies and machine translation research, Sputnik produced a desire to improve the efficiency of scientific information transfer. Stories that some company or other had spent $100,000, or perhaps $250,000, redoing some research because they could not find the original circulated as urban legends but helped to justify research into information retrieval.

The very first systems were built in the 1950s. These included the invention of KWIC indexes, concordances as used for information retrieval, by such researchers as H. P. Luhn. Figure 2, for example, is a part of the list of occurrences of “train” in the Sherlock Holmes stories.



```
 how far can a specially trained hound follow so pūnge
: himself.  That was the train of events as far as I c
of the first alarm.  His trained and experienced facul
, whom we had taught and trained, handling our own wea
 shall meet at the 10:30 train from Paddington."
d-bye," he added, as the train began to glide down the
```

Figure 2

Concordances are still used where a small text is to be studied intensively; they are too bulky to be printed for any kind of large collection. However, the simple "every occurrence of every word" kind of processing they represent survives today in many commercial retrieval systems.

There were also innovative attempts at alternate kinds of machinery, such as the use of overlapping codes on edge-notched cards by Calvin Mooers. The most famous piece of equipment of this period was the WRU Searching Selector, built by Allen Kent at Western Reserve University (now Case Western Reserve). All of this special purpose technology was swept away by digital systems, of course. I could no more find edge-notched cards today than I could find IBM punched cards.

THE SCHOOLBOY (1960s)

Shakespeare's next time of life is the schoolboy, and in fact the late 1950s and 1960s were a time of great experimentation in information retrieval systems. This period also saw the first large scale information systems built. Many of the current commercial library systems such as Dialog and BRS can be traced back to experiments done at this time. The early 1960s also saw the definition of recall and precision and the development of the technology for evaluating retrieval systems. It also saw the separation of the field from the mainstream of computer science. In fact, the 1960s were a boom time for information retrieval. More people attended the IR conferences then, for example, than have attended the SIGIR conferences in the last decade.

The first experiments that were done, with such systems as NASA's RECON, used straightforward and simple technology. They largely used mechanical searching of manual indexing, keying in the traditional indexes that had been printed on paper, but were much easier to search by machine. In particular, it is difficult with traditional indexes to do the “and” operation. Users wishing to find all articles in Chemical Abstracts which mention two particular compounds must look up both and then try to find the abstract numbers mentioned in both lists. This is something computers can do very well compared with people. Computers can also deal with quantity well. While the number of index terms assigned to each document in a database which is manually indexed and printed on paper must be limited to prevent the printed version becoming unwieldy, with computers this does not matter. Thus, the advent of computerized databases meant that indexing could become more detailed, although this promptly meant that indexing could become too expensive to do properly.

Thus, the idea of free-text searching arose. There could be complete retrieval of any document using a particular word, and there would be no cost for manual indexing. This idea became and has remained very popular, but immediately also attracted objections from those who pointed out that indexing was not just a way of selecting words but was a way of choosing the right words, either the words that were more important, or the ones that had been selected as the correct label for a given subject. The latter was enforced, in many indexing systems, by official vocabularies, so that people would not be frustrated by finding one document indexed with the word “cancer” while another was indexed “neoplasms.” Thesauri such as MeSH (Medical Subject Headings) or the Inspec Thesaurus had attempted to standardize nomenclature.

This created several questions. Was free-text searching of acceptable quality? If controlled vocabularies were preferable, was there a way of translating to them or creating them automatically? These problems stimulated the development of evaluation techniques, led by Cyril Cleverdon of the Cranfield College of Aeronautics (now Cranfield Institute of Technology). Cyril developed the mathematics of `recall' (fraction of relevant documents retrieved) and `precision' (fraction of retrieved documents that are relevant) as measures of IR systems, and built the first test collections for measuring them. There is a tradeoff between the two measures: if a system simply retrieves more documents, it is likely to increase recall (with more retrieved documents, a larger fraction of the relevant ones are likely to appear) but to decrease precision (since there are also more opportunities in a longer list of retrieved documents to retrieve non-relevant material). Figure 3 shows one of the early recall-precision graphs, taken from the Aslib Cranfield project report. [Cleverdon 1966].

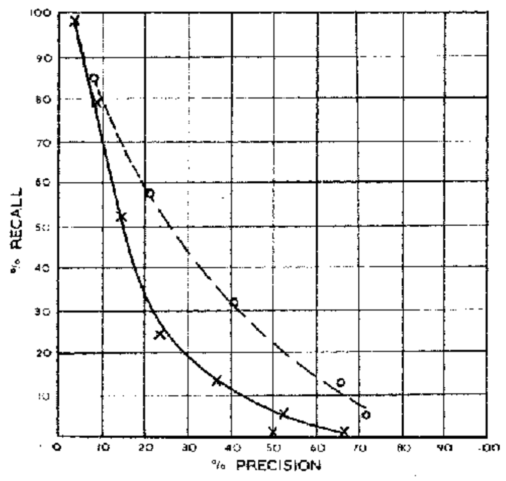


FIGURE 4. 814P INDEX LANGUAGE III. 5. a SEARCH E
200 DOCUMENTS
(Index Language III. 1. a Broken line)

Figure 3

A series of experiments followed, run not only by Cyril and his co-workers (notably E. Michael Keen, now at the College of Librarianship Wales in Aberystwyth and F. Wilf Lancaster, now at the University of Illinois) but also by the late Gerard Salton, at both Harvard and Cornell. They showed that the value of manual indexing was low, if any, at least on the small test collections being used. Free-text indexing was as effective, and implicitly a great deal cheaper. Efforts to see whether terms could be assigned to thesaurus categories automatically or whether better thesauri could be made, following a variety of methods. A long series of experiments was run, and in fact the test collections created at that time are still in use, even though the largest was only 1400 abstracts of aeronautical documents. [Cleverdon 1970].

As these experiments were being done, new retrieval techniques were invented. Key among these was relevance feedback, the idea of augmenting the user's query by adding terms from known relevant documents. This was one of the most effective techniques, since a problem from that day to this in retrieval experiments is that users often specify very short questions, and the systems can do better with longer queries. Another early experiment was multi-lingual retrieval, using a bilingual thesaurus and mapping words from both languages into the same concepts. [Salton 1968].

All these experiments started to establish information retrieval as an unusual software discipline, in which evaluated experiments, statistical methods, and the other paraphenalia of traditional science played a major role. In contrast, the 1960s also was the start of research into natural language question-answering. Artificial intelligence researchers wondered why retrieval systems had to be limited to retrieving documents which the user would still have to read to find the answers to questions. Researchers such as Terry Winograd, Daniel Bobrow, and others began building systems in which actual answers were retrieved from databases, while others looked at the problems of syntactically analyzing documents to either detect data elements or do phrase matching for retrieval. At this stage of development, these systems were fragile and ran only on a few examples. The information retrieval community had run its own experiments on the use of phrases, for example, and tended to reject these devices.

Throughout this period of the 1960s, there was relatively little actual computerized retrieval going on. Most of the work was still research and learning, and ambitious though some of it was in principle, there was no access to really large amounts of machine-readable text with which to build large systems. This was, however, about to change.

ADULTHOOD (1970s)

During the 1970s retrieval began to mature into real systems. Perhaps the most important cause was the development of computer typesetting, and later word processing, which meant that lots of text was now available in machine-readable form. Computer typesetting started in the mid 1960s, and even earlier than that there were paper-tape driven Monotype machines whose input tapes could be converted for use by computers, but the large scale adoption of cold type came in the 1970s. By the end of the decade both hot lead and conventional typewriting were obsolete, and most material that was printed was passing through a stage of computer input. This provided piles of material for retrieval systems.

The other key technology motivator was the availability of time-sharing systems. Instead of queries being processed in some background batch operation, it was now possible to present them directly from a terminal and get an answer immediately. This made retrieval much more practical, and systems developed to offer such services to librarians. Time-sharing, in turn, became possible as disk drives became cheaper. Figure 4 shows the price of disk space by time; as much as anything, this chart has driven the information retrieval industry.

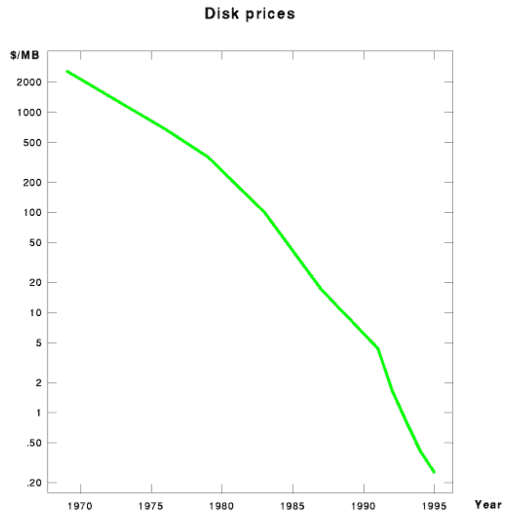


Figure 4

Among the earliest large-scale systems were commercial systems such as Dialog, Orbit and BRS. These fed off the computerization of abstracting and indexing services, which were early adopters of computer typesetting. At Chemical Abstracts, for example, the delay of combining many feet of yearly indexes into a five-yearly cumulative index was becoming intolerable; done manually, it looked like this might take longer than five years to do. Computerization made it faster and easier to combine first the weekly indexes into a yearly cumulation and then the yearly volumes into the five-yearly Collective indexes. The tapes used to do this could then be provided to services such as Dialog for access by professional librarians.

Another early system was OCLC, the Online Computer Library Center (although at first it expanded its acronym into Ohio College Library Center). Founded by Fred Kilgour, it used the output of the Library of Congress MARC program for machine-readable cataloging. The Library of Congress cataloged books; OCLC acquired tapes of this cataloging; and it then printed catalog cards for member libraries, saving them the work of preparing these cards themselves. OCLC also introduced the idea of cooperative work. The Library of Congress catalogs about 1/3 of the book titles bought by the OCLC member libraries, and about 2/3 are not cataloged by LC. These represent foreign publications and items (such as technical reports, dissertations, and so on) which are not conventional book publications. Member libraries cataloged these books themselves, provided the cataloging to OCLC, and other libraries could then access that record. This is, as far as I know, the first large-scale example of computerized cooperative intellectual organization of information, as Bush had envisaged.

All of these systems used relatively simple and limited searching systems. The OCLC search specification, for example, was the first four letters of the author's name and the first four letters of the title. The online systems were better, but they still limited themselves to Boolean searching on text-word matching. Since they often had index terms as their data content, they were in fact using manual content analysis, although often the abstracts were searched more than the index terms.

Although most of these systems operated off abstracting or indexing, this decade saw the start of full-text retrieval systems. Lexis, later absorbed by Mead Data Central (and more recently by Reed-Elsevier), provided the complete text of court decisions. Other early systems were derived from newspapers, among the first non-reference publishers to go to computerized typesetting: they included the Toronto Globe and Mail and the New York Times Information Bank.

During this decade, the kind of research that had been done in the previous decade went into a decline.

The 1970s and 1980s were a period when both databases and office automation research boomed; one might have thought that information retrieval, situation between the two areas, would have done well, but in fact it lost out. Part of this represented government politics; NSF, for example, had funded quite a bit of information retrieval research in the 1960s, but now started to think about what it meant to be responsible for the availability of scientific information. Did that mean that NSF should be funding research in the distribution of information, or the actual distribution of information? In practice some money was shifted away from research, and in addition there was some disillusion with promises that had been made in the 1960s and that did not seem to be coming true fast enough. Arthur Samuel, for example, wrote in 1964 that by 1984 libraries would have ceased to exist, except at museums. [Samuel 1964]. Other promises, such as those associated with machine translation, had also perhaps encouraged funders to look elsewhere. The rise of computer science departments also perhaps had negative effects on IR research, as it often did not fit into the categories of highest importance in those departments, and in the earlier anarchy it was perhaps easier for oddball researchers to find a place.

There was some research progress, of course, and perhaps the most important was the rise of probabilistic information retrieval, led by Keith van Rijsbergen (now at the University of Glasgow). This involved measuring the frequency of words in relevant and irrelevant documents, and using term frequency measures to adjust the weight given to different words. Although these algorithms had computational problems on the equipment of the day, a simpler variant (term weighting) was one of the few techniques that seemed to improve performance over the simple word matching that was prevalent.

What about the AI researchers who were trying to do intellectual analysis automatically? The problems of overpromising results for machine translation and computational linguistics affected them as well, and the key subjects in the 1970s were speech recognition and the beginning of expert systems. Expert systems, in particular, occupied the same niche in buzzword space that “intelligent agents” do today. Programs were going to go out and retrieve data in the way that a librarian did. For example, one author wrote “The 1980s are very probably going to be the era of the expert system, and by the end of the decade it is possible that each of us will telephone an expert system whenever we need to obtain advice and information on a range of technical, legal, or medical topics.” [Peat 1985]. In practice, there was a split between the AI community and the IR community. The AI researchers felt that they were attacking more fundamental and complex problems, and that there would be inherent limits in the IR string-searching approach. For example, I have an easy time looking for my name in retrieval systems; it is not an English word nor is it a common name, and when I search for it and find a hit that is not me it is most likely to be either my brother or my second cousin. But I worked once with a man named Jim Blue, and you can imagine the problems looking for his name in a dumb database. Even worse, consider the chemists named Bond or Gold. The AI researchers felt that without some kind of knowledge understanding these problems were insoluble. Meanwhile, the IR camp felt the AI researchers did not do evaluated experiments, and in fact built only prototypes which were at grave risk of not generalizing.

The best-known demonstrations of the AI systems were from the group at Yale led by Roger Shank (now at Northwestern University). These programs mapped information into standard patterns. For example, a typical medical drug evaluation experiment can be imagined as “we took some number of rats, suffering from disease X, and gave them the following quantities of the drug being tested, and the following numbers of rats at each dosage were cured.” Shank's group constructed such schemas for a number of common activities, e.g. ordering in restaurants, and then took natural language descriptions of these activities, picked out the information that appeared to fit slots in the frames, and thus constructed a semi-formal representation of the information. They could then take queries about such subjects, e.g. vehicle accidents on the United Press newswire, and retrieve actual answers. These programs ran on a restricted set of examples, and produced much argument about whether they were in the end going to develop into practical retrieval systems. Some of these systems, such as the LUNAR system of Bill Woods or the Transformation Question Answerer (TQA) of Stan Petrick and Warren Plath were evaluated, but they tended to operate off databases rather than text files. [Schank 1973].

Part of the work in AI was the elaboration of languages designed for the precise representation of information, called knowledge representation languages. These languages attempted to use logical notation to represent general knowledge, so that it could be processed by programs. Several such languages were defined, and their proponents suggested that large amounts of information could be written in them and used by expert systems. Unfortunately, as with earlier AI work, there were great arguments as to the reality of these proposals.

MATURITY (1980s)

During the 1980s, the steady increase in word processing and the steady decrease in the price of disk space meant that more and more information was available in machine-readable form and was kept that way. The use of online information retrieval expanded in two main ways. One was the availability of full-text instead of just abstracts and indexing; the other was the spread of online retrieval into use by non-specialists, as libraries replaced or supplemented their card catalogs with online public access catalogs. Meanwhile, the IR research community began to come back, with new work on term analysis and related areas. Artificial intelligence, boomed and then crashed. And the development of the CD-ROM provided an entirely new route for information delivery, which seemed for a while to threaten the whole idea of cooperative online work.

There was an enormous increase in the number of databases available on the online systems. Figure 5, from the Gale Research Directory edited by Martha Williams, shows the steady increase in the number of online databases. Not only did the number increase, but new kinds of databases, involving full text and numerical data, appeared on the online services.

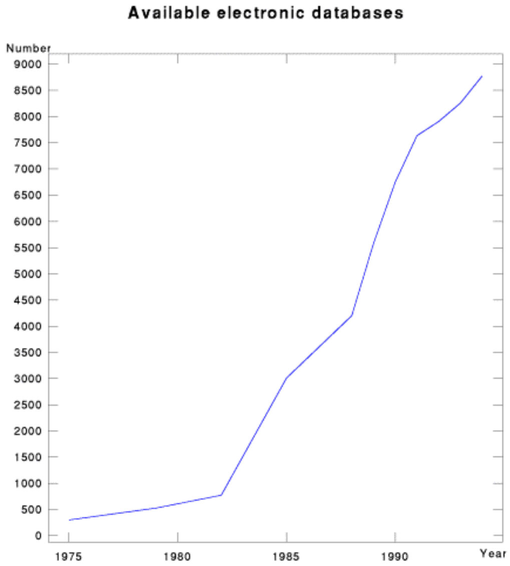


Figure 5

OPACS also developed during the 1980s. Many libraries, thanks to OCLC and similar cooperatives such as the Research Libraries Group (RLG), had full machine-readable records from the mid 1970s. They began making search systems available, and some even converted their historical catalogs to machine-readable form and provided full online catalogs. This included the Library of Congress with the REMARC project, providing an extremely large and complete book file. By the end of the decade, commercial vendors provided OPAC software (eg NOTIS and GEAC).

Full text online blossomed in this decade. Many current magazines and newspapers were now online, albeit with text only. The full text of American Chemistry Society journals, for example, was provided by the online system originally named Chemical Journals Online (CJO) and now called Science and Technology Network (STN).

In the research community, a resurgence of effort on information retrieval techniques started. The new material which was available, and the interest in smaller programs that would run on individual workstations, produced new research in retrieval algorithms in places like the University of Massachusetts (Bruce Croft) and Virginia Tech (Ed Fox). Although some were still looking at the mathematics of the vector space model defined by Salton in the 1960s, there was increasing interest in new kinds of retrieval methods. For example, research on sense disambiguation using machine-readable dictionaries, as a way of helping distinguish different meanings of the same words (the familiar problem of distinguishing “blind Venetian” from “Venetian blind”). There was work on bilingual retrieval, and on part of speech assignment. In fact, after many years of frustration, it began to appear that some computational linguistics might be useful in retrieval, but it was going to be the statistical kind of retrieval, rather than the AI approach.

Few of these techniques, however, were used in any of the commercial systems. Those went along with the same kinds of retrieval algorithms that had been used for decades now. In fact, in some ways the commercial systems appeared not to care very much about retrieval performance. If one looked at their advertisements, they boasted primarily about the number of databases they had, secondarily about their user interface, and not at all about retrieval performance. Part of the problem was that the whole retrieval evaluation paradigm was based on small test collections, and there were almost no evaluations of full-scale systems. In addition, the evaluation work had shown that there was huge scatter across questions in the performance of retrieval systems, so that it was not possible to offer any confidence, in talking about the performance of a system, that any particular question would be close to some average performance level. Whatever the reasons, the research community felt frustrated that their techniques were not being employed in the industry they had started.

The AI community continued expert systems and knowledge representation languages. One of the most ambitious plans was the CYC system of Doug Lenat at MCC; this was an attempt to encode all “common-sense” knowledge in a formal structure, so that a computer could do ordinary reasoning in any subject areas. This was part of an effort to respond to the “Fifth Generation” computer being studied by the Japanese, a machine that would specialize in logic and symbolic processing. In the early part of the 1980s there was great enthusiasm for expert systems and knowledge-based projects. Later in the decade, however, the general trend away from support for basic research plus the failure of expert systems to deliver on their initial promises (or over-promises) caused a movement away from this area, sometimes called “AI winter.” In fact, far from believing that any knowledge expressed in language can be translated into a single formalism, it is probably now respectable again to mention the Whorfian hypothesis which suggests that the specific natural language used to express information constrains the kind of information which can be described, and is really a part of that information.

A key technology change was the widespread use of the CD-ROM. By the end of the decade, most libraries had at least one CD-ROM drive, and CD-ROMs were being used regularly to distributed information. They fit the traditional publishing paradigm very well (stamped out, put in the mail, and hard for the users to copy), so they could be economically distributed and easily used. Their large size for the time (650 MBytes, with a very low production cost) meant that big databases, appropriate even for full-text files, could be distributed. Computer networking also continued to develop in this decade, but the CD-ROM fit so well with traditional information publishing economics that it would develop into a real threat to the online systems that had been growing rapidly for two decades; see Figure 6, with data collected by Martha Williams.

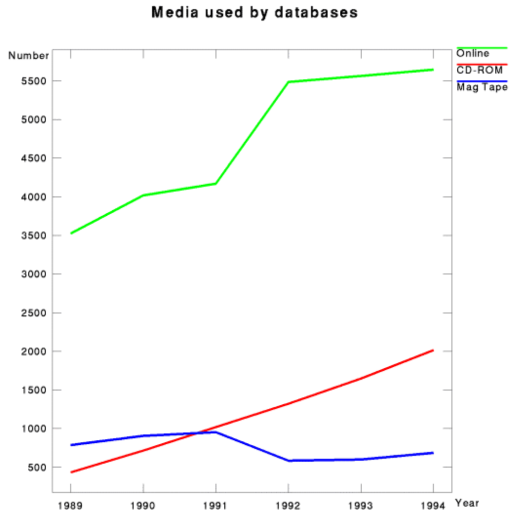


Figure 6

On balance, this was a decade in which online information became common, even if still not used by most people most of the time; and in which it appeared the Weaver followers, users of statistics, had basically routed the Bush followers, urging detailed content analysis.

MID-LIFE CRISIS (1990s)

By 1990 information retrieval could be said to be 45 years old, and it was time for a mid-life crisis. Things seemed to be progressing well: more and more text was available online, it was retrieved by full-text search algorithms, and end-users were using OPACs. Figure 7 shows the sales of online vendors in 1993.

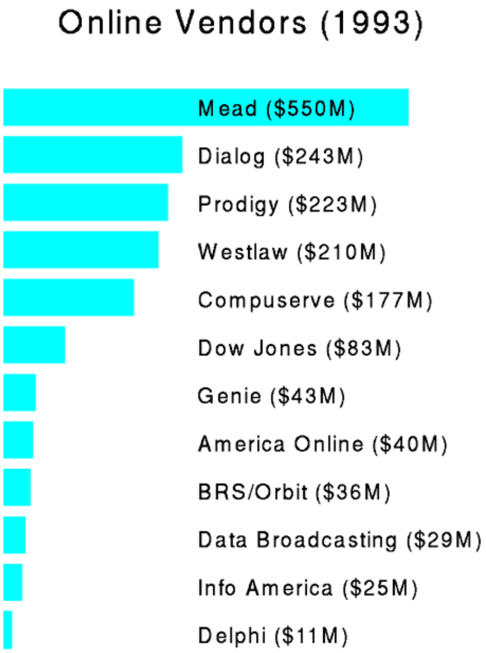


Figure 7

Nevertheless, in some ways the subject was still unfulfilled; it was still an area primarily of interest to specialists in libraries. A few efforts to sell online services to ordinary people (BRS After Dark, Dialog Knowledge Index) had not been big successes. And there was little use of advanced retrieval algorithms in the commercial systems. Word processing was everywhere, and yet most of the material keyed in was printed out and the machine-readable text not used any further.

This decade, however, has produced another technological revolution in the Internet. Even one year ago, in 1994, people were saying to me “only about 15% of home computers have a modem, who cares about online information?” Now we see forecasts that within ten years everyone will be on the net; one million people a month are signing up. What is remarkable is not that everyone is accessing information, but that everyone is providing information. For decades, information distribution had been a few large publishers sending information to many consumers. Now the consumers are generating their own information and classifying it themselves, as all sorts of people generate their own home page and link it to all kinds of resources they care about.

Comparing this with Bush's forecast, we see that his model of information storage is coming true. Each user is organizing information of personal interest, and trading this with others. What is also remarkable is that it is happening entirely on a free basis, without much support from industry. Admittedly much of the information is of low quality, but that does not seem to make it less attractive. In addition to the individual hyperlinks, we also have some classifications (e.g. Yahoo, `yet another hierarchical organization'). There are now over ten million pages on the net (a guess from the Lycos group). In the late 1980s it became normal to expect everyone to have a fax number; now everyone is expected to have a home page. Figure 8 shows the growth of the Web.

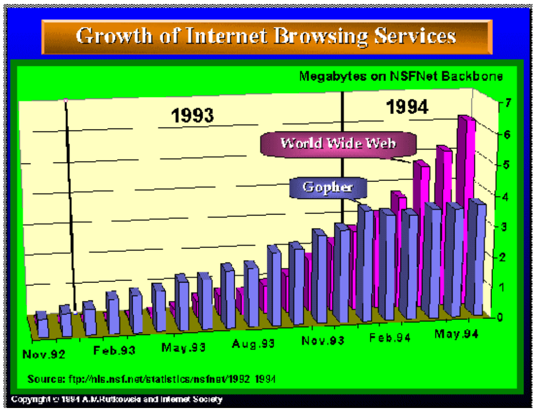


Figure 8

Although there are many free pages, there are also many groups planning to sell information on the net (and some already doing it). The Internet is now a standard medium for publishing, thanks to the development of Mosaic by Marc Andressen of NCSA (now at Netscape). The availability of pictures has provided an attractiveness, and thus a growth rate that neither “gopher” nor WAIS was able to achieve. Pictures do require a great deal more space, but as the curve of disk prices shown before in Figure 4 suggests, it is now affordable. This year, in fact, the world's hard disk industry will ship over 15 petabytes of disk; the equivalent of 2 MB for every person on earth (I can remember when I, as a researcher at one of the prime industrial labs, was hassled to keep my disk usage under 200 Kbytes).

Another technology that is lifting off in the 1990s is scanning. Until Mosaic, most online information was text-only. Publishers attempting to deal with the graphical content of their publications did not know what to do, until the advent of cheap disks and still cheaper CD-ROMs made it possible to simply scan all the pages into images and sell these images. Many publisher projects and library preservation projects rely on imaging, since scanning a page costs between a few cents and 25 cents per page, depending on paper and print quality, while keying it will cost $2-$3 per page.

Today we see publishers doing both kinds of information sales. Some put full text into an online system, and some sell page images via CD-ROM. Sometimes the same journal is available both ways. Even more interesting is the phenomenon of the electronic which is only published online. There are several hundred such today, most free but some charged for; OCLC, for example, is publishing scientific journals for AAAS and other publishers in online format. The economic advantages of this format may well mean that many smaller journals either move to this style of net distribution or disappear, as soon as the universities figure out how to deal with tenure cases based on online publication.

In the research world, the IR community was suddenly delighted to find the technologies with which they had experimented for decades actually in use. Bruce Croft's NSF Center for Intelligent Information Retrieval at the University of Massachusetts, for example, provides the software for the Thomas system at the Library of Congress (the online Congressional Record). Gerry Salton's software was used to make cross references for a CD-ROM encyclopedia. And relevance feedback is a key ingredient in the WAIS software of Brewster Kahle (now bought out by America Online). There is now an active industry in text retrieval software, as shown in Figure 9.

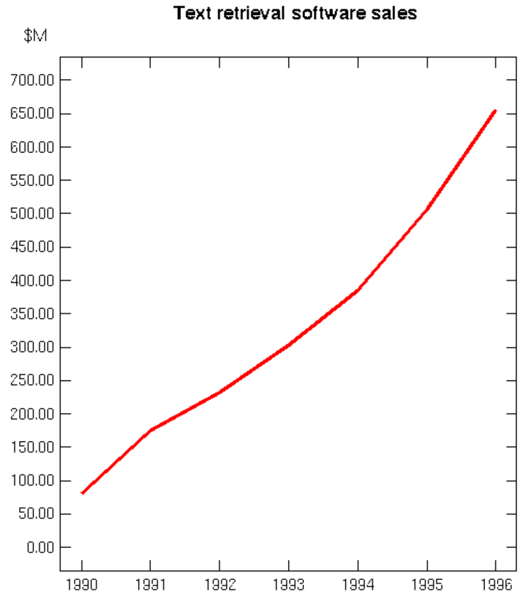
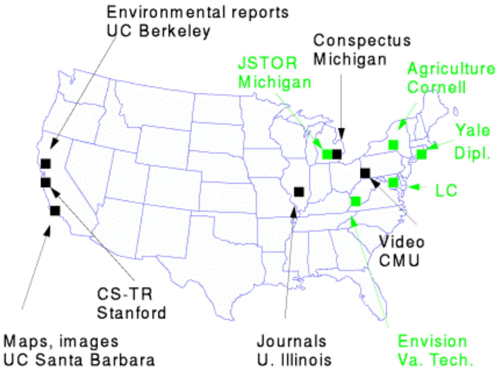


Figure 9

Politically, the opinion of information retrieval boomed as the Clinton administration started. Vice President Gore had been a long-time supporter of computer networking, and the new Administration looked at digital libraries as a way of coping with the needs of America for better education. Computer networks, the Vice President suggested, would bring information to every student; he described, for example, his “vision of a schoolchild in my home town of Carthage, Tennessee, being able to come home, turn on her computer and plug into the Library of Congress.” In order to produce this result, the Federal government started a Digital Library research initiative from NSF, NASA and ARPA. [Fox 1995]. Under it, six universities are leading consortia to study different aspects of digital library construction and use. Figure 10 shows the projects funded by the digital library initiative, and several other major digital library projects in the United States.

NSF, NASA, ARPA
Digital Library Projects



Other projects in green

And... IBM/Vatican library, Seville, ...
CS tech reports, 5 universities

Digital Library Foundation, Digital Preservation Consortium

Figure 10

On the side of retrieval evaluation, there was a sudden jump, after thirty years, to realistic collections for evaluation. The TREC conferences, run by Donna Harmon of NIST, distributed a gigabyte of text with hundreds of queries, as a way of judging the performance of retrieval systems. TREC results confirmed that there is still very large scatter in the performance of retrieval systems, not only by question but even over individual relevant documents within answer lists. A "best-hindsight" system (pick the right answers from different systems for each query) would outperform any of the actual systems by a wide margin. Thus, we still need research to try to understand how people phrase questions and how we can get questions composed so that we can provide good answers to them.

FULFILLMENT (2000s)

What should happen in the next decade, when IR will be 55 to 65? I have labeled this, optimistically, as `fulfillment;' Shakespeare had it as an age of a comic character. Which will it be? I believe that in this decade we will see not just Bush's goal of a 1M book library, so that most ordinary questions can be answered by reference to online materials rather than paper materials, but also the routine offering of new books online, and the routine retrospective conversion of library collections. We will also have enough guidance companies on the Web to satisfy anyone, so that the lack of any fundamental advances in knowledge organization will not matter.

Where we will need research is in the handling of images, sounds and video. Nearly the search algorithms we have today are based on text and words; we are all Whorfians with a vengeance, after the decline of controlled vocabularies and AI research into computational linguistics. So what will we do with the very large volumes of pictures and videos that need to be retrieved? Again, there is the Bush choice in which individuals will make indexes and pointers; and the Weaver choice, in which projects like the IBM QBIC system will be expanded and make it possible for us to do content retrieval on images. [Flickner 1995].

Looking at Bush's goal of a 1M volume library, we should note that Mead Data Central today has the equivalent amount of material, if not in the form of a general book collection. At 1 MB of ascii per book, the 2.4 TB at Mead already are the equivalent of more than 1 M books. But for a general collection we need the economic arrangements that will let publishers and authors provide each new book and journal in electronic form, perhaps in addition to the paper form for many years. For academic publishing, this may not be a problem (since the authors are not paid, and many university presses run at a loss, we may find that a transition to self-publishing is not very painful). But for commercial publishing, we need a way for all to be treated fairly: authors, publishers, illustrators, photographers, and readers. Many companies are working actively on this; see Figure 11. We desperately need some leadership in these economics areas; given the number of people working on it, we'll probably get some.

Some Internet Vendors

LEXIS/NEXIS
WHAT'S NEW
OPENMARKET
What's New!
VISIT THE OPEN MARKETPLACE
Doing Business on the Internet
What's New!
Re-Inventing Business for the Electronic Age
Doing Business on the Internet
FIRST VIRTUAL
HOLDINGS INCORPORATED
Electronic

First Virtual -- Bringing the World of Electronic Commerce to Your Computer -- TODAY!

FIRST VIRTUAL

First Virtual Holdings Incorporated presents the First Virtual Internet Payment System.

Figure 11

In addition to the provision of new material, we are seeing libraries converting their old material. The Cornell CLASS project set a cost figure of $30-$40 per book for scanning; [Kenney 1992]. in addition it costs perhaps $10 for the disk space to store the converted book. This is now comparable to the cost per book of new bookstacks built on campus (the new stack at Cornell is $20/book, and at Berkeley the stack under construction is $30/book). If several libraries could agree to scan an out-of-print book and move it to a warehouse, they would all save money over building stack space for it. Of course, they would save even more money just by moving the book to an off-campus warehouse. The online book, however, is likely to be perceived as much more usable. Experience with online catalogs, for example, is that when a university has about 1/3 of its books in the OPAC, and 2/3 still in an old card file, the students start ignoring the old card file, and some form of retrospective conversion becomes desirable. The Mellon Foundation is now supporting project JSTOR, which is scanning ten major economics and history journals back to the beginning as a replacement for shelf space.

So, we can see how, in the next few years, we'll have a large book library. What about pictures, sounds and video? All pose problems in indexing and searching, which may be overcome simply by people creating their own trails. Video poses, in addition, a storage problem. Going back thirty years, 2000 80-column punched cards held 160 Kbytes, weighed 10 lbs, and occupied 0.2 cubic foot, a storage density of 0.016 MB/lb or .8 MB/cu ft. Exabyte 8mm cartridges hold 5 GB, weigh 3 ounces (0.19 lb) and have a size of 5.5 cubic inches or 0.003 cubic ft; a storage density of 1100 GB/cu ft or 27 GB/lb. This means that there is an improvement in bytes/lb of a factor of 1.6 million and in bytes/cu ft of 1.3; comparing punched cards with CD-ROMs instead of Exabyte cartridges, the improvement is somewhat less, perhaps 300,000 to 500,000. But look at how much space video takes: a page which might contain 2000 bytes will take about 3 minutes to read aloud. A video of that reading, even compressed by MPEG-1 will be over 30 Mbytes, and for the quality imagined for decent video will be perhaps 100 Mbytes, or 50,000 times as much as the text. So we have given back much of the improvement in storage space. It will take another generation of storage space cost decreases before we can move from each researcher having a single, precious digitized video item to large, useful libraries of digitized video.

When we do have video libraries, we will need to build better systems for retrieving material from them. In research terms, we will need to move into more serious image recognition and sound recognition research, areas which have been more promising than computational linguistics (as one might expect, given that pigeons can be trained to do image recognition, but not language understanding). We also can look to some new techniques in cooperative browsing, in which the opinions of people about videos or sounds can be used to guide others. [Hill 1995].

RETIREMENT (2010)

Shakespeare gives the last age as senility, but we'll be more optimistic. At this point, 65 years after Bush's paper, we can imagine that the basic job of conversion to machine-readable form is done. It is not likely to be complete: there are still many manuscripts in European libraries that have never been printed, and there are likely to be many books that are never converted to computerized form. But these will not be the key items, but the obscure ones. We can imagine, for example, that central library buildings on campus have been reclaimed for other uses, as students access all the works they need from dormitory room computer. A few scholars, as with those who now consult manuscript and archive collections, go off to read the obscure books.

Most of the students are probably not just reading, anyway. By this time, multimedia information will be as easy to deal with as text. People will have been through the picture files, even the 9 million Library of Congress photographs, and noted which are relevant to which subjects, linking them in to pages of directory information. Most students, faced with a choice between reading a book and watching a TV program on a subject, will watch the TV program. Since there are likely to be fewer TV programs than books on any given subject, this limits their choice. The US has about 60,000 academic-type books published each year, while even if we allow ten primary TV channels (4 network and six cable) broadcasting 4 hours of prime time shows 365 days per year, that is only 20,000 or so TV shows.

Educators will probably bemoan this process. They will suggest that students are not getting enough diversity, not only because there is too much emphasis on multimedia shows, but because universities have been combining and getting more of their courses by remote learning. None the less, there will be no stopping it, any more than we have been able to keep the vaudeville theatres alive in small towns. The economics of university libraries show that there is a large potential savings if we can deliver material more directly from authors and publishers to readers; see Figure 12.

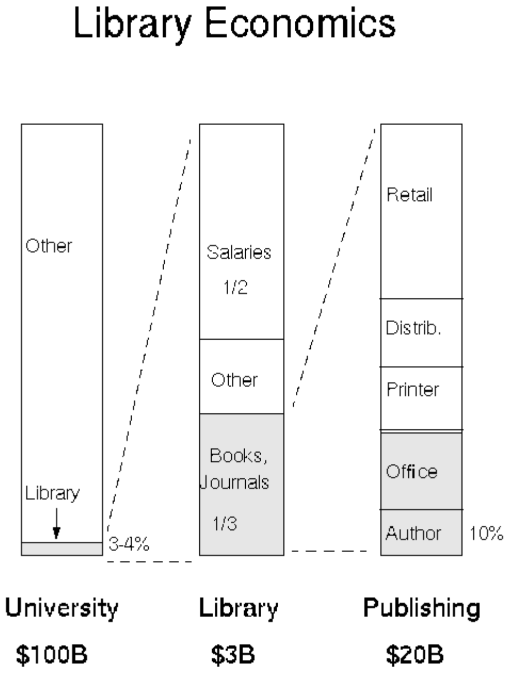


Figure 12

Internationalism will be a major issue by now, if not before. Although the Internet has been heavily US-based and English-language up through 1995, it will now include the entire world and there will be many multilingual trails. Methods for dealing with this, improving language teaching, and otherwise dealing with the fears of countries who think that their culture will be too peripheral in the Web, will be political issues in the next century.

As for the researchers, there will be engineering work in improving the systems, and there will be applications research as we learn new ways to use our new systems. Perhaps there will even be thriving schools giving PhDs in the arcane details of probabilistic retrieval. But many of the scientific researchers, I suspect, will have moved on to biotechnology. And only a few of us old fogies will remember that we were once promised fully automatic language analysis and question answering.

WHAT MIGHT GO WRONG?

This is all very optimistic, and suggests a long and happy life for our imaginary IR person. What are the alternatives? Is there an equivalent of an automobile accident that would leave IR stuck at its present state? We all know about quad sound and CB radio; what might happen to stop progress in online information availability? The most obvious problems relate to the need to publish current information with being overrun by illegal copying. This problem destroyed the computer games industry in the late 1970s, and is extremely serious for the software industry right now in many countries. However, there is a great deal of work to sort this out, and it is likely that acceptable solutions will be found.

The analogy of CB radio, however, is unfortunately relevant. Many people already complain that the Internet is so full of trash that it can no longer be used. Some of this is the same as suburbanites wishing that they had been the last people allowed to move into their town. But part of it is indeed a problem with abuse by commercial operators (such as the infamous Canter and Siegel `Green Card' postings), by the pornographers, and by people who simply do not think before they clutter a bulletin board going to many tens of thousands of people. Perhaps the easiest way to deal with some of these problems would be an ethical position that anonymous posting should not be allowed. This might remove a great deal of the junk. A more likely outcome is that moderated bulletin boards become more popular and unmoderated, unedited newsgroups lose their readers.

Another, very serious problem, is that of the copyright law. At the moment there are no procedures that involve low administrative overhead for clearing copyright. Proposed revisions to the copyright statutes would extend its scope, aggravating the issues. It is said, for example, that in preparing a commemorative CD-ROM for the 500th anniversary of the first Columbus voyage to America, IBM spent over $1M clearing rights, of which only about $10K went to the rights holders; everything else went into administrative and legal fees. It is possible that we could find ourselves in a strange world in which everything being published now was available, since we imagine that the economics of current publications is sorted out; and that material printed before 1920 was available, since it was out of copyright; but that the material printed between 1920 and 1995 is hard to get. We have to hope for either administrative or legal progress in these areas.

We might also find that the Net worked, and that people could get commercially published material, but that the freely contributed work of today was hard to obtain. For example, suppose we wind up adopting technology, such as that proposed by some cable companies, which provides a great deal of bandwidth to residences, but relatively little in the reverse direction. This might make it difficult for ordinary citizens to put forward their views. The Electronic Frontier Foundation has been arguing for some years against this danger and urging symmetric network designs for the NII.

Finally, there are some unfortunate analogies with activities like nuclear power, childhood vaccines, and birth control technologies. In these areas, various kinds of legal liability and public policy debates have hamstrung technological development and availability. There are undoubtedly readers of this article who think that for some of these topics, the naysayers are right and that technology should be stopped. We need to be aware that there are groups which would like to limit the Internet for a variety of reasons, whether the fear of net pornography, uncontrolled distribution of cryptographic technology, or the gains to be made from libel or strict liability lawsuits. We need to try to stave off the worst of these attacks on the Web and try to defend freedom of expression and openness as democratic values. And we must remember that there is no technological certainty, and that perfectly good technologies can be stopped by legal and political forces.

WHAT MIGHT GO RIGHT?

Not to end on an unhappy note, remember that it looks like this is all going to work. Already in 1995, university audiences asked where they got the answer to the last question they looked up usually say on a screen, not on paper. Given the number and rates of progress of conversion projects, it certainly looks like Bush's dream, as far as the availability of information on screens, will be achieved in one lifetime.

What about his view of the new profession of trailblazer? There were of course librarians in his day, although they had low status even then. Remember that many people, with no idea of the job classifications that exist in libraries, still think that a librarian is somebody who puts books on shelves in order. Will, in a future world of online information, the job of organizing information have higher status, whatever it is called? I am optimistic about this, by analogy with accountancy. Once upon a time accountants were thought of as people who were good at arithmetic. Nowadays calculators and computers have made arithmetical skill irrelevant; does this mean that accountants are unimportant? As we all know, the answer is the reverse and financial types are more likely to run corporations than before. So if computers make alphabetizing an irrelevant skill, this may well make librarians or their successors more important than before. If we think of information as a sea, the job of the librarian in the future will no longer be to provide the water, but to navigate the ship.

REFERENCES

[Bush 1945]. Vannevar Bush; "As We May Think," Atlantic Monthly 176 (1) pp. 101-108 (1945). On the Internet in http://ebbs.english.vt.edu/hthl/As_We_May_Think.html and several other sites.

[Cleverdon 1966]. C. W. Cleverdon, J. Mills, and E. M. Keen Factors Determining the Performance of Indexing Systems ASLIB Cranfield Research Project.

[Cleverdon 1970]. Cyril Cleverdon; "Progress in documentation, evaluation tests of information retrieval systems," J. Documentation 26 (1) pp. 55-67 (March 1970).

[Flickner 1995]. M. Flickner, H. Sawhneyi, W. Niblack, J. Ashley, Qian Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker; "Query by image and video content: the QBIC system," IEEE Computer 28 (9) pp. 23-32 (Sept. 1995).

[Fox 1995]. Ed Fox; "Special issue on digital libraries.," Communications of the ACM, New York (April 1995).

[Hill 1995]. Will Hill, Larry Stead, Mark Rosenstein, and George Furnas; "Recommending and Evaluating Choices in a Virtual Community of Use," CHI 95 pp. 194-201, Denver, Colorado (1995).

[Kenney 1992]. Anne Kenney, and Lynne Personius Joint Study in Digital Preservation Commission on Preservation and Access (1992). ISBN 1-887334-17-3. .

[Peat 1985]. F. David Peat Artificial Intelligence - How Machines Think Baen Enterprises (1985).

[Salton 1968]. G. Salton Automatic Information Organization and Retrieval McGraw-Hill (1968).

[Samuel 1964]. A. L. Samuel; "The banishment of paperwork," New Scientist 21 (380) pp. 529-530 ((27 February 1964).

[Schank 1973]. R. C. Schank, and K. Colby Computer Models of Thought and Language W. H. Freeman (1973).

[Schuchman 1981]. Hedvah L. Schuchman Information transfer in engineering. Report 461-46-27 The Futures Group (1981). ISBN 0-9605196-0-2.

[Shakespeare 1599]. William Shakespeare, As You Like It, Act 2, Scene 7, lines 143-166.

[Weaver 1955]. Warren Weaver; "Translation," pages 15-27 in Machine Translation of Languages, eds. W. N. Locke and A. D. Booth, John Wiley, New York (1955). Reprint of 1949 memo.

NOTES