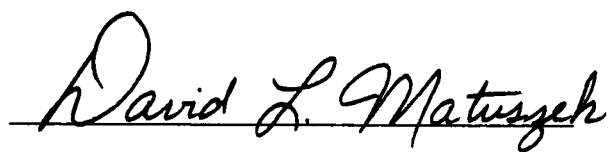


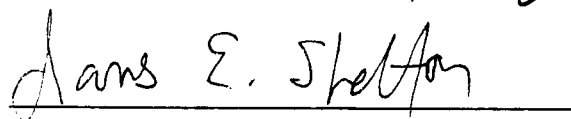
To the Graduate Council:

I am submitting herewith a thesis written by Hal Manley Harrison entitled "Advances in Information Retrieval through the Use of Full Text Databases." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.


Kevin C. O'Kane, Major Professor

We have read this thesis
and recommend its acceptance:







Accepted for the Council:


The Graduate School

ADVANCES IN INFORMATION RETRIEVAL THROUGH THE
USE OF FULL TEXT DATABASES

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

.

Hal Manley Harrison

August 1983

The data:

"Text was designed for people to read, and not
for indexing or for information retrieval."
--Lauren Doyle [1975, pg 324]

A user:

"I spend half my time looking for a picture of
a horse like the one I have. I find lots of
horses. Many of them look like mine, but I keep
on looking, hoping I will find one that has
written under it, 'This is a horse like
yours.'" --Justice of the North Carolina
Supreme Court, Sam Ervin [May, 1973, pg 81]

ABSTRACT

Information is exploding and can't be kept up with. Most will not be referenced, but, with declining hardware costs, storing all of it is cheaper than trying to predict what won't be needed. The most efficient use of human facilities is at the search and output stage. Let the user do the editing, based on a full text database.

The first and second of the four sections of this thesis form a survey-conclusions pair. The first section surveys several factors which have contributed to current possibilities in electronic publishing. But current access methods are inadequate. The second section draws some original distinctions which separate full text retrieval from other types of IR, namely bibliographic and question-answering systems.

The third and fourth sections form a similar pair. The third surveys the existing methods of processing text, while the fourth offers some techniques for full text retrieval which are a step beyond current practice. By processing relevancy feedback elicited from users, a future full text system will automatically expand and update its search vocabulary as it gets more use. The emphasis should be on making the collection more visible to the user.

TABLE OF CONTENTS

SECTION	PAGE
I. THE ADVENT OF FULL TEXT RETRIEVAL SYSTEMS	1
A. FT Systems Operating Today.	2
B. Definition	21
C. History of Search Strategies	26
D. The Rising Potential for FT	35
1. Photocomposition	37
2. The cost of mass memory	39
3. The wired community	41
4. The spread of personal computers	44
5. Other issues	45
E. Access Methods	51
1. Current methods on DIALOG, STAIRS, and LEXIS	52
2. Ranking	59
3. A simpler system and the public	61
4. Menus and the public	67
II. FT IN PERSPECTIVE	69
A. The Narrow Perspective	69
1. Distinguishing bibliographic services from FT	69
2. Distinguishing FT from question-answering systems	76
B. The Broad Perspective	84
1. A model of information cycles	85
2. Will changing information routes change society?	94
III. WAYS OF ANALYZING TEXT	102
A. Manually	108
B. Mechanically	126
C. Statistically	133
D. Linguistically	152
IV. PROPOSED LINGUISTIC ADDITIONS TO A STATISTICAL SYSTEM	158
A. Deeper Indexing through Phrase Formation	159
B. Automatic Thesaurus Maintenance	165
C. Learning from Queries	170
D. Automatic Relevancy Feedback	171
E. Conclusion	177
LIST OF REFERENCES	180
APPENDIX	190
INDEX	191
VITA	196

ABBREVIATIONS

FT - full test (retrieval (system(s))), whichever fits the context. Implies IR techniques possible when the full text of documents, rather than document descriptions only, is available in machine readable form.

IR - Information (Organization and) Retrieval, usually by subject. The process of linking a user's query to hopefully relevant documents. Distinguish from Data Processing which delivers information from formatted fields into other formatted fields.

online - access from a user's remote terminal to a database maintained and managed by others.

SDI - Selected Dissemination of Information. Rather than broadcast descriptions of all their new acquisitions, a library sends customized notifications of items likely to be of interest to individual users.

Q-A - Question-answering, Expert, or fact delivery systems. They use techniques of Artificial Intelligence.

videotex(t), viewdata, teletext - These terms are ill defined. Authors usually distinguish one-way systems (CEEFAX), where the viewer must select from whatever comes, and two-way systems (Prestel), where he orders interactively what he will see. But the technology of the two systems is so closely related that the distinction blurs; cf. FT.

[. . .] - used only for citations, either in the form
[Salton, 1983, pg 40] or Salton [1983, pg 40] speaks
. . . .

{. . .} - used only within a quote to separate insertions
by the present author from statements of the quoted
author.

1. THE ADVENT OF FULL TEXT RETRIEVAL SYSTEMS

The attack of current Information Retrieval (IR) methods upon the proverbial information explosion is like attempting to harness a waterfall with a box of Tinker Toys. To control today's flood of information there are only a few standard pieces to work with. Occasionally a researcher will fashion and add a new piece to the limited collection of components.

Some old pieces are: controlled vocabularies, Boolean searches, how to get digitized text, how to store digitized text, classification schemes, derivative and assignment indexing, thesauruses, and inverted lists. New pieces added to the box in the last 20 years are: interactive searching, CRT displays, and statistical analysis of word frequencies--computational linguistics. The golded key piece, which appears to be hidden at the end of the rainbow is linguistic analysis. Its absence is noted in many projects, and players are constantly reassembling the pieces at hand in hopes of proving that syntactic and semantic processing won't be needed after all.

Section I introduces the many components needed for the ensuing discussion and seeks to isolate from the array of IR systems the trend of full text

retrieval. FT facilities usually stem from other operations.

A. FT Systems Operating Today.

Descriptions of the following prominent full text systems:

Info Globe
Prestel
CompuServe
LEXIS
U.S. Congress
Newsletters, and
French Telephone Directories

should distinguish the role of vendor from that of database producer.

A database producer gathers information manually, converts it to machine readable form, and delivers a copy of his tape to a vendor. The vendor provides public access via his computers to the disk mounted database. In the world of print, book publishers and newspapers are database producers; bookstores and news stands are vendors.

Info Globe

The Toronto Globe and Mail, Canada's largest morning daily, advertises the completeness and timeliness of online access to their system as follows, "It gives you the full text of any article, any time you want--at the push of a button--and you don't need an index to slow you down!"

Besides the full text, several other fields are stored for each story from 1977 to present.

- Headlines;
- Date and column location in the paper;
- Byline, if published, or author's name (for staff access only) if not;
- Dateline or wire service source;
- Category--news, sports, review, editorial, letters to the editor, etc.;
- Illustrations flag for stories carrying pictures in the printed version (which see to view them);
- Memo of any errors or unusual features in the story;
- Notes of any misspelled words caught by morgue librarians; and
- Added Search Terms.

The whole database is searchable on most fields, including any significant word appearing in text or headlines. The Added-Search-Terms field holds manually assigned key words not appearing in the story. The Notes-of-any-misspelled-words field effectively deals with the problem, "Does the vendor change the document? or let a misspelling pass which may render the document irretrievable when Boolean matches are sought."

The access program from QL Systems Ltd. of Kingston, Ontario, the original vendor, supports the Boolean operators and truncation. For instance, a request for any article on Canadian unity between June 10 and October 7, 1978, would read

canad* & unity @DATE AFTER 780610 BEFORE 781107

Whole articles, or only selected fields may be displayed [Nash, 1979].

Although the command language is quite refined, it has difficulties. For instance, "781107" has to be awkward. Or, in reply to the above query, QL's software returns the number of

canad* AND unity
articles in the whole database while only sequentially displaying those in the specified date range. This leaves the user guessing the size of his document set until he has finished looking at it.

The full text legal database system Westlaw suffered the same brand of malady until its revision in the spring of 1983. The document set sizes returned always ignored refinements such as proximity relationships between terms or a time range. Before continuing query reformulation, the user had to give a special command and wait minutes to get the true size of his document set. The Westlaw system was also originally written by QL Systems [Voedisch, 1983].

To use Info Globe, the online information division of the Globe and Mail, a subscriber must have his own terminal and phone connection. Telecommunication charges, if by Datapacs in Canada or if via Tymnet or Telenet south of the border, are included in the \$105 per hour connect time charge. Hard copy printed offline will be mailed the next morning for \$0.003 per line, and an average client spends

\$2 per day, which doesn't allow much time for online reading of the full text.

Globe experiments on electrifying their "morgue" began in 1974. Keeping one copy of the database for staff use, the newspaper sent a second copy to the vendor all through the first year of service, 1978. Then Info Globe diverted public access to its own IBM 360/50 until 1982, when they again engaged a vendor, Canada Systems Group, Mississauga, Ontario. The database grows at the rate of one million characters per day, and subscriptions grew from 100 in 1979 to 300 in 1981 [Online Review 1982a].

In contrast, the New York Times, which has onlined an extensive abstracting service (of its paper and 55 other publications) with its own computers since 1970, only announced full text access to the paper in July of 1981, retrospective to June, 1980. Their database grows at 90,000 stories, or 105 million words, per year [Hogan, 1979; Inforum, 1981].

Through the years, the Time's Information Bank had developed over a thousand user sites around the world. But because of high hourly charges for both papers, making the full text available is really only an extension of a bibliographic, controlled vocabulary search. Instead of always using terms an indexer has previously assigned to an

article, the user may query on any terms he thinks might be in the full text itself, a free language search.

Prestel

The British Post Office was first with a more consumer oriented service. Being the agency of a government with a more efficient bureaucracy than some countries, the BPO was able to budget several million dollars to buy computers, bypass extended marketing tests, and offer Prestel subscriptions to the public in 1979, just 5 years after Sam Fedida, one of their video telephone researchers, demonstrated a model of the home oriented system. In the 1960's Fedida worked with an international hotel booking information service. He learned that 80% of the overhead went for clerks to relay information from the central computers to users on the telephone (just as airlines handle information and reservation services today). Fedida designed a system where a user could query the central computer himself through a home TV set modified for a small number pad.

After an automatic telephone connection is established, the first display looks something like:

Prestel	Page 0a
1. List of Information on Prestel	
News & Weather	Sports & Hobbies
Entertainment	Holidays & Travel
Marketplace	Jobs & Careers
Advice	Books & Reference

- House & Garden
- 2. Alphabetical List of Contents
- 3. Local Information
- 4. International Information
- 5. Business Prestel
- 6. Special Databases
- 7. Guide for Prestel Users
- 8. Guide for Information Providers

Entering a digit from the number pad produces another menu, and another, until a desired page containing about 100 words of text is reached. Included in one corner is the price of accessing the page, ranging from free for ads to \$0.50 or \$1 for business information. In addition to this charge, the user incurs a phone line and computer bill with the BPO.

While encouraging TV set manufacturers to produce modified sets, the government utility also leased access to the information providers. About 1100 users were set up with code numbers in 1979 and 150 organizations, from Pan Am Airline, the Barclay Bank, Financial Times newspaper, and MacDonald Educational publishers, to the Government Information Service were working to input 200,000 pages into the Post Office computers. Electronic mail is also offered:

Page 81a

BIRTHDAYS & ANNIVERSARIES

- 0. HAPPY BIRTHDAY
- 1. HAPPY ANNIVERSARY DEAR
- 2. YOUR WEDDING ANNIVERSARY
- 3. HAPPY ANNIVERSARY
- 4. HAPPY BIRTHDAY DAD
- 5. HAPPY BIRTHDAY MUM
- 6. HAPPY BIRTHDAY SON
- 7. HAPPY BIRTHDAY DAUGHTER

- 8. YOUR NEW BABY
 - 9. MERRY XMAS
- Key number required

[Sigel, 1980, pp 59-78].

The frames getting the most frequent, though brief, use are the menus, guiding users to what they really want. Any content page may be retrieved by entering its number, if the user knows it. Printed directories have been suggested to economize on the sometimes tedious menu searching [Storey, 1982]. Actual growth has failed expectations, with only 17,000 terminals in use, most in businesses rather than homes, by 1983.

Cost may be the key to growth in the consumer information market, because the BBC's CEEFAX service has boomed. One million sets, capable of displaying text transmitted in the vertical blanking interval of the TV signal, are expected to be in use by the end of 1983. [Anthony, 1983]. Perhaps the important difference with Prestel is that once the TV set is adapted for CEEFAX, there is no additional charge for receiving any of the 100 frames available. Size is an obvious limitation. The pages are filled with newspaper type material prepared by a department of the BBC started in 1975. Along with news, finance, sports, weather and entertainment listing pages, there are three set aside for a full alphabetical index and another which shows the current TV program subject to being captioned with news flashes. Britain's Independent

Television Authority also produces a compatible product, ORACLE, on its channels.

Consecutive pages are transmitted continuously at a rate of four per second. Again with a number pad, the viewer selects a page by number. The TV set waits 25 seconds or less for the selected page to come around again and stores it locally. So the average waiting time is twelve and a half seconds. An attraction of the system is that it uses an otherwise untapped channel, the 25 TV lines never shown for electronic reasons on the home screen. That one way channel is free 24 hours a day.

Prestel, on the other hand, makes additional demand on busy telephone lines. Originally the BPO aimed Prestel at the home information market to increase telephone usage at the slack times after office hours. Unfortunately, the greater variety of information available on Prestel attracts business subscribers, further loading the phone lines at their peak time. An additional bias in the British experience of broadcast versus phone channels is that 97% of Britain's homes have TV's while only 60% have telephones.

Japan's CEEFAX type system is complicated by the necessity of transmitting ideographs instead of characters, while the French ANTIOPE was designed to also be compatible with their Prestel like Teledon system.

In America, TV vertical interval systems have been tested in Salt Lake City and St. Louis, but the most widespread use is in the "closed captions" over PBS and ABC which transmit captions specially prepared for the deaf [Sigel, 1980].

CompuServe

Without the resources of the BPO, a computer time sharing company in Columbus, Ohio, may be accomplishing in the US via Ma Bell what Prestel intends in the UK. The American way is to market the telecommunications software for established home computers, to offer connect time cheaply at \$5 an hour, and to build a general appeal database.

CompuServe began in 1970 offering data processing services to the many insurance companies in Columbus. To keep its mainframe busy after office hours, CompuServe organized time sharing phone networks which now extend across the country. By 1981, 9,000 hobbyists and businesses used the net work to run programs, send electronic mail, and access information [Fortune, 1980; U.S. News & World Report, 1981]. After a quick telephone call to a local number in nearly any American city, the screen looks something like this:

CompuServe: CompuServe Information Service Page CIS-1

1. Home Services
2. Business and Finance
3. Personal Computing
4. Services for Professionals
5. User Information
6. Index

Enter your selection number
or H for more information.

User: 1

CS: CompuServe Information Service Page HOM-1

Home Service

1. News/Weather/Sports
2. Reference Library
3. Communications
4. Home Shopping/Banking
5. Groups and Clubs
6. Games and Entertainment
7. Education
8. Home Management
9. Travel

Last menu page. Key digit
or M for previous menu.

User: 1

CS: CompuServe Information Service Page Hom-10

News/Weather/Sports

News Services

1. The Washington Post
2. St. Louis Post Dispatch
3. Middlesex Daily News
4. AP Viewdata Wire
5. NOAA Weather Wire
6. Official PGA Tour Guide
7. Hollywood Hotline

User: 1

CS: Request Received
One minute please

Thank You for waiting

Washington Post Page TWP-1

1. The Administration
 2. Congress
 3. Business and Economics
 4. Science and Technology
 5. The Courts and the Law
 6. Politics
 7. World and Nation
 8. Editorial and Commentary
 9. News Summary and Calendar
 10. Associated Press News
- Copyright (c) 1983
The Washington Post Co.

User: 8

CS: Washington Post Page TWP-17

1. Editorial News Digest
2. Editorials
3. If They Lock Up Lech Walesa
4. Day of Reckoning
5. . . . And the Day After
6. The Adelman Vote
7. Raspberry on Fear and Taxes
8. Mann on Taxation
9. Rosenfeld on Central America

User: 3

CS: Washington Post Page TWP-125

The Polish government could
lock up Lech Walesa again, of
course, or arrange to have him
run down by a truck, or
whatever. But then what? Gen.
Jaruzelski must calculate

The first time through such a menu tree at 30 characters per second is interesting, but backtracking gets boring quickly. Once the newness of choosing among the Washington Post, the St. Louis Post-Dispatch, etc. has worn off, the reader begins to feel that he might be getting more news information with less searching from the hard copy on his doorstep. The articles in the printed version are rewritten

for the capsulated CRT screen [U.S. News & World Report 1981] and only today's paper is online. Under

7. Education

will be found pages from Popular Mechanics Magazine, and under

2. Reference Library

an electronic version of the World Book Encyclopedia. Somehow it's not the same without the pictures, but it's there for anyone with a home computer and an acoustic coupler.

Radio Shack offers a "Videotex Package" for \$30. Besides the network software in a ROM cartridge and command language instructions, the kit contains account numbers for both CompuServe and the Dow Jones' News/Retrieval service and an initial free hour of connect time to both. In late March, 1983, a salesman at their Cedar Bluff store in Knoxville, Dick Vogler, said, "All last year we didn't have much interest in this network. But in the past 90 days we've sold dozens of these Packages. The idea of home information service may be about to take off." However, salesmen in other stores tie interest in this kit to locally available home banking, which has waned considerably since early 1982.

Large wall charts are sometimes available from CompuServe showing hundreds of page numbers. The user may

avoid the menus with a GOTO. Notably absent from the whole service is any other means of searching. One computer hobbyist familiar with CompuServe expressed an, "I wish there were some way to" and proceeded to describe a hunch he had for a search strategy. It could have been done with uniterms and Boolean operators which he had not heard of before [Chamberlain, 1983]. Of course adding such capability is only a matter of a software in Columbus when users are ready for it.

While CompuServe is an example of a computer company moving into FT vending, many database producing newspapers, are looking at the same business. Several are feeding cable TV systems to gain experience in the electronic medium, and both the New York Times and the Miami Herald are information providers for Britain's Prestel.

LEXIS

Legal computer aided research has been full text from the beginning, datable from the American Bar Association's 1960 convention in Washington D.C. The University of Pittsburgh successfully demonstrated Boolean searches using punched cards and a full text of Health Law statutes. After donating transportation of the University's 650 computer to its headquarters in Washington, IBM discovered that due to

rough handling in shipment all the vacuum tubes in the machine needed replacement--5,000 of them [Lawlor, 1973].

The Ohio State Bar Association continued development of computer searching with the OBAR system containing statutes and case law from Ohio and New York States plus some Federal Tax material [McGonigal, 1973]. In 1969, Jerome Rubin, as president of what is now Mead Data Central, Inc., Dayton, Ohio, offered the LEXIS system to lawyers in his state and New York. The more prominent ones, including the Big Eight accounting firms, accepted the offer at \$85 per hour of connect time. [Business Week, 1974]. From the two state law libraries the data base grew to approximately 20 states by 1979, and then to all 50 states in 1981. It's thousands of millions of words also include extensive Federal material in such areas as Tax, Securities, Bankruptcy, Patents, Communications, Labor, Trade Regulations and British and French law [LEXIS, 1982]. User sites have grown from the two states to hundreds across America and in Europe. Over 20,000 searches are performed each day on Mead Data Central's Amdahl computers [News, 1980, Sept, pg 313].

Described as "flexible and forgiving" by Mr. Rubin [Hogan, 1979], the LEXIS service depends upon a leased terminal providing such special function keys as:

NEW SEARCH--new query coming,
TRANSMIT--for Send,

KWIC--specifying display of relevant passages with
keywords highlighted,
MAIL IT--for online ordering of offline hard copy,
CHANGE LIB, CHANGE FILE--for changing segments of
the database to be searched, and
SIGN OFF--for logging off [LEXIS, 1980].

According to Mead Data Central this is a more user friendly approach than expecting infrequent users to remember sequences of commands.

With LEXIS a commercial success, Rubin launched a new service in 1979, NEXIS. Beginning with retrospective full texts of the Washington Post, some news wires, and Newsweek magazine, the NEXIS database has now expanded to dozens of other current affairs publications such as U.S. News & World Report, Christian Science Monitor, and the English wire services of Japan and China. In 1982, Mead Data Central added the New York Times databases to NEXIS, including the full text, the Information Bank, geopolitical material organized by countries, and abstracts of advertising and marketing periodicals compiled in cooperation with the J. Walter Thompson ad agency.

Mead Data Central is an example of a vendor who has also worked to create its data base. Before 1969 and MDC's efforts with the Ohio Bar Association, the LEXIS database did not exist in electronic form. In contrast, the Westlaw system emanates from a century old database producer, the West Publishing Co., largest publisher of law books in the world. The full text of all cases published plus the West

editors' "headnotes" (abstracts of each case) and "key number" searching (over 400 classification headings) are available online [Westlaw, 1983]. The Westlaw system, running since 1975, resembles LEXIS in using special function keys. The University of Tennessee's Law Library recently changed from LEXIS to Westlaw because of the more favorable educational rates on the latter.

Congress

The US Congress probably comprises the world's largest concentration of policy makers using online IR systems. Members first needed word processors to control correspondence. But beginning with a test of seven terminals in the offices of Senators on the Committee on Rule and Administration in 1973, use of IR has mushroomed.

Initially the Library of Congress, as a database producer, offered on the seven terminals access to the printing tapes of the Digest of Bills and Resolutions, a summary of past and pending legislation. The terminals spread to the offices of most Senators and the access expanded to the New York Times' Information Bank and various government produced statistical databases.

Meanwhile the House of Representatives restricted its IR service to the Bill Status Office, to which a Representative could telephone for current information on

the status of any pending bill or resolution. Clerks accessed the elaborate, bill tracking database with IBM's STAIRS, and the system produced daily SDI to Representatives and Senators. In 1975 policy changed to allow online access to the database. Beyond those services already mentioned, the system is providing online:

- abstracts of the Congressional Record,
- committee reports,
- background briefs on current subjects,
- the Library of Congress' computerized card catalogue,
- an elaborate tracking of expenditures of the Federal budget, and
- a written summary of floor activities with a delay of only 10 minutes

(cable TV provides live coverage of the House floor).

In theory most of this information is public property, and outsiders may gain access to some of it by telephoning their Congressmen, using public terminals in the Library of Congress reading rooms, or online through various vendors [Gregory, 1979; Hamilton, 1973].

The spread of computers to State legislations is less conspicuous. SYSTEM 50 was an early batch retrieval system written by Aspen Systems Corporation containing statutes from the 50 States. Users submitted keywords and received printouts of statutes containing the words, or of the keywords imbedded in 70 characters of adjacent text, or of citations only [Doyle, 1975, pg 297]. Typically the laws of one State will run to 40 million characters. Over half of the States have some sort of online statute retrieval

for their legislators. Most have also extended this service to include online modification of the database when legislation changes existing statutes. Bill Status Reporting is also popular in State assemblies [Encyclopedia of Library and Information Science, 1975].

Newsletters

Perhaps one of the first places FT comes into its own is with newsletters. They are numerous, dated, important to a few, diverse, and often thrown away.

Newsnet, in Bryn Mauser, Pennsylvania, has twenty newsletters online and is aiming at over a hundred. Titles include

the Official FCC Daily Digest,
Fiber/Laser News,
Satellite Week, and
Viewdata/Videotex Report.

The publisher delivers his newsletter in machine readable form, and Newsnet indexes on significant words in titles and text. Subscriptions to the database are \$15 per month and the hourly rate for access varies with bandwidth--\$24 for a 300 baud line, \$48 for 1200 baud. Electronic mail to any of the publishers is also provided [News, 1982, Aug, pg 350].

BRS, one of the big three vendors, has another variant on the newsletter theme, which approaches the operation of Prestel. For \$200 per month, a newsletter producer

subscribes to the Electronic Newsletter Service, and maintains his text on BRS equipment as he wishes. Readers connect to BRS at \$10 per hour and read any newsletter not restricted to specific viewers [News, 1980, Sept, pg 313].

French telephone directories

And all the possibilities aren't exhausted yet. One of those, "Is it really true?" ideas is: give telephone subscribers simple terminals instead of telephone books and teach them to access the one, large, constantly updated central listing which the 411 operators use. The money saved by not having to print all those telephone books every year would eventually more than pay for all those terminals. AT&T tried this with 100 test homes in Albany, NY, but the larger experiment is in France.

There the government phone company purchased 300,000 home terminals and began giving them away. After a successful trial of 1500 homes in Rennes, 10,000 terminals a month were being given to subscribers in 1982 [Time, 1982]. But problems have developed so that local authority and press support in a given area must first be secured before continuing the give-away program. Nevertheless, any individual may obtain his terminal for a rental of about \$120 per year [Anthony, 1983].

In summary, FT in 1983 acts like an adolescent industry with experimental pricing policies and unsettled supply lines. Prices range from \$90 per hour on NEXIS and \$5 per hour on CompuServe to \$120 (or zero sometimes) per year for French telephone books.

Can video displays be created that will allow standard computer typeset materials to be used (Info Globe, Westlaw), or must text be rewritten for the brevity necessitated by 100-word, low resolution pages (Prestel, CEEFAX, the Washington Post on CompuServe)? And, can text be presented electronically in new colors or display technologies so that readers will not tire, or will hard copy always be in demand?

Such questions await further study of user habits and more market place experience, but the part of the confusion which comes from failure to distinguish FT from other services may be attacked through some definitions which follow.

B. Definition

Before continuing the discussion of FT, it will be helpful to give a preliminary definition. The elements of FT introduced here will be discussed more fully in later sections. The term, "storage and retrieval of full texts" is almost self explanatory. The full text of a document, as

written by the original author, is fed into an IR system; and, in response to queries, relevant full text is returned. In contrast, bibliographic systems provide the same service except processing only titles, abstracts, or some list of indexing terms meant to describe the referenced documents.

Though this definition of FT is simply stated, the ramifications are far reaching because of the kind of beasts computers are: they don't process natural language the way humans naturally do. Documents are designed for reading, not for indexing or as data for IR systems. Two implications will be discussed, the necessity of document processing prior to query processing and cost.

The larger-context problem

First, an author and his reader unconsciously share many concepts not explicitly stated in the document passed between them. A

. . . seemingly unlimited pool of knowledge,
standing ready to embrace every . . . sentence
heard by the human ear. [Doyle, 1975, pg 288].

Any system referencing documents by their meanings, as IR systems intend to do, must take this larger, implicit context into account. Human classifiers and abstractors operate within this seemingly unlimited context, describing documents in terms more readily understandable by those who

have not yet read them. Current bibliographic systems work well on documents that have been manually described in easily computable terms.

The storage of full text implies that searching will be done on it rather than only on descriptions as in the bibliographic system. A system may support this view to an FT user, but internally something must be done for the larger context-document description problem described above. In addition, processing time will be a problem.

Though Order(n) string searching algorithms exist, the huge size of an FT database makes even $O(n)$ processing across the whole database take too long. Therefore, when documents enter the system they must be processed to prepare structures, indexes, which are more quickly searched than the full text.

Even if new types of computers were available, creating searchable structures would still be useful. In Content Addressable Memory (CAM) architecture every memory location contains the few logical gates necessary to test the setting of its memory cell. Memory is referenced by broadcasting a command to all cells simultaneously, "The memory cell containing pattern ____ set your tag bit." Neighboring contents may then be retrieved by reference to which tag bits are set. CAM contains about 1.5 times the

number of logic gates as normal RAM and could be produced for 1.5 times the cost [Foster, 1976].

Or, a fixed-head disk could provide a similar capability. Every track on the disk has its own head, and every head has enough logic to process the broadcast command, "If your track contains ____ pattern, set your flag." The entire disk pack may be searched in one revolution [Boral, 1982]. Though such parallel processing promises acceptable search times across raw text, some sort of processing on each document would still have to be done to relate it to the larger context. Therefore the fact that VonNeumann computers require the intervention of indexes to gain access speed is not the disadvantage it might seem to be.

Cost

The other implication of the basic FT definition concerns cost. Practical FT depends upon cheap storage for billions of characters. Text compressors may act as back-end storage processors, but they will only reduce memory requirements by a constant factor, not by an order of magnitude. Many writers, e.g. [Sparck Jones and Kay, 1973, pp 4, 28], restrain all their discussion with concern for the limitations of computer capacity and budgets, but in fact the cost of hardware may decline to the point that

it is cheaper to save everything in a constantly growing system than it is to pay humans to decide what to keep and what to throw away.

If storage is cheap, then processing is also likely to be cheap; and an FT system need not shrink from the complex processing invited by its natural language data. However, an empirical argument can be made against computational complexity in the system.

Salton, Cleverdon, and other experimenters have demonstrated that adding complexity to a simple computerized IR system does not improve its performance, as surprising as this may seem [Sparck Jones and Kay, 1973, pg 126]. But this conclusion is challenged on the basis that the experimental collections were too small, too specialized, and contained a too precise vocabulary (scientific) to reveal the advantages complexity would have brought [Austin, 1976b, pg 49].

However this discussion is resolved, if the user is to be shown full text, he must have the leisure to evaluate it. To say, "Here's the full text of a document which is probably relevant to your query, but it will cost you \$100 an hour to read it," is unacceptable. Whether FT is delivered in a time sharing environment or on a dedicated personal computer, the cost of viewing its output must be small.

Therefore FT is a system which processes the full text of documents to create document descriptions used to satisfy relevance queries. It will then display full text at reasonable cost.

But note that FT output is a convenient byproduct of the facts that (1) the full text was entered into the system in machine readable form, and (2) storage is cheap enough to retain it in that form. Therefore the essence of FT is not the kind of output, but the significant improvement in searchable structures gained by using full text to build them.

C. History of Search Strategies

Current FT is cut to the same pattern as computerized bibliographic search services. Understanding where the shape of that pattern came from will show how FT may be plucked from the circumstances hemming it in.

Lockheed

In the 1950's research libraries encountered a tough problem. An example is the Scientific and Technical Information Department of the Lockheed Corporation in California.

The quantity of technical documents has mushroomed to such an extent in recent years that manual indexing and retrieval have become

unbearably time-consuming and costly. We have had to find a more expedient and less-expensive method of handling these printed resources and have found that machine documentation and retrieval is the best answer. [Bauer, 1963, pg 570].

Two similar functions were needed: satisfying one-time user queries across the total collection--retrospective searches, and notifying research workers of library acquisitions in their respective fields without flooding them with notifications of all acquisitions--the selected dissimination of information (SDI).

The Lockheed librarians' first answer was the mechanical creation of Keyword in Context (KWIC) indexes of their document titles. When occasional titles were not adequately descriptive of article content, they added keywords or model numbers to them. When they extracted keywords to the left margin and printed the whole title to the right, obtaining a Keyword Out of Context (KWOC) index, they were approaching a uniterm classification system pioneered by Mortimer Taube in the same decade.

Traditional pre-conceived classification headings tend to be compound terms, especially when overcrowding in a category leads to subdivision of its subject headings. Taube observed that the same words in different combinations were repeated throughout a scheme. Why not revert to use of single words, "uniterms," only? [Doyle, 1975, pg 173]. A KWOC index does just this, the

classification vocabulary being nothing other than the "significant" words appearing in document titles.

From this approach each Lockheed user gained a personal copy of the KWOC listing containing only those terms of his interest, SDI. Also multiple copies of the total KWOC listing were distributed to satellite libraries and a user no longer had to travel to the centrally maintained card catalogue to make a retrospective search of the total collection.

Uniterms

But uniterm, or "coordinate classification" has more to offer. Taube developed the idea of reconstructing the compound classification headings at query time by retrieving the logical intersections of the sets of documents indexed under the individual uniterms. For example, rather than classify and retrieve under such a pre-conceived heading as

Wheat cultivation, Madras

(a heading which implies numerous others, "Rye propagation, Texas," etc.) assign the uniterms

cultivation
Madras
wheat

then retrieve

cultivation AND wheat AND Madras.

Without resorting to computing machines, Jonker Business Machines Inc. developed the Termatrex line of Peek-a-boo cards in 1960. There is one card per uniterm in the system. A tiny hole may be punched in any of 10,000 places on a card, the (x, y) coordinates of the punch specifying the identification number of a document possessing that uniterm. Eventually the "Madras" card would contain several holes, one for each document in the system concerning that State in India. When the "Madras," the "cultivation," and the "wheat" cards are superimposed and held up, light peeps through at just those coordinates identifying documents satisfying the ANDed query above.

Jonker went so far as to develop a printer to strike negative images of the Peek-a-boo cards. When stacked with other original cards, this allowed OR and NOT operators in the query [Jonker, 1963]. The system may support as many search terms as one cares to make cards for, but the number of possible documents per card set (10,000) is miniscule compared to contemporary collection sizes.

Punched cards

Hollerith card processing handles up to 100,000-document collections. [Sharp, 1964, pg 492]. It also alters techniques, speeds up searches, increases turnaround time, may cost more, but doesn't change the

searching strategy one bit. The identification number of each document is punched in the first field of the IBM card. Succeeding columns are punched with numbers signifying the uniterms found in that document:

010356 / 3149, 5278, 1143,

where 010356 may be a serial acquisition number identifying the document and the three other numbers represent the terms "Madras," "wheat," and "cultivation." (This conjunction of numbers representing uniterms on a single card representing a document duplicates edge-notched cards of the preceding decade.) The whole document deck consists of one card for each item in the library.

The equivalent of a KWOC index paragraph is obtained by passing the document deck through either a sorting or collator machine and selecting those cards with a certain subject number in any of its uniterm fields. Depending on the machine, more than one pass of the deck may be necessary. The equivalent of Peek-a-boo retrievals is obtained by passing the deck of selected cards through again, comparing each with the number of the next uniterm. The cards finally selected after n passes contain:

term 1 AND term 2 AND . . . AND term n.

ORs may be obtained by processing the master document deck repeatedly against different uniterms in turn. When all the cards selected are stacked together, that deck contains:

term 1 OR term 2 OR . . . OR term n.

NOTs are obtained by matching any deck against the operand, and keeping the rejected cards rather than the selected ones.

It turns out that the processing of the decks is simplified if a document's different uniterms are put on successive cards rather than in successive fields of the same card. All cards will contain only two fields, a document ID and one uniterm number:

010356 / 3149
010356 / 5278
010356 / 1143

If the cards are sorted on the first field, the deck is a document list showing all subjects assigned to any one document together. If the deck is sorted on the second field, the document list is "inverted" into a subject list, identifying together all the documents about a single subject, i.e. a KWOC type index when the numbers are translated. [Faden, 1959].

When IR entered the machine room, it began to look like modern retrieval, but search strategies remained identical to the hand operated systems. Bibliographies made in the machine room differed from those compiled manually from the old card catalogue classification schemes. The reason was not because machines were being used, but because uniterms had replaced subject headings. Assigning

the latter is termed "pre-coordination," multiple search terms coordinated at the time of processing the document; while the use of uniterms is called "post-coordination," search terms coordinated by the user at search time. Note that pre-coordinated expressions carry syntactical meaning ("horse chestnut" is clearly different from "chestnut horse"), but post-coordination carries none ("horse AND chestnut" retrieves both of the above).

Interactive computing

The next step appeared to be natural language queries to replace Boolean connectors and uniterms. In the days of batch processing, a strong linguistic analyzer was desired to resolve the "few" ambiguities that creep into some natural language statements. But a linguistic analyzer wouldn't work and this next step proved formidable until a way around the difficulty appeared.

. . . one change of major importance which opens the way to the construction of various kinds of semiautomatic language processors that might be of immediate practical use. . . . interactive computer systems may be such as to compensate sufficiently for the ambiguities and inaccuracies {in free language} about which science has so far found nothing to say and which render fully automatic systems infeasible. [Sparck Jones and Kay, 1973, pg 42]

After punched cards had expanded the scale of IR, online access increased its speed. The pioneer was

Lockheed; the user NASA. After experiments on the East coast with up to 23 terminals, NASA revised specifications and contracted Lockheed in 1966 to build the system which eventually became RECON (Remote Console system), still in use at the Oak Ridge National Laboratory and several other installations [Doyle, 1975, pg 280]. Starting with one terminal in 1966, the program at Lockheed became DIALOG, by far the largest of the big three vendors, with subscribers around the world [News, 1981, Aug, pg 352].

The government subsidized National Library of Medicine began equally early experiments resulting in the Abridged Index Medicus, a database of medical titles and abstracts, becoming available on the Teletypewriter Exchange Network (AIM-TWX) in 1967. The popularity of this service spurred other online enterprises, but MEDLINE continues to be probably the most used bibliographic database in the world. It is available at low rates through the NLM, and also from BRS (the Bibliographic Retrieval System) where it is called Medlars [Fenichel, 1981, pp 19-20] and, since 1981, also on DIALOG [News, 1980, 4(3), pg 313].

Perhaps from the inertia of current installation investments, perhaps because IR workers come from a Data Processing background--but the queries submitted today to these high speed interactive systems duplicate the punched card strategies of 20 years ago. But user friendly

interfaces are gaining ground in IR. Through interactive dialogue with the computer, the user can convert his natural language query into a form the computer can process. The linguistic analyzer won't be needed on the user end of an FT system.

This review of the continuity of search strategy from hand operated systems to complex online systems is not to claim that post-coordinate searching is unsuited to its new electronic environment. Rather it is to understand it better, and to ask this question: With the dramatically greater computing power available now than in the early days, is there not some way to gain a dramatic improvement in services to the user without losing the power of the old methods?

With its added speed, interactive IR adds the power of iteration to these post-coordinate search strategies. Online networks potentially extend the reach of this service to all who have computers. Finally, here may be a technology to contain the exponential expansion of information publication. While world population grew at such a rate as to double in 50 years (middle 20th Century), the number of new book titles doubled more quickly in seven years. Pulp trees hardly grow that fast, but computer installation capacities have grown at an even quicker rate so as to double in only one year [Doyle, 1975, pg 4].

Information transfer via the bibliographic systems discussed in this section still depends upon the existence of print on paper libraries, but FT, not thus dependent, may be able to harness the computer to this problem which has been escalating since Lord Rayleigh observed in 1884:

What has been once published is spoken of as known, and it is too often forgotten that rediscovery in the library may be a more difficult and uncertain process than the first discovery in the laboratory. [Sharp, 1964, pg 211].

D. The Rising Potential for FT

Suppose an existing, printed book, with its table of contents and index, were keyed onto the floppy disk(s) of a personal computing system. Could it have any advantages over the book? Some simple commands would be:

Show Table of Contents.
Display pages starting at page ____.
Show index.

But to bring some database management power to the system, page numbers could be made optionally invisible to the user. Words go in the blanks instead.

Display pages starting at ____ heading
(as seen in Table of Contents)
Find ____ in index and show entry
(to observe any subheadings)
Display next occurrence of ____ in text
(where term must be a valid index entry, and the system follows page number pointers without ever showing index entry).

Some page thumbing and number remembering are saved, especially for a user who wanted to run several references in the index. But no IR capability is gained beyond that already possessed by the printed volume.

Or consider a user who enters a command on his home terminal which means

Check out ____ book from my public library.

Either the library's computerized catalogue and circulation system determine that the book is presently unavailable or the book may be delivered to him by mail. But no service is gained beyond what the user might do in a simple visit to his library.

All facilities for both these computer applications exist today in various stages of development. Alone neither service seems particularly attractive, but put them together, enabling commands from the home like:

Display next occurrence of ____ in ____ book
at my public library

and the whole pattern of user access to information could change. Some problems, though, appear to be (1) the availability of books in machine readable form, (2) the cost of disks and disk drives, (3) the installation of high band width telecommunications to each home, and (4) the availability of computer terminals.

Four technologies answer these problems and impact the practicality of FT systems:

Photocomposition
Cheap digital memory
The wired community, and
Personal computers.

The first two of these were essential to the development of the present online, bibliographic searching industry.

1. Photocomposition

The replacement among publishers of hot-metal typesetting by photocomposition sustained the growth of online searching [Fenichel, 1981, pg 1-2]. Offset replaced letter press printing as standard technology in the 1960's. Since offset masters are prepared photographically, not mechanically, the application of computer text manipulation to the preparation of such masters was straight forward. Pages to be printed no longer needed to be composed, or "laid out," on paper first. Text, and even pictures, are prepared in digital form only, quickly displayed on high resolution CRTs, and then photographed from the screen to produce the master plates for printing. A hard copy image of a page may never be produced until the first one rolls from the printing press [Doyle, 1975, pp 17-27].

After making the high capital investment in photocomposition equipment, a publisher saves money in several ways. Proofs don't have to be read to catch typesetter's errors. Once a text is corrected in the

computer, it tends to stay correct. A CRT more quickly displays a font that is higher in quality than that producible from an impact printer. Photocomposition is particularly popular with "secondary" publishers--producers of abstracting and indexing volumes--because they must deal in the constant updating of a difficult type of text which is prone to typographical errors.

Now, a free byproduct of photocomposition is a machine readable version of the reference book being published. The only direct expense involved in searching it by computer is that to the vendor in mounting the database and processing queries. Note that the database comes free as long as the sale of the paper product continues to support the text preparation process.

Photocomposition's contribution to the rise of bibliographic databases contributes just as well to FT. Not only are nearly all books, magazines, and newspapers produced in this way today, but authors are writing on word processors. The author with a "third generation" word processor (one which uses a standard computer storage medium rather than cassettes or IBM magnetic cards) uploads his disks to his publisher's system, which adds page formatting commands, and the presses roll [Buckland, 1980]. Today's nearly universal processing of text in electronic form suggests that FT will proliferate as soon as

electronic distribution is cheaper than publishing on paper.

2. The cost of mass memory

The bibliographic record of one document requires from one to two kilobytes for its storage, depending on the length of its abstract. The other fields--identification number, title, author, source, subject keywords, etc.--require less than 500 characters or bytes. Punched card IR systems had very little memory. Operators moved decks from bin to hopper. Tape storage allowed the same batch processing from tape to tape, electronically.

As the size of secondary storage grew, turn-around time decreased. With 750K magnetic drums, it was practical to search and select relevant references from a preliminary document set of several hundred records instantaneously. Yesterday's improvement from drums to 300 megabyte disk packs meant moving the whole database into quick, random access secondary storage. Bibliographic searches become interactive. Tomorrow's expansion of secondary memory capacity, whether through optical disk, holographic fiche, bubble memory, or semi-conductor RAM technology, will do as much for FT as the last magnitude of expansion did for the bibliographic industry.

For example, the removable, 300M disk pack costs about \$600. The disk drive on which it plays costs thousands of dollars, but let price of the medium, about \$2 per million characters, indicate the total investment. Though most books are less than one million characters long, the computer playback of even the average personal library would be rather expensive. In contrast, one million characters may be stored on a length of computer tape costing about \$0.06 [Salton, 1983, pg 306-307].

As an example of the beginning of the new technology, Mnemos System 6000 uses laser technology to record glass masters of a 12 inch disk which holds over 1000M bytes. The glass master makes multiple, read only copies resembling LP phonograph records in appearance and cost [Computer, 1983, pg 91]. The application of this technology may go two ways: regional information centers containing the Library of Congress on disks, or 1000 book disks in the home. A. E. Caldwell [1980], surveying the British scene, expects economics to favor the latter, while developments in telecommunications beyond the telephone may favour the former. As the disk pack didn't just make an old search cheaper but made a new service possible--the interactive search; so the new, cheap storage may make a qualitative difference in IR.

3. The wired community

The only problem with using the telephone for bibliographic searching is the expense of the long distance phone call. Therefore a factor in the popularity of online searching was the availability of cheap, long haul, high speed telecommunications. Phone rates to a local node of TYMNET or TELENET, plus their low tariff, brings a total cross-country telecommunications' bill to \$5-\$8 per hour [Finichel, 1981, pg 41]. The slow, 30 characters per second (300 baud) rate of the local phone circuit allows manageable bibliographic searches because the volume of material transmitted is small.

But FT at that rate won't do. The content of a single typed page takes several seconds of transmission time. A broader band-width channel from local node to home or library is needed, and the entertainment industry has already provided one. The usual cable TV system in any American city has a channel resembling FT constantly photographing weather information, a news wire service, or perhaps community announcements. With its 6 MHz TV band-width, that single channel could provide all the digital telecommunications speed a single terminal could ever use--and TV on the other channels too.

The most outstanding one of these cable systems is Qube in Columbus, Ohio. Perhaps because of the value of

that city as a test market, its cable communicated two ways from the beginning in 1977. (As another instance, the franchise granted by the City of Knoxville in September, 1973, to Athena required the cable company to construct a two-way facility [Knoxville City Ordinances]. But it never has.)

A two-way facility means that TV viewers can send yes/no responses back to the "headend" while a program is in progress. Qube viewers have called the plays for one side in a local football game (viewer's team lost), chosen magazine covers which became best sellers, rated one of President Carter's speeches (generally satisfactory), and starred in NBC's 1980 show, "Speak Up, America," giving instant public opinions.

That same year the Warner owned cable company replaced the yes/no button in 100 test homes with Warner's Atari terminals so that CompuServe, the Columbus based phone network, could offer its many services, including FT of the Columbus Dispatch, via cable [Fortune, 1980]. To the same users OCLC, the Ohio computerized library network, offered Channel 2000, which included a machine readable edition of the Academic American Encyclopedia. Like the book-on-floppy-disk example at the beginning of this subsection, and like CompuServe's World Book, this encyclopedia offered no more access points or sorting

routines than does the printed version [Harter, 1981]. Still this old fashioned information source was the most used database in the OCLC experiment; though, as it turns out, Shop-at-Home transactional services will probably be a more popular application of this technology than FT [Nash, 1982].

However the Qube experience is an exception among cable systems. The basic obstacle to using these cables for high performance data transmission is not technical, and probably not even economic, but inertia and widespread ignorance of the potential [Tanenbalm, 1981, pg 102]. As Johnny-come-lately broadcasters, cable managers are very unadventurous. While the many one-city operations are being bought up by large cable companies such as Warner and Westinghouse, a giant looms to snatch whatever digital transmission market they might have.

AT&T, under its new diversification policy, has experimented in Albany, NY, with electronic phone books and in Coral Gable, Florida, with delivery of newspapers to subscribers [Rochell, 1982, pg 1813; U.S. News & World Report, 1981]. Boosting band-width to the home is a matter of replacing the common twisted pair local phone loop with coaxial or optical fiber lines. Which ever technology arrives first at the home or office door with the new wire,

someone will place high speed telecommunications capacity within the reach of every reader.

4. The spread of personal computers

In 1983, the microcomputer market is booming. Every sale of a personal computer adds one to the number of potential FT terminals. The emergence of this mass market begins to realize the 1945 dream of Dr. Vannevar Bush, President of Carnegie Institute, Director of the WW II Office of Scientific R&D, Dean of Engineering at MIT, etc. His Atlantic Monthly article, "As We May Think," is by far the most highly cited single article in the field of IR.

With uncanny insight into the scientific trends of the 1940's, Bush proposed "Memex," a personal, full text information storage and retrieval system, which housed, in a single piece of furniture the size of a desk, enough capacity to hold documents entered at the rate of 5000 pages per day for hundreds of years. He only missed current technology by failing to see past microforms to digital storage. Basic access was by "trails." Pointer fields reserved at the edges of microfilmed pages could be marked with a keystroke from the user to link different pages he was reading on multiple projection screens. The list of pages just linked could be named, the name stored in a

displayable directory, and the linked list called up again at any time.

Ralph R. Shaw used microfilm to build a prototype of Memex at the Engineering Research Associates in St. Paul, Minnesota, 30 years ago [Doyle, 1975, pg 39-40]. The technology survives today in Kodak's "Oracle" microfilmer and retrieval terminal machines [Kodak, 1976].

The early 1980's has marveled at the complexity of the large scale integrated circuit (LSI) chip which made the personal computer practical. Bush saw such a potential when he marveled at the vacuum tube.

Its gossamer parts, the precise location and alignment involved in its construction, would have occupied a master craftsman of the guild for months; now it is built for thirty cents. The world has arrived at an age of cheap complex devices of great reliability; and something is bound to come of it. [Bush, 1945, pg 22].

5. Other issues

Two routes to FT

Most Memex contents would be purchased already microfilmed, along with pre-associated "trails," ready for insertion. The rest could be entered locally from full size document pages as a modern xerox machine copies.

Though also mentioning facsimile transmission, Bush did not envision Memex content arriving by wire to his

researcher's desk. But now it is obvious that an FT database may either be owned centrally and require telecommunications with the reader, as in current online systems; or be user owned and reside locally, as in Memex. The Radio Shack "Videotex Package" described on page 13, is an example of the first, and disk or other memory expansion to the personal system would provide the second. Which comes into use depends on economics.

A third alternative is suggested by CompuServe's online service of ordering a book to be delivered by mail. An FT system could just as well deliver the full text of the book online to be stored locally. Many terminals, including Radio Shack's, are upgraded by adding the ability to work offline. Queries may be prepared before logging on and pages returned from the vendor may be stored and read later after logging off. The idea is to save connect time.

These two possible residences for the database suggest three possible modes of FT operation, each with its economic implications:

(1) Online databases, implying low telecommunication and connect time charges. Presently the former is cheaper than the latter.

(2) Memex databases, implying cheap secondary storage for personal computer systems. Being a matter of hardware, this mode is more likely to decline in cost than (1) But as

Bush explained, to own the database locally is to run the access system locally and have the trouble of logging and managing the files locally.

(3) The information utility, a compromise between (1) and (2). Suppose connect charges are too expensive for a user to spend time reading online, but personal storage costs are too high to store a great deal permanently. A sought document is located and down loaded as quickly as possible into limited local storage which, after the document is read, will be recycled to receive the next document.

Mode (1) corresponds to library usage (which is in fact where online access has arisen). A user never owns the database, but only has access to it. Mode (2) is like purchasing a book. Mode (3) is a new development of the CRT screen, a way of reading text without buying paper. The user has the option to keep or not, depending on how highly he values the particular document.

Mode (3) may not happen because (a) eventually read-only storage should be cheaper than read/write equipment (though it hasn't happened that way with entertainment tape recorders), and (b) without an access program running locally, downloading may not be so attractive [Kutz, 1983].

Electronic publishers have several options to explore here. For example, excerpts of a new book may be obtained free online as advertising [Horowitz, 1982, pg 91]. Prestel has free advertisement pages. Or, a Memex copy might be purchased for \$10, royalties and all. Or, a full copy may be down loaded, or extravagantly read online, for \$5, refundable when a Memex copy is purchased.

Cost

Economics will determine the further development of FT. The above analysis was based, as are present bibliographic searching habits, on the desire to minimize connect-time, the standard measure for pricing. However, time-based pricing has disadvantages. It penalizes the slow typist and the slow (30 character/sec.) terminal, factors of really little consequence to the vendor or database producer. And it discourages use of the very essence of the service--interaction. The steps of searching, browsing the results, refining the query, and searching again, all consume connect time.

The quickest search makes use of the controlled vocabulary. Here the database producer expends most of his resources in document processing--manually developing the controlled indexing vocabulary and pains-takingly assigning the controlled terms to each document in the collection.

The searcher who comes with a query on the controlled vocabulary is processed quickly but charged little for use of all that resource. But a query of post-coordinated, free text terms requires longer processing, some of which may be string matching. This searcher makes no use of the controlled vocabulary resource, yet receives a higher bill for connect time. Other approaches are to charge for the amount of output received, the number of postings used, or a simple, lump sum subscription to the service.

From November, 1980, the Chemical Abstracts Service shifted its pricing towards: (1) a low connect time fee of \$25 per hour, (2) a typical charge of \$0.10-\$0.25 per item printed offline and mailed, and (3) a new "search fee" for the execution of certain commands, \$14 per search of a sample (5% of the file) database and \$77 per search of the whole database. In 1982, as a further encouragement to interactivity, the sample search was offered free if a full search were made in the same session [Boyle, 1982].

CAS knows that their pricing policy determines how much of the market comes to the information provider and how much to the manufacturer of connect-time reducing equipment, such as 120 chars/sec. terminals.

Migration from print to electronic technology

Another, perhaps greater economic factor, is the impact computer access will have on the sale of documents in print. Database producers will charge more for the former if the revenue from the latter declines. The publisher must recoup a fixed, "first copy" cost from whichever medium. The migration to online use from print has not been readily demonstrated [Lancaster, 1981].

Libraries cancelling their print subscriptions to secondary works don't respond well to publisher surveys about why they stopped [Trubkin, 1980]. Yet the threat of higher royalties for online users and of falling revenues for secondary publishers hangs over the head of the industry.

Copyright

There is also much discussion of problems with copyright protection as digital texts become widely available. The discussion seems uninformed. Computers can easily keep records of who accesses what, bill accordingly, and reimburse authors. Private digital reproduction can scarcely be worse than the present xeroxing of books and pirating of video cassettes.

The problem with video cassette duplication is that studios are attempting to sell their pre-recorded programs on the same medium (magnetic tape) widely available in blank form for individual copying. The video disk promised relief by making the medium (mass reproducible laser disk) plus royalties cheaper than a blank tape. This new approach may not work because it is not cheap enough.

In FT the same situation does not hold. Perhaps a spool of tape can be made to retain as much information as a laser disk, but the disk has one great advantage--random access. That facility is of very little value in program viewing, but of very great value in FT. If these disks prove to be the next FT storage medium, then the economics of the situation may take care of copyright concerns quite easily.

In any case, the technology of text storage and manipulation has gone beyond the requirements of bibliographic searching and towards an area where the retrieval of the full text of anything by everybody with a personal computer is possible.

E. Access Methods

Most users of current bibliographic search strategies are intermediaries. Since more people want full text to read than want bibliographic references, FT must cater to a

wider segment of the public than do current access methods. After a review of DIALOG, STAIRS, and LEXIS, other approaches are described.

1. Current methods on DIALOG, STAIRS, and LEXIS

DIALOG's access depends on two commands, SELECT and COMBINE. The philosophy is for the user to give a command followed by parameters.

User: SELECT bus

Dialog: 1 732 BUS

meaning that the first document set contains 732 documents using the term "bus," which is exactly the same information a user would get by counting 732 pin holes in the "bus" Peek-a-boo card or by counting the number of entries under "bus" in a KWOC index.

User: SELECT buses

Dialog: 2 461 BUSES

User: SELECT fare

Dialog: 3 116 FARE

User: COMBINE (1 OR 2) AND 3

Dialog: 4 53 (1 OR 2) AND 3

There are 53 documents from the chosen database in the indicated intersection of these three document sets. A tedious comparison of different KWOC index entries could have provided the same information. And stacking

Peek-a-boo cards would have yielded 53 pin holes immediately (or two subsets, "bus AND fare" plus "buses AND fare" which adds to 53).

STAIRS takes a presumptive display philosophy, consistent with other IBM interactive setups. The command is assumed and only parameters are typed. After entering the SEARCH mode:

STAIRS: 1

User: bus OR buses AND fare

Stairs: BUS	1104 OCCURRENCES	732 DOCUMENTS
BUSES	895 OCCURRENCES	461 DOCUMENTS
FARE	515 OCCURRENCES	116 DOCUMENTS
RESULT	53 OCCURRENCES	53 DOCUMENTS

"Occurrences" meaning that the specified term occurs n times in all m documents [Hall, 1977, pg 191].

If LEXIS were accessing the same database used above instead of court cases, then the same query would run like this:

Lexis: What you transmit will be Search Level 1

User :bus OR buses

Lexis: Your search has found 1193 CASES through Level 1
To MODIFY your search request, press the M key.

User: M

Lexis: Your search request is:
BUS OR BUSES
Number of CASES found with your search request
through:
LEVEL 1 . . . 1193
REMEMBER to start your modification with a CONNECTOR.

User: AND fare

Lexis: Your search request has found 53 CASES through Level 2.

"Level n" meaning the n-th expression in a compound query. [adapted from LEXIS, 1980, pp 10-11].

All the systems are more flexible than these examples show, for example, DIALOG will take:

SELECT (bus OR buses) AND fare

and return all the information it gave above. Or STAIRS will prompt with statement numbers, accept the terms one at a time, and then process:

User: (1 OR 2) AND 3

These examples are not to illustrate the differences in command structure but to illustrate the amount of information returned during a search and to show that the basic strategies are identical.

The first two have Spartan displays. DIALOG looks like a direct descendant of the manual techniques. The numbers of occurrences given by STAIRS could not be obtained from manual systems, but nothing else is new. The wordiness of LEXIS leads the infrequent user by the hand through his search and he doesn't have to keyboard document set numbers. The "Level 1", "Level 2" idea seems obscure, but may require less training for the novice than would the use of parentheses in Boolean expressions. But in its efforts to become user friendly, LEXIS, has really only dropped some information while changing the search strategy not one

bit. A LEXIS user must still struggle with the application of Boolean operators to natural language text. Note two crucial problems with the Boolean search strategy common to all these access methods:

(1) It searches only for words, not concepts. Yet searching is

. . . essentially a job of handling concepts which, unfortunately, are expressed in words. I say 'unfortunately' because words are ambiguous, change in meaning, creep into the language, and are capable of being most mischievous. For instance, the simple word 'fix' has grown by usage to reflect totally different ideas, e.g., fasten, repair, bribe, narcotics. [Henkle, 1973, pg 121].

When a user has expressed his concept in words, it is very difficult for him to look again at his words and see that another concept lurks there. Unless the user can easily review the different concepts retrieved by his words (i.e. look at some retrieved documents as in interactive searching) his search grows quickly more frustrating as his subject becomes more complex.

(2) Think of the right word, or sink. If a word does not appear in the text of a document or as a manually assigned descriptor attached to it, then that document will never be retrieved for a concept depending on that word. The user must, somehow, come up with the right terms before any searching can be done.

Getting the right word

If the literature to be searched uses its vocabulary precisely, as does law and science, then anyone familiar with the literature, a lawyer or a scientist, will probably be thinking of his query in the "right" terms. Because humanities are not so precisely written, the post-coordinate search has less chance of success in those fields. For example, contrast the probable success of a retrieval using the terms:

Hydraulic system. Brakes. Failure. Leakage.

versus one using:

Women. Managers. Attitudes. Employees.

[Austin, 1976, pg 47].

But the infrequent user is nearly always less effective than the intermediary familiar with the vocabulary of the particular database. This author listened to several research librarians and computer researching lawyers discuss the following information need: A girl on salary had been sent home early from her job one day when her boss said, "There's not much work to do, go on home." Later the boss docked the girl's pay for the time off. Question: form a query to extract cases or statutes from a legal database dealing with a similar situation.

Neither "firing" nor "termination," "disciplinary action" nor "sick leave" are relevant. It's not "labor

relations" because the girl is on salary; "vacation pay" might be relevant. The phrase "sent home" was the best query these experts could come up with. Depending upon the concept involved, a user may be quite at a loss to guess the right word unless he has some listing of authorized search terms in the database.

Both earlier systems provide ways of getting thesaurus help online. DIALOG supports the EXPAND command.

User: EXPAND electric vehicles

Dialog:	REF	INDEX-TERM	ITEMS	RT
	E1	ELECTRIC VEHICLE DEVELOPMENT	1	
	E2	ELECTRIC VEHICLE DRIVE	3	
	E3	ELECTRIC VEHICLE DRIVES	1	
	E4	ELECTRIC VEHICLE PERFORMANCE SIMULATION	1	
	E5	ELECTRIC VEHICLE PROPULSION	2	
	E6	ELECTRIC VEHICLES	288	8
	E7	ELECTRIC VEHICLES APPLICATIONS	1	

This page from the alphabetical thesaurus of the selected database shows that one of the terms is generic, therefore:

User: EXPAND e6

Dialog:	REF	INDEX-TERM	TYPE	ITEMS	RT
	R1	ELECTRIC VEHICLES		288	8
	R2	ELECTRIC DRIVES	R	1177	7
	R3	ELECTRIC LOCOMOTIVES	N	249	6
	R4	ELECTRIC PROPULSION	R	179	5
	R5	ENERGY STORAGE	R	333	4
	R6	LINEAR MOTORS	R	478	3
	R7	MOTORCOACHES	N	30	3
	R8	VEHICLES	B	3180	6

showing not only terms related horizontally, "R", to "Electric Vehicles," but also those which are "B"roader and "N"rrower than it in the hierarchy of the database's thesaurus. [Hall, 1977, pg 165]. The "EXPAND term" shows

an alphabetized section of the thesaurus, while the "EXPAND REF n" shows a section of the classification tree of the thesaurus. "Items" in either display is the number of documents using that term.

STAIRS does not offer the second expansion, only the first by the round-about use of a truncation search:

User: comput\$

Stairs: COMPUT\$	4280 OCCURRENCES	
COMPUTE	59 OCCURRENCES	57 DOCUMENTS
COMPUTER	1104 OCCURRENCES	625 DOCUMENTS
COMPUTERISATION	1 OCCURRENCE	1 DOCUMENT
COMPUTERS	2987 OCCURRENCES	601 DOCUMENTS
COMPUTES	1 OCCURRENCE	1 DOCUMENT
COMPUTING	128 OCCURRENCES	111 DOCUMENTS
RESULT	4280 OCCURRENCES	806 DOCUMENTS

[Hall, 1977, pg 194]. Thesaurus help is a very valuable aid to user comprehension of what is in the database, but getting and understanding that help is not straight forward in the typical examples shown here.

Using the right word

All three systems have operators specifying proximity in one of two ways: (1) that terms be in the same prose unit (ADJ--terms immediately adjacent, WITH--in the same sentence, or SAME--in the same paragraph) or (2) that terms be within a given number of words of each other, regardless of intervening punctuation (e.g. W/7). Luhn [1957] used both of these approaches in his pioneering work on automatic abstracting.

Also, all three systems support some Data Processing-type searching on formatted fields. In DIALOG one may write as his term "AU=O'Kane" to specify the desired value in an author field, or "Mumps/TI" to specify that the word "Mumps" must appear in a title. These expressions may appear anywhere terms appear and the document sets created join easily into the COMBINE logic. [Hall, 1977, pg 167; Salton, 1983, pg 33].

STAIRS, however, requires the user to enter the SELECT mode, extract a numbered document set from his formatted field search, then enter the SEARCH mode where the above commands work and where the carried-over document set number may be used in search logic. LEXIS allows restricting searches to opinions of certain judges or to cases appearing within user specified date ranges.

2. Ranking

The newest features of these systems is the ranking of documents on the basis of word frequency counts. When it displays the number of OCCURRENCES, STAIRS is reminding the user to specify the RANK command and to choose one of the several available ranking functions to determine the order in which his documents will be presented. The ranking function may be as simple as, "Show first the documents with the highest number of occurrences of the query terms,"

or as complex as a cosine coefficient function which treats documents as vectors in Euclidean space and calculates the angles between them. STAIRS has four of the simpler functions built in and allows the IR manager of an installation to specify other functions. Such a variety is excellent for a community of users who will work carefully to determine by experiment which function is best for their particular database and needs. But, in fact, most STAIRS users forget to invoke RANK [Quinn, 1980] or do not want to wait the extra 30 seconds required for the output document set to be sorted [Martin, 1973, pg 58].

Neither DIALOG nor LEXIS have a RANK command, which means that system designers are free to establish whatever ranking function they like as part of every TYPE or PRINT command. Most users never know whether the service is being provided or not. The difference is that the service appears to work better when its best guess as to relevancy is presented first. Westlaw, which has ranked retrievals automatically all along and always labeled its documents as "Rank 1", "Rank 2", etc. instituted, in its sweeping 1983 changes, the option of the old ranking based on word frequency or of a new ranking in reverse chronological order.

By use of its ranking algorithms, a system can process any query and find the one best document, the second best,

and so on, independently of query complexity. If only one term were given, the system could rank all documents in the collection on the basis of the frequency of that one term in each document. If ranking algorithms become good enough and fast enough, a system could rank every document in its collection by its presumed relevance to any query.

Ultimately a query could dispense with Boolean operators and consist of lists of terms plotted by the system as a point in n -dimensioned space. Relevancy would consist in shortest distances between the query point and document points plotted in the same way. Algorithms and functions for doing this have been well developed in the field of Pattern Recognition. The techniques will be discussed later in Section III C., Statistical Analysis of Text; but their applicability depends upon the assumption that content can be represented as points and relevancy represented as some function of those vectors.

3. A simpler system and the public

Hardware is improving constantly, but search strategies haven't changed much since the days of punched cards. A former president of the ACM said:

The processors, the fantastic memory devices, the inexpensive but sophisticated terminals, are available right now, and will become even better year after year for a decade or more. The appropriate communications channels are

somewhat more problematical
 Nevertheless, what is needed will come, and at attractive prices--in the end. What is much, much harder is the software Each year the hardware improves; each year the software seems less satisfying. Nor do I see much improvement ahead. There is an almost impenetrable barrier between man and machine [Herbert Grosch in forward to Hiltz, 1978].

Individuals, called programmers, make up the difference, but to the public, where lies the future of FT, the machine must penetrate the barrier. A common, observed reaction of a naive user upon entering a query into an IR terminal is this: Scan the computer's response for an immediate contribution to his information need--a highlighted query term or the size of his document set. As soon as something appears resembling either (1) the answer sought, or (2) anything else; the verbal responses are, respectively: (1) "What do I do next?" or (2) "It doesn't work." They will not take training courses voluntarily. They will not read manuals. They will not even read documentation when it appears on the CRT in front of them. Only a system which performs "naturally" according to the expectations of the naive user will succeed in communicating with the public.

An example of a system which does this is Search Helper, installed winter, 1983, in the Undergraduate Library at UT. Running on a small Commodore personal

computer, the program from Information Access, Menlo Park, Calif., accesses DIALOG by making a single telephone call:

Search Helper: Hello. I am Search Helper.
I can help you with your research needs.
See your librarian for a password.
{Currently passwords are free.}
Type password and press Return.
{Similar such instructions appear
in each frame, but are omitted here.}

User: (password)

Search Helper: Would you like information about a
1. Person
2. Place
3. Subject
4. Brand Name?

User: 3

SH: Type the first word of your subject:

User: bus

SH: Type the second word of your subject:

User: (Return)
{Though undocumented, experience has shown that this
prompt expects the second word of a two-word term.
Queries with one-word terms leave this line blank.}

SH: Would you like to add another
1. Person
2. Place
3. Subject
4. Brand Name
to your query?

User: 3

SH: Type the first word of your subject:

User: fares

SH: Type the second word of your subject:

User: (Return)

SH: Would you like to add another . . .
 . . . Type 6. to continue.

User: 6

SH: Would you like to search
 1. Index of Magazines
 . . .
 6. Most recent additions from the latest month.

User: 1

SH: In your next search, other references
 might be found by selecting Newspapers
 including the New York Times and the
 Wall Street Journal (1979-last month)

During which the Commodore is telephoning DIALOG. Usually
 the user must enter the following loop.

SH: Telephone lines are busy.
 Would you like to try again with the same search?
 Type "Yes" or "No"

User: Yes

SH: In your next search, other references might be found . .

And the user sits, feeding "Yes"s every 30 seconds until
 the call is completed. Not to do so means termination of
 the program and starting over later. Finally,

SH: Congratualtions!! 53 references have been found.
 Finding them took the computer 28 seconds.
 The most recent 15 will appear on the printer.

SH: Take the print out to the librarian
 {who will help you find the documents.}

SH: Hello, I am Search Helper.
 I can help you with your research needs

There are a couple of variations in this exciting plot. The
 user's query may net zero retrievals:

SH: No references were found.
Try again. Better luck next time.

SH: Hello, I am Search Helper

Or, when given the option of

SH: Telephone lines are busy.
Would you like to try again with the same search?
Type "Yes" or "No"

User: no

SH: Hello, I am Search Helper

Advantages to this interface are significant. The query formulation process has been broken down into several small tasks, each of which is prompted with an instruction so short that the user has read the whole screen before he can help himself. Then a user response is necessary to continue, whereupon another short instruction appears. Also dead time, such as waiting for the phone connection, is used to offer another bit of instruction or advertisement, "On your next search" And giving every user 15 references, thermally printed on a small form, simplifies output considerably. While free, Search Helper is popular with patrons of our Undergraduate Library.

The disadvantages are serious. Not only are the displayed instructions brief, the initial documentation added nothing to them. If a two-word query term is entered in response to the "Type the second word of your subject:" prompt, the system evidently constructs a "word1 PROXIMITY CONNECTOR word2" clause. No one has yet discovered the

exact proximity distance. As entered above, the actual query was probably "bus AND fare".

But that is the total capability of this little system. No ORs available. And, though Search Helper uses a computer and looks like an online, interactive system, it is not. The user interacts with a simple local program to prepare a DIALOG query off line. Then the local program connects to California and submits the invisible query for a batch run. Though useful for introducing the idea of computer searching to the public, Search Helper is frustrating to anyone with any background in his literature or in computer searching.

For instance, if one does happen to get a good query that produces 15 good references, how will he get more instead of the same 15 again? As it happens, Search Helper supports the entry of DIALOG's "PY=" terms. If a user will read forthcoming documentation, he can supplement his working query each time with a different publication year parameter, getting 15 from each year. But "PY=" is not very user friendly.

Search Helper is, perhaps, an ultimate example of sacrificing power and flexibility for the sake of naturalness. However, it is not as restrictive as CompuServe's menu approach.

4. Menus and the public

In her analysis of Prestel and CompuServe type access methods (examples on page 6 and 13), Nash sets a wise goal for FT access methods:

The future panacea is thought to be keyword access; whether this will be so remains to be seen, as many synonyms, antonyms, singulars and plurals and like words would have to be indexed. Perhaps a solution will be to offer both methods of access, menu and keyword, geared to the level of expertise of the user of the system. [Nash, 1982, pg 295].

Is menu selection a useful solution? Exploring CompuServe's menu tree the first time is fun, but becomes tedious when a known page is wanted but its page number (to use a in GOTO command) is not known. More frustrating is the knowledge that each word in the database is machine readable, but there are no tools which make use of that fact. A user cannot get the machine to scan a whole magazine article for one word; he must read it all for himself.

Menus are pre-coordinate classifications of the pages they represent, and the number of subcategories at any one node is limited by the number of menu entries listable on one CRT page. Neither can any steps in a menu search be skipped. Therefore, the menu way of presenting a classified document collection is more restrictive than a card catalogue, but also more "natural" to use.

Using a pay telephone is, by now, natural for almost anyone; but the new bank teller machines were not when tested in 1971. A typical reaction was, "They will need a third generation of machines that are no more complicated than making a long distance telephone call." This was a machine requiring only one choice of function followed by the keying of two sets of numbers.

Librarians planning to present the wonders of key-word searching and Boolean operators may need to think again if they have to produce machines that are 'no more complicated than making a long distance telephone call.'
[Lawrence, 1983, pg 106-107].

Only by his amazing problem solving ability does an uninformed human ever get any satisfaction from an unnatural computer interface. The naive user will listen to another person give specific directions about how to solve his immediate problem. He might listen to a synthetic voice. But composing Boolean search expressions is not "natural." FT systems must find ways, such as Search Helper has attempted, of enlisting the power of the Boolean search strategy without demanding its discipline of the user.

However, today these second generation bank teller machines do a thriving business. The public can learn. Perhaps President Grosch (page 61) was wrong. The humans may win after all.

II. FT IN PERSPECTIVE

This Section compares and contrasts full text systems with related IR concepts. After describing the information handling power which could be developed, it speculates on the implications of this power.

A. The Narrow Perspective

FT's two nearest neighbors are bibliographic searching, which returns a reference to a hopefully relevant document, and question-answering (Q-A), which returns answers written by the machine from a knowledge-representing data structure. [Salton, 1983, pg 264, 287]. An FT system stores strings of characters which the user interprets as documents written by humans. A bibliographic system stores strings of characters which are interpreted as proxies or descriptions of documents which have an independent existence outside the system. A Q-A system stores special data structures which the human designer interprets as representing ideas.

1. Distinguishing bibliographic services from FT

DIALOG, Search Helper (see pages 52 and 63), and most systems operating today access bibliographic databases. The user receives pointers which he must "take to his librarian

(see page 64)" with the hope that the documents themselves may be found. To save storage in the computer, bibliographic systems reduce each document, upon its entry into the system, to an essence of title, abstract, or keyword list. This document processing is usually done manually but could be done mechanically.

The pioneers had other ideas. Bush's [1945] Memex assumed a full text database and proposed page-wise access methods which took no account of the words on the page. Luhn [1957] advocated statistical processing of word frequency counts derived from encoded full text with the purpose of automatically deriving abstracts. The bibliographic system has been a profitable exercise in striking at the heart of the IR problem--relevancy processing--without fulfilling these pioneers' intentions--full text retrieval.

Having been present during the introduction of an on-line literature searching service in an academic institution, I vividly recall the initial enthusiasm of the prospective users of the service at the possibility of obtaining references automatically by computers. Their enthusiasm diminished quite markedly when it was explained that only references to documents, rather than documents themselves, would be retrieved." [McMurdo, 1980, pg 349].

It would be logically trivial for any existing bibliographic system to follow its pointers and retrieve the full document. But it would be a practical nightmare of mammoth storage additions and horrendous text acquisitions.

The new facility would be called online document delivery," distinguishing it from today's mail, facsimile, or interlibrary loan document deliveries.

Two types of FT system

The full text CEEFAX, CompuServe, and Prestel are Memex type systems. They make online delivery of known documents. No real relevancy processing takes place. Menu items are chosen until a document or page becomes "known." Selective display of a sort is available when showing only the first page of each document.

By allowing DIALOG style relevancy processing, the full text Info Globe, LEXIS, and NEXIS achieve Luhn's advance of processing words rather than pages. Search terms are no longer limited to the dozen or so words assigned to each document in a DIALOG type system. Any "significant" word in the text may appear in a query. Such "free language" searching actually lays more burden on the searcher to guess what words actually occur in his target document. He may have more or less success than with a controlled vocabulary search.

In addition, this second type supports the proximity operator. Rather than selecting documents which contain "reactor" and "core" anywhere in their text, the proximity

operator requires that terms appear within a certain distance of each other. For example:

reactor W/3 core

would pick up both "core of the reactor: and "reactor core" but not "The core of the matter is that a reactor" Without the proximity operator, the greater number of terms FT puts into the description of each document would yield many more false drops [Lancaster, 1976, pg 33].

Other advantages of FT

Working from FT, users may select for display only those pages containing the query terms, and those terms may be highlighted. Note that highlighting means that the same page is displayed differently in response to different queries.

The user values the system more highly because he no longer has to chase his document elsewhere, with the real possibility his library won't have it. When he finds the document, he doesn't have to manually search every page to find the keyword passages.

Since all systems offer online or offline hard copy, archiving the printed document becomes superfluous. Then libraries have the opportunity to escape their storage problem. Universities don't have to build more library

extensions, which in the past have been inadequate before completion anyway [Gore, 1976, pg 165]. Electronic publishers will save costs all down the line, from ink and press personnel to forest management and logging operations. In other words, such a conceptually simple extension as adding full text to current systems could solve the "paper explosion," while supporting more exhaustive searching. Besides

the greater convenience of selective, online document deliveries,
more exhaustive indexing by use of every significant word in the document,
the greater precision of proximity searching, and storage space saving,

a new capability, only possible with FT, is now suggested by this author.

A new possibility

Just as the user makes more use of the system by remaining online to read his full text; so the system may gain more input from the human by observing his response to its output. The system may get an accurate evaluation of its own retrieval performance by timing the user's interest in each page of FT, or by noting if hard copy is ordered, or by requesting simple evaluations:

System: Is the above page
1. definitely useful,
2. possibly useful,
3. not useful, or
4. already known?

And the user makes his evaluations on the basis of what he really wants--the full text--rather than on the basis of the partial information contained in titles and abstracts.

Each instance of automatic relevancy feedback is, admittedly, very specific to one query using a few terms; but it is of very high quality, coming as it does from a human who is aware of the shifting subtleties of current natural language usage. To profit from this type feedback, the system will automatically modify its thesaurus or classification scheme. This means assimilating lots of tiny details, but such processing is the forte of computers. Automatically processing this relevancy feedback (see Section V.) probably adds an order of magnitude to the computational complexity of a bibliographic access system. But the definition of FT (see page 22) expects such complexity.

Salton has experimented with actually changing the document descriptions to conform more nearly to query statements to which they are judged relevant by users. This is also an automatic processing of relevancy feedback, but with a different purpose. To shift the description of a document is to record the shift in its interpretation by users. [Salton, 1983, pp 145, 244]. The automatic processing here suggested would instead shift the classification scheme so that all documents, new or old,

become more accessible. Access points change, not single document descriptions.

Online search intermediaries presently solicit relevancy feedback when they send week-after questionnaires to their users. But whatever its accuracy, this feedback never reaches the retrieval system, only the intermediary [Schabas, 1976]. Not to incorporate relevancy feedback is to overlook one of Luhn's leads:

It seems to be an inescapable fact that the less disciplined the language, the greater the human effort that must be expended somewhere in the system. . . . Excessive editing (on input) obviously increases the likelihood of bias due to current interests, experiences, and points of view. In consequence the usefulness of the system will be reduced as emphases and interests change. It would therefore appear that the less information is classified and contracted at the input, the more it will lend itself to dynamic interpretation at the output phase. [Luhn, 1957, pg 310].

The least contraction of information at the input is to deal in full texts. The most direct way of assaying changes in "emphases and interests" is to let user response to today's search influence tomorrow's. When the access system changes under usage, it is helping the contemporary user expend "the greater human effort" needed to analyze its database. Each user comes to a collection of flexibly classified, uncontracted documents. The human effort he expends updates classification, helping the next user,

while at the same time he gains satisfaction in the most dynamic way of his own information need.

Therefore, the logically trivial distinction between retrieval of references and retrieval of FT means an inescapable order of magnitude increase in storage requirements, and a possible order of magnitude increase in computation. The user gains (1) an obvious extension of retrieval service, (2) more flexible query processing, (3) selective display, plus (4) an invisible facility of access method updating. All this gives FT a potentially qualitative advantage over bibliographic retrieval.

2. Distinguishing FT from question-answering systems

Just as FT exceeds full text Memex by processing the words on the page instead of whole pages, so Q-A systems exceed FT by processing the meanings expressed by the words. Bush [1945] would capture a page containing the words containing the idea. FT would capture the words containing the idea. Q-A systems would capture the ideas themselves. An FT system displays documents; Q-A digests and evaluates them. This, at least, is how the user views the difference.

Properties of Q-A

When a user submits a question to the Q-A system, he receives a statement which answers his question, much as he would from a human expert. The system has converted his question into a representation of the ideas it contains, fetched stored representations of the same ideas, and combined these logically to transform the query into a statement. This new statement is translated into user readable output.

Salton [1983, pp 287-8] rightfully considers Q-As an extension of Data Processing but wrongfully considers them an extension of IR. Q-As are really different animals from bibliographic and FT systems [Hafner, 1981, pp 6-7]. These two respond with originals: "Here is a place where the author says ____"; while the inventory program, the payroll program, or Q-A will answer with a processed version of the original input: "The computer says ____." Many typographical errors will not affect IR systems, but the Data Processing and Q-A system must be careful not to lose a decimal point or a negation sign. [Lancaster, 1975, pg 34].

The problems of Q-A

All the above is more easily explained than done. The MLT experimenters of the 1960's, who started so

enthusiastically to build machine language translators and failed so completely, did succeed in proving the enormous complexity of converting a natural language statement into a representation of the ideas it contains.

However several limited "expert systems" have been constructed. LSNLIS answered questions about mineral content in rocks brought back from the moon by Apollo astronauts [Salton, 1983, pg 294]. INTERNIST and MYCIN aid in medical diagnosis, and DIPMETER ADVISOR and PROSPECTOR analyze geological experiments to locate oil and mineral deposits [Nau, 1983, pg 83]. LIRS uses the same principle to analyze questions of the law concerning negotiable paper. The fact that LIRS outputs not machine written answers but document references and summaries makes it a forerunner of future document retrieval systems which use Q-A principles [Hafner, 1981, pp 6-7].

One problem with "converting a user's question into a representation of the ideas it contains" is the ambiguity in many natural language expressions. For example: "Tomorrow they will fly the packages to Atlanta." The human quickly disambiguates such sentences by knowing who and where "they" are and knowing if the packages contain airplanes. For an expert system to process such expressions, it may have to devote more of its cognitive data structure to such entities as airplanes, cities, and

other common knowledge (the larger-context, see page 22) than to the field in which it is supposed to be an expert. [Salton, 1983, pg 266].

By restricting themselves to a small number of entities and to a few relations among them, expert systems, including the six mentioned above, eliminate many of the ambiguities which otherwise would plague their conversations. For instance, to the lunar system, LSNLIS, the word "sample" must refer to one of the rocks on which it has data, and it wisely accepts no other meaning [Salton, 1983, pg 294].

Then why not restrict Q-A input to a controlled language and avoid the work of disambiguation? The legal system, LIRS, for example, accepts such queries as

FIND CASES (DESC DISHONORED) AND NOT (PL HOLDER)
which looks like a query to a standard Data Processing database. DESC means describing and PL stands for plaintiff, so that the initiated will understand this request to mean: Find cases which describe dishonored instruments but do not have the holder of one as the plaintiff [Hafner, 1981, pg 92]. By forcing the human to take the first steps in translating his request into machine understandable form, the system has reduced its ambiguity problems to a manageable level but has also

restricted its clientele to those who will do this sort of pre-translation.

Q-A is more ambitious than FT

A Q-A system should strive to accept natural language requests, produce natural language output, and hide from the user translations, conversions, logical manipulations, classifying, and searching. This level of service would elevate an SDI service to a "surrogate memory." The "knowledge" base continually grows as the system processes incoming documents; but the user remains "unconscious" of them, depending upon his profile program to bring to his attention, or "consciousness," items in which he would be interested. He also depends on the system to provide adequate access points so that he can be sure of success in any retrospective search. In using such a system,

The ability of a human to reason is enhanced--he is freed from distracting, nonproductive activity {such as filing and retrieving}, and given an aid to one of the weakest parts of {his} cognitive system--long term memory. [Sauvain, 1970].

The user would depend upon the "surrogate memory" Q-A system to replace his attention to the complex activities of browsing through literature and associating new ideas with old.

FT does not attempt to provide such a complete service. The user is expected to use a query language

easily interpreted by the machine. He is looking for passages which use term 1, term 2, etc., not for passages which say "something like ____." When he finds, through system term expansion or through browsing, another term which also might do well in his query, he adds it, intending that the system will now seek passages which also contain that word specifically. (To the extent that an FT system automatically introduces additional search terms related, by the system, to the given query terms, to that extent does it attempt to imitate by table look-up the cognitive work of Q-As.)

The user must approach the two systems differently. Q-A intends to take the user's question and give him an answer [Hafner, 1981, pg 3; Sparck Jones and Kay 1973, pg 175]. Some FT users approach their systems with questions, but they must translate their questions into requests for documents containing certain words. Not, "What is ____?" but "Get me a document with talks about ____." When the passage is retrieved, and the full text quickly displayed, the context of the sought words may in fact answer the user's original question, but not because the system was clever. When an FT appears to operate as a Q-A, two things have happened accidentally, (1) the system happened to contain a passage which in a few words would answer the question, and (2) the user happened to probe with enough of

the right terms to distinguish that particular passage from all the others the system could have retrieved.

If the FT system had not contained that particular passage, the human would have had to synthesize several passages to get his answer. In contrast, the Q-A system specializes in synthesizing. It may add to or modify its representation of an idea when it receives input containing new information about that idea. And it will do nothing when the content of new input merely duplicates what the system has received before. This need to synthesize is another reason current expert systems restrict their subject matters. The more concepts a system holds, the more complex are the modifications in its storage when new information arrives [Hafner, 1981, pg 12]. Likewise, the fact that Q-A equates its one representation of an idea with the several ways of saying it in natural language means that Q-A storage may be much smaller than FT storage. However, the processing complexity is greater.

Reasonable expectations for current IR

Such a synthesizing system has been sought from the beginning of computer IR.

Machines will be really useful when they can do some analysis, interpretation and synthesis of information. An example would be finding new relationships not brought out in any single document, but putting together information from two or more documents . . . grouping items of

information in new ways not conceived by either the original author or the person who entered the information in the machine. The encouraging thing is that there is good reason to believe that machines can be made to do this sort of thing. [Goodwin, 1959, pg 48].

{The 1960's} information-retrieval methods . . . stop short of determining how items of information are interrelated and how such items can be fitted together for making decisions. The situation is particularly difficult when the information is highly fragmented and when the absence of a previously well-established picture means that there is no framework of reference to which the individual fragments may be related

The inability to correlate facts by machine is most felt in areas of long standing problems (e.g. leukemia) or quick facts (e.g. satellite cloud pictures). [Perry, 1960, pg 301].

Who will organize the flood of knowledge? Needed is a system for the experimenter on the frontier of knowledge, attempting to draw together strands of different disciplines for some new theory or application.

The Library of Congress or Dewey can't keep up anymore. The penchant of modern science for inter-disciplinary work has wrecked some categories in the traditional classification schemes.

Artificial intelligence techniques may someday synthesize materials, but current Q-A systems are far from the goal of "putting together information . . . in ways not conceived by . . . the person who entered the information in the machine." Today's Q-A designer manually fabricates

the data structure by which his machine understands its ideas. The system cannot associate ideas the originating human did not design its data structure to associate.

This leaves people as the only currently available information organizers. Memex aimed to preserve at the touch of a button the new association just discovered by its human user. What is currently possible is a system which will help users draw together ideas which have never been associated before. FT offers the flexibility Luhn suggested (see page 75):

. . . the less information is classified and contracted at the input, the more it will lend itself to dynamic interpretation at the output phase.

FT's freedom from pre-classification and contraction supports a user's desire to draw dispersed documents together. FT will not attempt to organize the flood, but will free the user from index scanning and page turning so that he may organize.

B. The Broad Perspective

The narrow perspective on FT has discussed the technology of what can be done. Here will be discussed what is, should be, or will be done.

1. A model of information cycles

Among nearly every community of information users there is a subgroup who produce literature of interest to their peers. Therefore the two-way driving force of any information flow is an information producer seeking to speak to a group of information consumers, and a subgroup of those consumers seeking to receive some of that information. (And, producer and consumer often exchange roles.) FT is only one of many possible routes from author to reader.

Route One

Personal communication is the oldest route. By letter, meeting, phone call, and other informal means, colleagues exchange information. These contacts form a network called the "invisible college" which, after centuries of existence, continues to operate very efficiently among scientists. In IR terms, a user consults a known producer.

The closeness of this college depends upon the kind of science and its stage of metamorphosis [Doyle, 1975, pg 386-7]. Personal communication is really internal to the user community, but Figure 1 draws this route externally to show its relation to other routes.

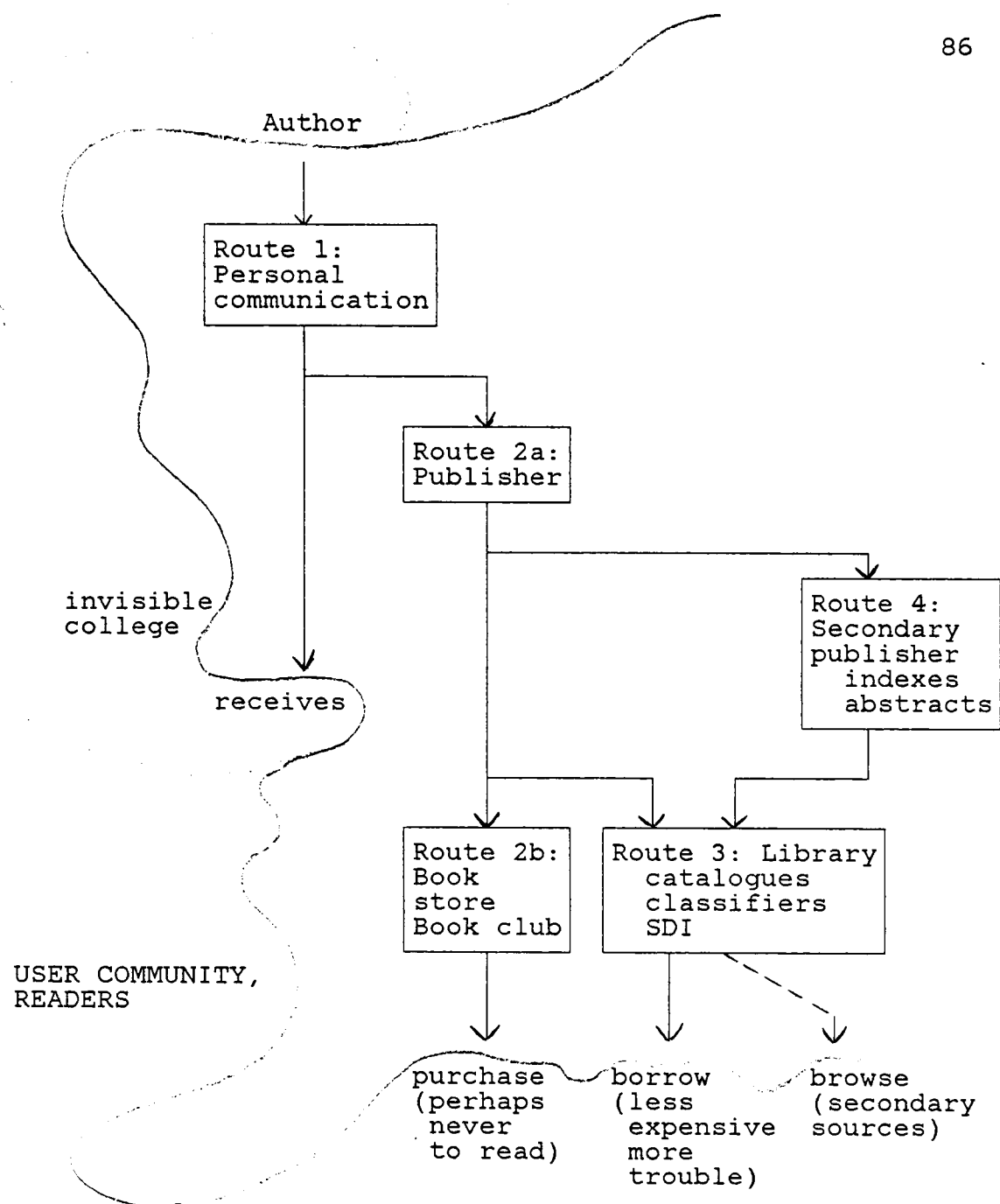


Figure 1. Existing print information cycle.
(with reference to Lancaster [1977, pg 20])

Route Two

A second route emerges when an author includes a third party, a publisher, on his mailing list; and a buyer or subscriber reads the publication because he values the author or journal. These readers may or may not be members of the invisible college. This second route is one step removed from the directness of personal communication, but it opens the author's communication for access by the public.

The word "publish" comes from Latin publico, to make public. In Colonial days, the information producer might not just write his statement, he might also typeset it and operate his own printing press. Then the technology of the process encouraged authors to leave the printing to others. With a monopoly on large scale reproduction of images of the author's original statement, the publisher became the author's sole gateway to the public, as seen in Figure 1 He also provides adjudication and advertising services which encourage some readers to purchase materials beyond what they might otherwise do on the strength of the author's name alone. In IR terms, a Route two user is acquiring a known document.

Route Three

A third route of communication may occur because a library purchases a book or journal and users receive access there. A fourth party has entered the cycle: a document holding and cataloguing library, which may even issue SDI notices. As an additional advantage, at the library a user may stumble across related material in the process of locating a known document. Books shelved nearby or cards associated in the card catalogue may be of interest to a reader. Also, the library, through its classifying and cataloguing, provides additional access points through which a user may obtain relevant material.

Route Four

A fourth route from information producer to consumer requires more active searching on the part of the consumer. Some are not satisfied by the limited catalogue resources of their library and wish to consult additional indexing and abstracting services for pointers to documents of potential interest. This route is the most complex. A secondary publisher becomes a user in Route two and, from several primary documents, produces a new document. Nearly all of these indexing and abstracting works are now so expensive that only libraries buy them for shared use [Collison, 1971, pp 76-77]. A reader will come to the

library, consult the secondary volumes, and then may or may not complete the information cycle by obtaining the document. Routes three and four make retrieval by subject practical.

Stability of Routes

Studies of different groups show that patterns of information use are not intuitively obvious. Engineers in one study had the following preferences for satisfying specific queries: (1) piggybacking--hoping the answer would come down their usual information channels, (2) ask a friend, even if an expert were available, and (3) commence a literature search. [Christie, 1981, pg 159]. Of the top ten sources of information for crop farmers, the first four were people, company field men, neighbors, merchants, and government extension agents, in that order, followed by media, of which government publications received first preference. [Christie, 1981, pg 161]. Perceived ease of use versus expected value of information obtained also affects a user's choice of channel. For instance, housewives in search of a receipt chose their source on the basis of expected return, while professionals chose on the basis of ease of use. [Christie, 1981, pg 165]. Therefore, routes of information flow will change if computer technology makes all sources equally easy to use.

However, in a more surprising study, Kahen found that first-rate scholars obtain most of their references from memory, and only engage literature searches to verify their memories. Kahen suggested that the availability of online searching makes no significant difference in productivity or scholarship [Kahen, 1981].

Routes one to four developed in roughly that historic order, with the rise of secondary publishers being most recent. The progressive complexity of these access routes indicates what other factors show: information exchange via print is increasingly inefficient. One estimate of the volume of literature is 50 million book titles, total, printed since Gutenberg, but with additions currently running at the rate of 400,000 books and 300,000 journal volumes every year [Gore, 1976, pp 171, 173]. Considering only annual additions in English, 50,000 books and a higher number of periodicals appear [Collison, 1971, pg 4]. Journal publication is agonizingly slow, especially in the social sciences and humanities. Only two out of ten good articles see print; and, when accepted, an article is still one year away from publication. The situation is bad, both for information flow and for careers [Seiler, 1981, pg 76]. Another expression of the same opinion is in the popular maxim, "If it costs less than \$100,000 to find out by experiment, then do that rather than try to find it in the

literature." [Boodson, 1976, pg 275]. The move to FT will come because print doesn't work well anymore.

Picturing the first application of computers to Figure 1 would show a simple enlargement of Route four, Secondary Publishers. Vendors now secure their bibliographic databases from secondary publishers, or similar agencies, and a user goes to his library for access. Route four is already mechanized.

As a holder of full text, FT threatens to replace the library. If machine readable texts can be secured from publishers and access programs made powerful enough from enough terminals, then Route three is mechanized. If the access programs are really powerful enough, operating on, or seeming to operate on, the full text, then the role of secondary publisher has vanished. Lancaster predicts that primary and secondary publishers will merge so that by 1990 secondary services will be derived automatically from the primary publisher's photocomposition database. If electronic publishing develops, then libraries may be replaced by vendors who maintain access to the output of several publishers [Lancaster, 1982].

In Figure 2 an author's personal communication with his invisible college remains constant. Electronic mail may add one more to the number of informal means already practiced. But now the author, using a computer

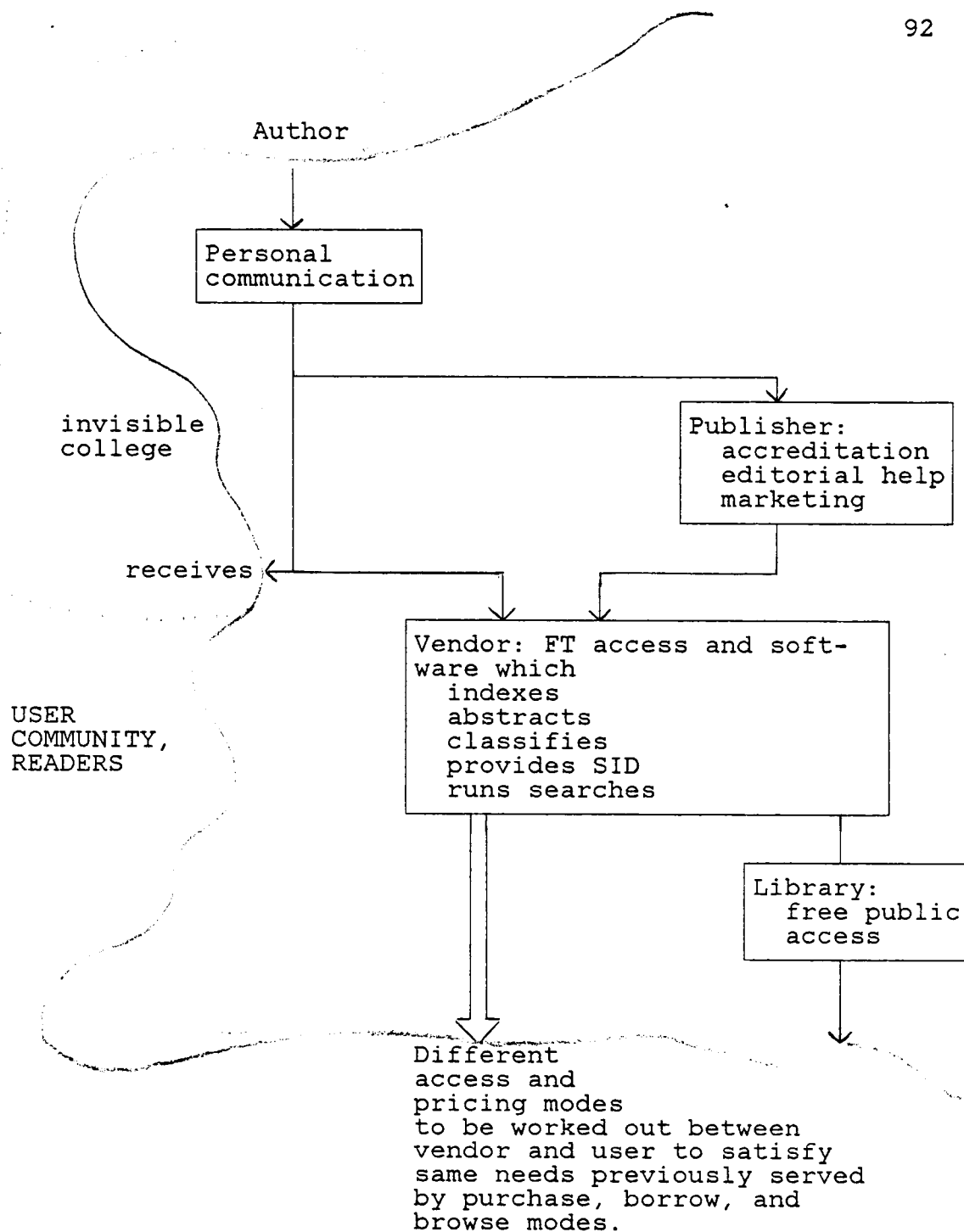


Figure 2. Possible electronic information cycle.

compatible word processor, has additional options. He may upload to a publisher for editing, proofreading, and other services, including vendor interface; or he may upload directly to a vendor (e.g. Newsnet, see page 19), with perhaps the option that the vendor will even handle, via restricted access, some of the personal communication with his invisible college.

The vendor is now gateway to the public, and his services are (1) maintaining the databases and access software, and (2) retailing to the public. Of course publishers are free to attempt to interface with the public, but the complexities of non-standard hardware may require the specialty of a vendor to make the interface profitable.

All the old habits and traditions of the library may continue, and the size of the drop in their clientele will depend upon the number of users willing to pay a fee for the greater convenience of dealing with a vendor. To participate in the electronic information age, the public library will probably offer some sort of free access to vendors, using, as it does now, public money.

I see an expanded role for the public library, not necessarily serving more users, but providing more and more complete services for the users it does serve. [Rochell, 1982, 1815].

Note that between Figures 1 (see page &next.) and 2 the user community and its needs have not changed. That community, in a free economy, will see to it that all the services it received from the print technology are also provided by the electronic technology, and probably more.

2. Will changing information routes change society?

But what about the people who are too poor to pay for information access but who may need it the most?

In all this talk of paying for information there lurks the specter of a "Counter Enlightenment." As the Age of the Enlightenment, heritage of our modern society, sprang from a new freedom of information (e.g. public schools and libraries), so a new feudalism may spring from discriminatory access caused by pay-per-use information services. Those who have money to buy computer services will receive better information more efficiently than ever before. Those who cannot afford to pay will have less access than they have now. Public libraries, free to everyone, may shrivel as their more affluent clients retire to the privacies of access via NEXIS and CompuServe terminals at work and at home.

Knowing how to use computers enables the information "haves" and their children to secure the best jobs and to participate in the coming post-industrial, information

society. Being denied computer access, and skills, the "have nots" are relegated to a permanent class of know-nots [Rochell, 1982, 1815]. An Abe Lincoln could never become President because he could not afford access to the information he needs to get him started. (Instead of readin', writin', cypherin', survey, and law books try CAI, PROSPECTOR, and LEXIS.)

Though such a view certainly seems plausible in our day of concern for equal rights, perhaps more analysis is needed before condemning a pay-per-use information policy.

The problem of cost

The "free" availability at the public library is really an illusion.

Information is not free and the real question is who should pay to make information available to whom. Libraries' budgets have traditionally made information available to individual patrons at no charge, partly because it would be impractical, if not impossible, to determine direct costs based on the patron's need for information. [Markee, 1981, pg 440].

Also, it is only available as long as concurrent clients want different documents. At a point of conflict (two people want the same book on the same day), availability is on a first come first serve basis, which is a more limiting restriction than pay-per-use. In time-shared FT, lots of people get to use the same document concurrently.

It was not libraries which opened the Middle Ages, but the mass production of information-containing images by printing which made the difference. Access is expanding again when FT removes the requirement that somebody possess something (a book) before he can access it. An instance of freer access is geographic leveling. A student will no longer need to live near a great research library to access its materials [Lancaster, 1982, pg 184].

Even after the Enlightenment, when Abe Lincoln walked miles to borrow a book, it was not from a public library but from a friend willing to lend a personal copy to him. Their exchange was exactly in the pattern of information access practiced by the 14th Century Humanists in their feudalistic society. Petrarch, Boccacio, and Salutati, exhumers of the Classics and forerunners of the Renaissance, would lend their manuscripts only to personal friends [Deuel, 1965, pg 3-51]. Mechanical book production released information from the cloister of the invisible college to a wider public.

One interpretation might say that printing brought books, previously free for the copying, under the control of publishers and copyright interests. But in fact, printing made books, previously very expensive because of the great labor of manual copying, much less expensive and therefore much more widely available (though not for free).

Are there any real examples of a commodity which became cheaper and more widely available also becoming less available to some?

The problem of quantity

Abe Lincoln was able in his poverty to get books because he had friends, somewhat less poor, who could afford the cheaper books. If he had lived in Petrarch's Florence, he would have needed rich friends. If he were to live tomorrow, he would need a friend with an online terminal. In all cases his access is subsidized.

The difference is that in the modern world there is many orders of magnitude more information available than in either Petrarch's or Lincoln's worlds. Besides the problem of storing it, the problem of keeping up with it requires more and more specialization by researchers. For present publishers

. . . marketing today involves informing large numbers about the availability of a book or journal, but selling to only a small peripheral group after one has sold to the core group--often less than one percent of those informed. As the market place becomes increasingly segmented, reflecting the development of science and scholarship itself, {new IR} technology may enable publishers to communicate effectively and at acceptable costs with the segmented marketplace
[Horowitz, 1982, pg 91]

and no longer trying to communicate with the peripheral 99%. This analysis describes current publishing trends with

its high printing and marketing costs, not FT. Where the Enlightenment freed information from its Medieval cloister, these trends may be creating

. . . a kind of new feudalism." And
 " . . . marketing people are becoming involved in publishing decisions at earlier phases.
 . . . increasingly it encompasses interpreting . . . the needs of scholars In most businesses, marketing people play a major role in product development, in figuring out how to shape a product that people will want to buy. To some extent, marketing personnel in the business of scholarly information are also going to begin participating in shaping products that their customers will buy . . . [Horowitz, 1982, pg 92].

The real problem here is that marketing professionals, not scholars or subject specialists begin to determine what will be published. The same trend has been noted in other communication processes, for instance, TV news:

When local stations create and choreograph entire news programs along guidelines supplied by researchers--toward the end of gratifying the audience's surface whims, not supplying its deeper information needs--an insidious and corrosive hoax is being perpetrated on American viewers. . . . Lost in this concept is that a professional journalist should have the ability and news judgement to determine what is important and significant Instead of striving to impart information to the viewers, the salesmen-managers of television stations were engaged in a tacit conspiracy to extract information from the viewers--information that would serve the managers in their efforts to maximize audience size [Powers, 1977, pp 234, 88, 3].

Or in the academic balance between faculty and libraries at universities. Former English professor and college

president, Ray Heffner, responded as follows in a panel discussion on how libraries should weed out books:

I have been a little upset here to hear so many times that the function of the library is to support the faculty and the academic program of the institution. Well, I hope that in many cases libraries are subverting the faculty and the academic program because it could very well be possible that the faculty at any moment in time is terribly limited in its vision. They don't really understand very deeply the subject that they are professing to teach, and only a well-run, well-organized library collection will correct all the misinformation that they are feeding the students in the classroom.

So if you tie a library too closely to your academic program, you may be putting the students into tunnels, giving them all kinds of cattle chutes which will lead them to slaughter but not to the kind of comprehensive education you hoped for. [Gore, 1976, pg 145].

The constraint in each case--library shelf space, minutes in a TV newscast, the limited advertising channel between publisher and reader--decrees that some information must be discarded.

The solution, according to FT, is not better methods of choosing what to discard, but is keeping it all.

. . . we could acquire the habit of recording nearly everything, because recording devices will be connected to memories so cheap that it will be easier to record any given transaction or observation than to decide whether or how to record it. Today, a considerable amount of human intellectual power is devoted to the task of deciding when to record something, or what to record: this after all is what a cataloger or abstractor does, as well as a physician writing a medical history, or a news reporter making notes for a story. I foresee at least the possibility of a world of holistic recording, in which the intellectual effort of

deciding what is worth seeing comes after the recording. [Meadow, 1979, pg 218].

Though acknowledging his contrast to Lewis Branscomb's [1979] vision of regathering data rather than accessing an archive, Meadow describes a future chemist who records all parameters of an experiment automatically and reports, not constituents, environments, etc., but the address of the automatically made record. A "world of holistic recording" continues Bush's [1945, pg 32] vision of the scientist who wears a walnut sized camera on his forehead and takes pictures of all his experiments and observations throughout his day. Suppose medical records could be kept "holisticly" and still be retrieved quickly. Memex was a means of huge storage and attempts to access it.

As in post-coordinate searching, the user must now assume more responsibility in retrieval, but he also gains the power of "dynamic interpretation at the output phase" over the whole collection, to use Luhn's phrase (see page 75).

There will be a struggle to define who is allowed to access what kind of information. The nineteenth century was largely defined by the struggle for control of the means of production; the twentieth century is being defined as a struggle over the means of communication and information. . . . who controls the hardware, software, and marketing--of information about the information--may ultimately control the information itself. We may begin to talk about democracy in terms of how widely a society permits access to information, to computers, to photocopying equipment. . . . Authoritarian

regimes restrict access to information, while at the same time they are plagued by the need to use advanced information. . . . In the past, information was the monopoly of the few. . . . as more persons become involved in the publishing process, there is an opportunity to maximize information as a liberating, democratizing activity. . . . democratic societies will be advantaged by the new technology and by its openness to all. [Horowitz, 1982, pg 95].

III. WAYS OF ANALYZING TEXT

Sections I and II have explained how FT can expand the services of present IR systems. FT was defined (see page 22) as processing incoming documents and producing indexes which describe their contents. Now it is time to defend that building and searching of indexes even for systems which could search the FT just as quickly.

The quality of output from traditional IR systems is restricted to the quality of indexing or classifying applied to the documents entering the system. No matter how much a target document may be "about" a certain concept, if that concept is not somehow represented in the indexing, then that document will never be properly retrieved. Also, the nature of their searches is restricted by the nature of their indexes. Should FT, by relying on document descriptions, be under these restrictions?

Searching without document descriptions

Now it has been suggested, mainly by more euphoric, computationally oriented workers in information retrieval, that retrieval systems should be based on full document texts, functioning as their own descriptions. Sooner or later, they say, the hardware will be available that will make the storage of, and rapid access to, such vast quantities of matter feasible. A scheme of this kind supposedly has the advantage that it avoids the omissions and distortions inevitably associated with index descriptions. But it ignores the positive role of the summary index description It

must, however, be emphasized that though this role of an index description as a summary is widely accepted, sophisticated analyses of the relation between such a summary and its source text, or indeed between summaries and sources in general, are lacking. [Sparck Jones and Kay, 1973, pg 47].

The euphoric suggestion has actually missed the whole principle of IR. Confusion easily arises because the document on the one hand and its description on the other are both words. And in FT both sets of words reside in machine readable form; but it makes about as much sense to say that a document can function as its own description as to say that a photographic subject can function as his own photograph. In a very few circumstances, the subject and photograph may serve equally well, but in most applications they are not interchangeable.

In other words, a system which searches full text without recourse to indexes derived from that text does not escape the indexing paradigm; it only postpones document processing. Consider the following contrived example.

The euphoric approach for finding the documents in a collection which contain the phrase, "Little Red Riding Hood" is to run a string matching search across the whole database. But it makes absolutely no difference in the concept or effectiveness of the search to (1) consult a prepared index of document descriptions for "Red" and "Hood"--"Little" and "Ride" having been considered too common to be indexed; (2) retrieve the FT of those

documents containing both terms; and (3) run the string search only on those. For this query the difference between the self-descriptor system and the index-descriptor approach is entirely internal to the machine. The outputs will be identical, though the second depends upon prior document processing while the first does not.

Not only are the strategies of the two searches identical, but the first approach has utilized a document description just as much as has the second. While claiming to operate without restrictive document descriptions, the euphoric approach has actually hidden the document description in the query, as follows: (1) The user has stated that he is interested in five items--four words and the syntactic content of this specific ordering. (2) Though a uniterm system could catch the four words, the syntactic content traditionally would require human analysis and assignment of a pre-coordinate subject heading (see page 31). (3) The document description in this example is a simple, binary flag attached to each document in the collection, i.e. does it contain the pre-coordinated phrase "Little Red Riding Hood" or not? Therefore, the euphoric search has processed documents--but at search time. Swanson described his FT experiments as "indexing at the time of search" [Stevens, 1970, pg 134].

This similarity between self-descriptor and index-descriptor systems holds regardless of how complex the query and resultant "hidden" document description may be. What the euphoric workers really desire is the ability to customize their view of the collection's classification. Whether an FT system prepares searchable structures ahead of query time or not is immaterial, as long as the system performs at least as well as a self-descriptor. The distinction of FT is not when documents are processed--before or at query time--but the fact that document processing may be altered without affecting input to the system (FT) or the user's expectation of output from the system (like a self-descriptor's).

In fact, the index-descriptor system may offer improvement over the other because it has put more work into the problem to develop some resources for further processing (e.g. term substitution suggestions). This analogy places a lower bound on the performance of an FT system, namely, it must never appear to do less than the self-descriptor system would. Any index upon which an FT system depends must represent every concept extractable from a search of the full text itself. A little of the perversity of processing natural language by these means appears in this contrived example. If the query had been

"The Little Ride", that index would have failed on a perfectly reasonable request.

Document descriptions as summaries

In the context of the Sparck Jones and Kay passage above, the "role of the summary description" is positive because it saves storage space. Summary means shorter. Salton also assumes that a major purpose of document analysis is to shorten what is stored [Salton, 1983, pg 53]. With its presumption of cheap storage, FT relaxes this heretofore universal restriction in IR systems. Even if a document's description were as bulky as the document itself, that would only double the amount of storage needed. If nonsummary document descriptions significantly improve the performance of systems meant to contain the exponential increase in publications, then the $O(2n)$ increase in storage is worth it.

Sparck Jones and Kay continue by identifying three valid relationships to be investigated:

- (A) between the unit of a document description and the portion of the document from which it comes (document processing);
- (B) between the description of one document and the description of another (the larger-context problem, see page 22); and
- (C) between the description of a document and the description of a query (query processing or searching).

Throughout the rest of this Section a letter in parentheses will indicate where one of these relationships is being used to analyze text processing. The first (A) and third (C) are obviously treated in bibliographic systems.

Investigating the second (B) seems to require Question-Answering techniques, but FT systems using automatic relevancy feedback may make some contribution here.

Another challenge to the goodness of index building comes from research suggesting that indexing on the full text of a document gives no better retrieval performance than indexing on title and abstract only [Salton, 1983, pp 71, 111], but these studies may be premature until methods of IR evaluation are above challenge [Doyle, 1975, pp 353-358; Sparck Jones and Kay, 1973, pg 50].

Swanson, however, is certain that the lack of FT is detrimental, even to the point of losing much of the material in a collection.

It is another obvious but seldom remarked fact that the extent of such information loss from existing libraries is not only unknown but has never {been} defined in measurable terms.
[Stevens, 1970, pg 146]

Therefore continuing with the presumption that FT will process documents to create indexes, how may text best be processed? The only known agents capable of this task are

humans, the traditional document processors, and computers, the experimental ones.

A. Manually

Sparck Jones and Kay [1973, pp 10-12] advise against attacking IR problems by computer before considering manual approaches. Humans have produced for public consumption only two kinds of document descriptions: (1) abstracts and (2) classifications or indexes.

The abstract

To show the relationship between the first kind of document description and its document (A), an abstract

. . . is the terse presentation in (as far as possible) the author's own language, of all the points made, in the same order as in the original [Collison, 1971, pg 3].

Except for the qualification, "(as far as possible)," this instruction to a human abstractor could just as well begin one's instructions to a machine which derives the terms of its document description from the vocabulary of the document itself. But his prerogative to become an editor is closely controlled.

It is important that the abstractor should use the vocabulary of the author as far as possible; paraphrase is dangerous and can lead the reader to channels of thought not intended by the author. Here of course the abstractor must use common sense, for not every author is a master of language or communication . . . so

the abstractor will need to add definitions of those terms where he deems it necessary.
[Collison, 1971, pg 11]

Most usually, then, the abstractor is not a translator but a selector.

An abstract is intended to convey all the original contribution to knowledge made by the original contributor--and nothing else.
[Collison, 1971, pg 11]

Though an article should commence with a summary of existing, pertinent knowledge, and make reference to the previous work of others, this material has no place in the abstract. As a subject specialist himself, the abstractor identifies what the user already knows and leaves it unsaid. The relation of one document description to another document description (B) is one of non-duplication.

Therefore, humanly produced abstracts suffer more from the larger-context problem (see page 22) than original documents. Consequently, IR experiments on specialized collections of abstracts, such as the Cranfield aerodynamic library [Salton, 1983, pg 101] may not discover the seriousness of the larger-context problem because the processing of abstractors has removed from the system the terms which would introduce it, and the narrowness of the collection's subject discourages users from trying broad queries. An FT system, on the other hand, aims to transparently transmit all original clues to the user (see

Premise One) and let him deal with the larger-context problem.

Finally, the abstract

. . . is much more than this, too, for it is intended to stand as a readable and complete item in its own right, and is therefore a separate work of scholarship that can be fully indexed and exploited for the benefit of scholars and research workers. [Collison, 1971, pg 3]

In other words, an abstract should function as a legitimate substitute or proxie for the document. A user may gain much of the information of the original by consulting only the abstract. This kind of document description should be a substantial, independent representation of the content of the document to which it relates (A).

Indexing and classification

For the second form of output from human document processors:

We will not make any distinction between subject classification and subject indexing because, for all practical purposes, the processes are identical. [Lancaster, 1976, pg 22]

Back-of-the-book indexing is the restricted but still typical case of subject indexing in which a single effort simultaneously indexes the total collection (the book). Whenever larger works are indexed, e.g. an encyclopedia, portions of the "book" begin acting like individual

documents in a library and the staff of indexers begin having typical classification problems, such as scope definitions and consistency [Posner, 1972; Shaftesley, 1964]. Lancaster continues:

Despite a considerable amount of wooly thinking in some quarters, it clearly makes no difference whether we represent documents on "feathered, egg-laying creatures with wings" by the word BIRDS or by some notation, say 598.2. In either case we are engaged in forming classes of documents dealing with similar subject matter, i.e., with classification. [Lancaster, 1976, pg 22]

Two classification formations very familiar to Americans are Dewey and the Library of Congress. Inventors of such schemes begin by dividing and subdividing the whole universe of knowledge, in a tedious exercise in "top down design," until individual nodes are formed. A classifier decides on a document's major subject, locates that subject in the invented tree of knowledge, and assigns the description of that node, usually represented by a number, as that document's description (A). A user who can also locate his query in one of the nodes in that tree immediately finds some relevant documents--or learns that there are none under that heading in the collection (C). (Incidentally, assuring a user that the system has no references relevant to his query is probably the most difficult task in IR.) A scheme external to both document and query is imposed on both to get them together. In order

to prepare documents so that they may be related to queries, the system also processes documents so that they are related to each other, either laterally or hierarchically, in the tree (B). This approach follows the comfortable Platonic path of conforming reality to a priori ideals.

But reality has a way of sometimes not conforming. Such traditional systems suffer three interrelated difficulties: (1) Not exhaustive; a document enters the system with only its major subject and a few extra subject cards as access points. Classifiers, abstractors, and even authors concentrate on the main subject of their documents. But for every reader who is interested in the subject as a whole, there are many more only interested in perhaps fringe aspects applicable to their own interests [Collison, 1971, pg 12]. (2) Scattering; a document relates different ideas to each other, but classification arbitrarily disassociates some of them to make the rest fit the system (A) [Doyle, 1961, pg 555]. Consequently, related documents go to different locations in the library (B) as classifiers attempt to fit new subjects into old categories. (3) Closed categories; when a document's main subject misses all the nodes in the tree (A) [Doyle, 1975, pg 255]. Classifiers may invent new nodes by adding further subdivisions to established categories, but the proliferation of

"inter-disciplinary" projects suggests at least that the invented supra-categories have missed the real and useful divisions of knowledge or, more likely, that the universe of knowledge is not really susceptible to a priori division without analysis of the documents to be classified.

Uniterm "classification"

A more modern approach looked at documents as natural phenomenon--perhaps a natural perspective occurring to those much penalized by the failures of traditional cataloguing--scientists. If the property of interest, content, could be measured as documents enter the system (A), then this easily handled list of content indicators could be consulted later to retrieve relevant documents (C) and provide the authority list of terms recognized by the system. The Data Processing techniques of counting, sorting, and listing were at hand to manipulate the data. In IR terms, the classifier has a chance to derive a more objective document description; "What subjects does this document discuss?" not, "Where does this document fit in somebody else's tree of knowledge?" When postings to a single keyword grow large, the DP machines (or Peek-a-boo cards up to a certain collection size) easily perform Boolean set intersections. This post-coordination of uniterms (see page 28) reconstructs something like the

traditional subject headings without the disadvantages of using the whole classification scheme.

It is now known that concentrating on relationships (A) and (c)--document to document description and document description to query--does not make relationship (B)--document description to document description--go away; it just hides. In other words, the collection is still being classified, but uncontrollably. For example, to search on

cigarette AND production

is to look under all the documents about cigarettes for the subcategory dealing with production, and also to look at all the documents about things in production for the subcategory of one of those things, cigarettes. By post-coordinating these two terms, the user searches as if the collection had been classified using one of these pre-coordinate headings.

A coordinate system is a 'do it yourself' system of forming subcategories, and though subordination is disclaimed by advocates of such systems, it is nevertheless inherent [Doyle, 1961, pg 560].

Instead of finding a subcategory under only one category, as in a traditional scheme, a user may search for a post-coordinated subcategory under a variety of categories, a boost of Luhn's desire (see page 75) for dynamic interpretation at output. However,

. . . when the librarians hand control of subcategory formation over to the customer, they relinquish two other kinds of control--control over the names of the subcategories and control over the quantities of material in them. Because of this the librarian assigning uniterms to a document can have no clear idea of how these assignments are going to affect its relationship to other topically close documents. [Doyle, 1961, pg 560]

Important properties of the collection are being obscured. Viewing a library as a whole or viewing the organization of some of its parts (as one could the traditional scheme, if not the collection) becomes very difficult. To diagram such relationships, where one subcategory may have multiple supra-categories, would require that maze of pointers which is familiar to those who have tried to draw networked database schemes. Neither can books be linearly shelved under such a network (B). But note that, even though the scheme is unsuitable for paper representation or filing, it is very easily navigated with simple menu routines on an interactive CRT. But does the lack of subcategory control cause any harm?

The poor performance of the extensive uniterm experiment at the Los Alamos Scientific Laboratory in the 1950's has been widely quoted [Doyle, 1975, pg 176]. Their disappointment may have been due to failing to understand the relationship between the number of documents retrieved and the number which will be irrelevant (recall versus

precision). In any case, they and others have attempted to improve the performance (Precision) of post-coordinate uniterm systems (1) by using multiple word terms ("cooling water"), (2) by adding role indicators ("passive-cooling water" for water being cooled, "active-cooling water" when water is the coolant), and (3) by treating a multi-subject document as separate documents [Cushing, 1963].

Improvements (1) and (2) attempt to attach syntactic meanings to this semantic based (really morphologically based) system while (3) attempts to avoid false associations by establishing proximity criterion [Lancaster, 1976, pg 33]. These modifications (A) improve performance, but perhaps not enough to justify the significant document processing complexity added to a system which was originally attractive because of its simplicity.

Perhaps of help will be some sort of overt classification scheme which is as scientifically objective as uniterm processing but is not so ambitious as the traditional.

Faceted classification

India's premier librarian, S. R. Ranganathan (Rang'-Na-Than) pioneered an inductive approach with his

Colon Classification, and Vickery developed the idea further in Britain. A faceted classification scheme begins inductively by looking at the documents in hand and builds, from the "bottom up," a classification scheme broad enough to encompass that collection without including the whole universe of knowledge. Each library or field of study writes the terms necessary to describe its own collections. Classifiers organize these terms into "facets" or dimensions along which the primary subject of the collection is modified by a subset of the terms.

For instance, soils may be Arctic, Temperate, or Humid or they may be Sand, Loam, or Clay. The first three describe soils by the climate in which they occur, the second by their textures. "By climate" and "by texture" become two facets in a Soils' schedule. When put in any convenient order, assigning a call number (A) amounts to writing a notation for each topic covered in a document in that order. New terms may be added under existing facets, and new facets may be added. Because of the prearranged ordering of facets and terms, documents on the same subject come together on the shelves, general treatments before detailed treatments (B) [Vickery, 1960]. Searching (C) still functions in the traditional way, but faceted classification allows local manipulation of terms and

schedules, making his controlled vocabulary more responsive to the classifier's needs.

Problems with a controlled vocabulary

The three preceding systems differ in where the organization of concepts (B) occurs: before any documents are touched (traditional), as preliminary to document processing (faceted), or at query time (uniterm). But all have the problem of consistency when more than one librarian classifies for the same collection, and even when the same librarian classifies on different days. To meet this problem classifiers compile and consult authority lists of accepted words, and their meanings, to control the vocabulary of their document descriptions.

It is important that concentration on inter-indexer consistency, which is easily studied, not obscure the more important issue of indexer-user consistency. The use of controlled vocabularies has been demonstrated to improve the former more than the latter [Doyle, 1975, pp 359-60].

Lancaster [1973, pg 22] reveals the opportunities for error by distinguishing three steps in human document processing:

- (1) The content analysis; when the classifier reads the document, he may or may not understand it fully.
- (2) The conceptual analysis; he relates his "content analysis" the potential interests of his users, which is the uncertain business of predicting

the future. And

- (3) translation; he expresses his "conceptual analysis" in the vocabulary of the retrieval system. Whatever his accuracy up to this point, he may find that the controlled vocabulary does not contain the terms necessary to describe the document as he believes it should be described.

A more effective squelching of Luhn's dynamic interpretation of the collection at retrieval time can hardly be imagined.

On the one hand comes the user (C) groping to associate concepts not previously drawn together. He has an inter-disciplinary query trying to synthesize new developments scattered in the literature and he needs all the variety of terms possible. The free vocabulary of the documents themselves may barely be enough. On the other hand stands the controlled vocabulary (A), which by the nature of committees and standardization, will evolve more slowly than the free vocabulary of the documents.

The indexer, assuming he recognizes a ground-breaking document in hand, must squeeze new concepts into an old, inflexible vocabulary. New terms appear more and more often in the literature and old meanings change. Unlike the dried wine-skins of ancient Palestine, the controlled vocabulary will not burst. Instead, perfectly modern documents fall into the system with antiquated document descriptions (A), never to be heard from again. Controlled vocabularies served well in a day of slower pace, but the

expansion of research writing also means an acceleration in language evolution.

PRECIS

A newer system of indexing goes a step further by not controlling the input vocabulary at all. The "concept analysis" itself is written in something like a natural language phrase, often using terms from the document itself (A). The phrase is reversed, tagged, and fed into the computer. For example:

- (0) Britain.
- (1) Aircraft industries.
- (p) Skilled personnel.
- (2) Training.

The PRECIS computer programs developed by Derek Austin for the British National Bibliography produce a rotated type index giving each term as an access point without losing its context:

Britain.

Aircraft industries. Skilled personnel. Training.

Aircraft industries. Britain.

Skilled personnel. Training.

Skilled personnel. Aircraft industries. Britain.

Training.

Training. Skilled personnel. Aircraft industries. Britain.

Entries from all the documents are sorted and printed as a classified bibliography (B). Any duplications of the top

line of a couplet are suppressed leaving the second lines as subheadings.

PRECIS depends upon term rotation concepts (of which the KWIC index is an example), upon principles of articulated indexing [Townley, 1969, pg 350], and upon Ranganathan's chain indexing, which, unlike PRECIS, creates generic headings by dropping the most specific term at each rotation (placing a pointer back to the full entry) [Austin, 1976, pg 51].

Consider again the basic document description input to the system:

Britain. Aircraft industries. Skilled personnel.
Training.

Though resembling a string of uniterms, this description carries syntactical content as well because the human indexer has pre-coordinated these terms in an order which makes sense to him.

Several analyses reveal why this type string is intelligible and unambiguous. The terms are in context-dependent order meaning that each term sets the next in its context, e.g. "Britain" tells where the "Aircraft industries" are [Austin, 1976a, pg 6]. Or, each stands in an inclusive relationship with the preceding term, e.g. Training is one of the many activities of skilled personnel, etc. But the easiest way of understanding how the syntactic content of their ordering

is preserved for the human user is to look at the terms written in reverse order:

Training. Skilled personnel. Aircraft industries.
Britain.

The mind, conditioned by years of reading, immediately fills in the operators:

the Training of Skilled personnel in the Aircraft
industries in Britain.

But another order, as might be produced by a less sophisticated rotation scheme doesn't preserve the syntax:

Skilled personnel. Britain. Aircraft industries.
Training.

The mind again attempts to fill in operators, but with ambiguity. Are Skilled personnel in Britain being retrained for the Aircraft industries?

Through deeper analysis, PRECIS handles much more complex document descriptions than exemplified here. But note that the system has demonstrated how a pre-coordinate document description, once generated, may be manipulated while preserving the syntactic meaning of its context. A great deal of work in Britain has contributed to understanding these principles [Wellisch, 1977]. This work also illuminates what was happening all along in traditional classification methods. A listing of the route through the tree of knowledge to the node where a document had been classified would yield a string of terms with the same syntactic properties as the example above. The

traditional ordering of terms was good, but the closed categories were bad. Experience with the British National Bibliography has shown that 85% of the PRECIS documents descriptions produced cannot be matched to Dewey notation [Austin, 1976a, pg 67].

The thesaurus

To control vocabulary, first control synonyms. Choose one as the favored term and enter all others as cross references to it:

Tennessee SEE Volunteers

What begins as an authority list to help indexers and searchers use the same terms to mean the same things becomes a more or less complete network of related meanings. Precise definition creates a semantic hierarchy in which specifics are subsummed under inclusives and other relationships develop between the favored terms. Three types of relationships are commonly identified (see page 58):

"BT" before a term that is broader, meaning higher in the tree than the term being defined;
"NT" before a narrower term, lower in the tree; and
"RT" before a term at the same level but related.

For example:

Smokey - BT mountains; NT Gatlinburg, Cades Cove;
RT Blue Ridge

which defines a tree where "mountains" is the parent; "Smokey," "Blue Ridge," and other ranges the sons; and localities are the grandsons. These relations evoke SEE ALSO cross references in an index.

However, natural language doesn't work this precisely. Traditional thesaurus-controlled systems require that the indexer translate from the natural language of his "conceptual analysis" (see page 119) into the indexing language of the system (using SEE references). PRECIS, on the other hand, maintains the thesaurus as one of the steps in classification. The indexer must fit into the semantic network of the PRECIS thesaurus database any new terms he has used, plus any BTs and RTs he can discover which are not already there [Austin, 1976, pp 35-36]. Then the thesaurus generation program supplies SEE and SEE ALSO references for the user who doesn't think of the right term the first time and for the inter-disciplinary searcher needing new access points for his uncertain query. In other words, a PRECIS thesaurus is something like a simple Q-A database which the indexers assemble manually term by term.

By letting the process of indexing generate the thesaurus, it will contain all the terms the indexers want to use, probably those the authors are using too, and eventually should contain most of the words the users would

be thinking of. The size of such an inductive, collection-oriented thesaurus will grow rapidly at first, but stabilize after 40,000 to 50,000 documents have been indexed [Doyle, 1975, pg 176]. The PRECIS thesaurus helps the user who is not naturally using the same vocabulary as the indexers. A researcher who consults the thesaurus before submitting his query uses what Lancaster would call a post-controlled vocabulary. The PRECIS user has the uniterm's advantage of a record of the collection's actual (not theoretical) access points and has the traditional ability of vertical movement in a hierarchy to generalize or specificize his query [Lancaster, 1975, pg 40].

Still the euphoric researcher who desires a customized classification of the collection (see page 105) will not be satisfied. All of these manual document processing systems, whether computer assisted or not, output a set of access points--index entries, shelf locations--which the user is expected to find applicable to his query. He tries his different query terms, one at a time, browsing in that area of index or shelves, hoping that a "horse like his" query (see page ii) will jump out.

Inescapably, human document processing relates document to document description (A) subjectively. It is as if the system says to the user, "Find your document in my categories, or don't find it (C)." Hopefully, through the

power of FT, a future system can appear to reclassify its whole collection from the subjective viewpoint of each user [Stevens, 1970, pg 231].

B. Mechanically

Method becomes a function of volume. If the volume of publishing were not so great, the method of computer analysis would probably not be pursued.

The history of computerized language processing may be looked upon as a monumnet to getting the cart before the horse. Normally, a new line of research and development will begin its attack on two levels: (1) the mobilization of what is already known for the sake of solving practical problems, and (2) the pursuit of basic lines of research, serving the longer-range objectives of enabling the profession to solve more difficult practical problems. In fact, language processing did not begin with either path, but, in dead earnest proceeded at once toward attaining practical solution of what has turned out to be one of the most difficult of language-processing tasks: machine translation of languages. [Doyle, 1975, pg 316].

KWIC

Contributing to the early enthusiasm may have been the easy success of one type of mechanical document processing, the KWIC index (see page 27). It was invented by Luhn and others, and introduced to the indexing of scientific periodicals in the late 1950's.

Of the two major problems in indexing, synonymy and homography, KWIC leaves it to the reader to solve the first. References on the same concept are scattered under different headings because authors use differently spelled words with the same meaning (synonyms). Users are expected to be familiar with their field, to learn which alternate access points to consult, and to collect the scattered references.

KWIC gives the reader more help in solving the second, at least in short indexes. Enough context is retained on the 80 to 120 character line entry so that the meaning of different words with the same spelling (homographs) may quickly be distinguished. But, for a large collection, single headings, such as light," may run for pages and the user must scan all, deciding for each line which homograph appears.

. . . no attempt is made to separate words which are orthographically identical, but distinct in grammatical function or in meaning To overcome this is a problem of an entirely different magnitude from that of producing the concordance or word index in the first place. [Bott, 1970, pg 222].

This subsection will return to that problem of different magnitude" after delving into some of the other problems KWIC avoids by its simplicity.

Title lists keyed by words have value for quick distribution and fast use since time is often a very important element in the obtaining of

information. {But} such lists do not serve adequately for thorough searching A title concordance may be more useful than {it} would seem from the . . . data on index entries. However, it must obviously be incomplete, must have many unnecessary entries, and would not prove suggestive enough to users who lack background in the subjects sought. [Bernier, 1962].

In other words, KWIC indexes don't have extra helps for users who don't know what they want.

The KWIC system uses a document's title as the document's description. The quality of that document description (A) depends upon the author and his subject. Many technical fields encourage the writing of titles which are specific and informative. But other titles may be whimsical or enigmatic [Doyle, 1975, pg 295-6].

About half the time a document's title will contain the words also assigned by indexers [Sparck Jones and Kay, 1973, pg 133] which makes it equivalent to a manual document description half the time. Scientific titling may have as high as 90% overlap--or as low as 52% in a new field where terminology is still changing. The titles of the general material covered by the Reader's Guide to Periodical Literature contain the words of the subject headings under which the articles are manually classified about 42% of the time [Stevens, 1970, pg 59].

KWIC gives some help in identifying related terms by the way compound terms group together. Just as all the titles containing the word "Equations" fall together, all

the subset of those containing the term "Linear Equations" also fall together. An alert user may consult "Linear" as another access point in his search. [Doyle, 1975, pg 296]. As cheap and quick, requiring a minimum of keypunching (titles only), a KWIC index is unsurpassed.

But in keeping with what must be a constant perversity of natural language computing, gaining any improvement seems to require solving problems of increased magnitude. As a simple example, which list of headings would be more helpful to a user seeking titles on heat treating metals?

temper
temperate
temperature
tempered
tempering

or

temper(ed,ing)
temperate
temperature.

To achieve the latter the computer must remove endings such as "ed" and "ing" and equate the remaining word stems. But this stemming algorithm must be sophisticated enough to remove the "ate" in "differentiate" without removing the same from "temperate" (as in climate). It should remove the "er" from "costlier" but not from "temper." All this can be done [Salton, 1983, pp 72-73] but the system is no longer as simple.

Two kinds of indexing

KWIC is the most rudimentary of mechanical document (really document description) processors. In fact, PRECIS makes more use of the computer than KWIC. The difference is that KWIC processing involves no intellectual indexing effort. The intellect must be applied at the output phase (Luhn's phrase, see page 75). The difference is best expressed in the distinction between a description that is derived from the text of an item itself (derivative indexing) and one that is assigned to it from a separately existing thesaurus or schedule of subject categories (assignment indexing) [Stevens, 1970, pg 13]. The manual methods already discussed use assignment indexes. The title and author cards in a card catalogue, the KWIC entry, or an abstract (most of the time, see page 109) are derivative document descriptions.

For a computer, derivative indexing seems most appropriate. To do more will require a Q-A type system to hold and apply outside material. But Stevens, after surveying derivative indexing, spends an entire chapter on assignment indexing by computer, quoting Bar-Hillel to the effect that any less will never be commensurate with human indexing [Stevens, 1970, pg 90]. On the other hand Lancaster decides on the basis of results:

. . . assignment indexing by computer seems a singularly inappropriate activity. That is if we are going to use automatic processing rather than human processing we should not be concerned in trying to faithfully reproduce the activities now performed by humans. We should only be concerned with results attainable by automatic processing that are as good as or better than or cheaper than the results achieved by humans using alternative approaches. [Lancaster, 1976, pg 31].

But improvement in document processing by computer involves use of material outside the document. Even KWIC uses an independent stop list to prevent common words from becoming lead terms in index entries.

When processing FT, it is not so much a question of whether a document's description contains document-based terms (derivative indexing) or system-based terms (assignment indexing), but rather the question: What data shall the system extract from documents in order to produce document descriptions? [Sparck Jones and Kay, 1973, pg 131].

Obviously all mechanical document processing must begin derivatively with word forms extracted from the text. Equally obviously, users are interested in concepts. This leads to the "different magnitude" problem of natural language. The correspondence between words and the ideas they represent is not one to one. The same words in different discussions or in different orders represent different ideas. Linguistic document processing attempts to

extract the meaning of the ordering, while statistical document processing uses the indirect clue of word frequency to differentiate discussions. The next two subsections treat these two methods of mechanical document processing.

Parenthetically

Citation indexing uses another indirect clue to document content. The citing author is pointing to another article like his. Weak links in this content indicator are authors prone to self citation [Stevens, 1970, pg 32] and those who cite works at hand rather than better ones that would have to be looked up [Doyle, 1975, pg 379].

Citation indexing is subject to the same problem as keyword indexing, namely that indirect clues are relatively opaque: a specific citation may not in fact have much to do with the prime topic of a document, any more than an individual word. The whole hypothesis on which this kind of approach to document characterization depends, however, is that there is a correlation, and that in general enough information about documents is obtained in this way to characterize them for retrieval purposes. [Sparck Jones and Kay, 1973, pg 135].

Though citation indexes are certainly useful and ought to be part of an FT system, they will not be further considered here because they are tangential to the main problem of getting meaning from words. However, note that

citations become more useful to a system the more deeply it indexes its documents.

C. Statistically

With FT's allowance of lengthy document descriptions (see page 106), indexing every significant" word in the text is a simple, direct way of providing all possible access points for later searchers. But the hundreds of tags assigned to each document by such maximum-depth indexing may cause the system to find at least one of the three or four terms of a normal query in nearly every document in the collection. Normal indexing assigns only a half-dozen or so terms to each document [Doyle, 1975, pg 336; Salton, 1983, pg 86-7]. When FT places many terms from each document in its index, some criterion must be used to limit or rank the documents retrieved by a normal query. Theoretically the words most indicative of a document's important contents are those which occur in the document most frequently. But the theory to explain why this is so has not arisen.

A beginning

Luhn's method for automatically extracting abstracts from FT laid a foundation for statistical processing of

text: (1) isolate the "significant" vocabulary of the document; (2) calculate the "significance factor" of each sentence as a function of the number of "significant words" it contains and how close those words are to each other; and (3) print as an abstract those sentences with the highest significance factors in their order of occurrence in the document.

The most obvious flaw in the resulting "abstracts" is the lack of continuity among sentences. For instance, pronouns are likely to have the wrong antecedents. But Luhn demonstrated that this simple orthographic counting did estimate meaning. A machine could be directed, by simply counting word forms, to select sentences which a human would also judge to be more indicative of a document's total content than those the computer rejected. And there was no fear of the automatic abstractor changing the words of the author. The result is a machine generated, derivative document description meant to represent an article to a reader.

Luhn carried his KWIC stop list into his automatic abstracting procedure. Words with the greatest "resolving power" were in the middle range of frequency of occurrence in the document. The high frequency, common words were identified either by looking them up in the stop list or striking a cut off point on the upper end of the ranked

frequency list of the document's vocabulary. The lower frequency, minor words were also eliminated by their rank in the vocabulary. Luhn explained the rationale for using frequency of occurrence in the document as an indicator of content as follows:

The justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject. This means of emphasis is taken as an indicator of significance Within a technical discussion, there is a very small probability that a given word is used to reflect more than one notion. The probability is also small that an author will use different words to reflect the same notion. Even if the author makes a reasonable effort to select synonyms for stylistic reasons, he soon runs out of legitimate alternatives and falls into repetition if the notion being expressed was potentially significant in the first place. [Luhn, 1958, pg 378]

This principle of least effort was formalized in the 1940's by George Zipf. First rank the word frequency list and assign the rank 1" to the most frequent word, "2" to the next most frequent, etc. Plotting the frequency of each words against its rank number, using logarithmic scales, will yield points very close to a straight line with a slope of -1 [Wylllys, 1975, pg 30]. In other words, one word's frequency times its rank number is approximately equal to any other word's frequency times its rank number [Salton, 1983, pg 60]. Although it seems that good authors expend considerable effort avoiding duplication, Zipf's law

has been approximately verified, among other places, on the 1-million word Brown University corpus of modern English [Kucera, 1967, pg 357, Table 59].

Was Luhn right?

Intuitively it still appears that this method will not work well on the better authors with exciting styles (there are a few, even in technical writing). Sparck Jones and Kay give a more satisfying analysis of the problem of isolating the "significant" vocabulary:

The main difficulty is sorting sheep from goats Most frequent nontrivial words are good content indicators, but not all; and equally, most infrequent words are bad content indicators, but not always. In other words, topics of no distinctive interest are sometimes referred to repeatedly in the same terms, and others of interest are referred to repeatedly, but in different terms. [Sparck Jones and Kay, 1973, pg 136].

Both quotes seem reasonable. Perhaps neither the true import of word statistics in a document nor the best way of using them is intuitively obvious.

Relative frequency

Bar-Hillel was the first to criticize, on the basis of information theory, Luhn's assumption [Stevens, 1970, pg 81]. Others soon followed:

Very general considerations from information theory suggest that a word's information should vary inversely with its frequency rather than

directly, its lower probability evidencing greater selectivity, or deliberation, in its use. It is the rare, special, or technical word that will indicate most strongly the subject of an author's discussion. Here, however, it is clear that by 'rare' we must mean rare in general usage, not rare within the document itself. In fact it would seem natural to regard the contrast between the word's relative frequency F within the document and its relative frequency r in general use . . . as a more revealing indication of the word's value in indicating the subject matter of a document. [Edmundson, 1961, pg 393-4].

Just as it appears some theoretical progress is being made, the perversity of natural language arises:

Unfortunately, the frequency ratio is an example of an idea which, though beautiful in concept, is not easy to carry out in practice. . . . there would remain several difficult problems regarding the method of totaling. The most severe of these is how to deal with homography.

The "set," "square," "curve," and "integration" of mathematics; the "weight," "bond," and "heat" of chemistry; and "mass," "motion," "wave," and (visible) "light" from physics are all examples of terms carefully defined in particular disciplines yet also in general use with more common meanings.

Any such words would tend to be downgraded by use of the frequency ratio and be relatively unavailable for indexing The only kinds of general usage counts that could cause the relative frequency approach to work as intended would be one which permitted a computer to decide: Is this usage of the word 'bond' as a chemical bond or some other kind of bond? [Doyle, 1975, pg 309-10].

But the machine which could answer that question of meaning could probably also deal with the meaning of the other passages in the document, and a powerful Q-A system has arrived.

Doyle continues by pointing out that the most important contribution of the idea of relative frequency was that attention shifted away from self contained document descriptions (A) to the relation of description to description (B).

Back to basics

Though systems based on either absolute or relative frequency of occurrence work, better performance might be obtained if the theory of why they work were better understood. Here are some observations more fundamental than Luhn's:

. . . there is more information in word arrangement than in mere word frequency in an article's text. This point probably ought to be discussed at length, since it comes up repeatedly in criticism of statistical analysis of text. The author will, however, confine himself to the observation that the amount of information inherent in the document is ordinarily more than enough to permit accurate classification of the article. [Doyle, 1965, pg 487].

The general assumption behind the extraction of words on a statistical basis . . . is that conspicuous words are significant content indicators. It is not necessary to make any more concerted attempt to discover what a document is about, because a document wears its

heart on its sleeve, and any nontrivial word that occurs sufficiently frequently must be a valid content indicator, or it would not be used so often. Looking only at the surface of a document, it is clear that prominent physical features reflect important features of its content, so we need not examine the latter directly. [Sparck Jones and Kay, 1973, pg 134].

Since a document contains more than enough" information for its classification, only a part of that information should be adequate to produce a usable content indicator. Taken this way, as a reasonable appeal to the nature of documents (c.f. Zipf's law, page 135) rather than to the mechanics of writing, then statistical document processing is simply a matter of counting words and manipulating the numbers obtained. Will this approach provide an adequate theoretical basis for document processing?

Understanding an empirical approach

Salton is the most prolific IR statistician, and his SMART project has evolved since 1964 to the use of pattern recognition techniques for manipulating references. An arithmetic weight for each significant term in a document is calculated from its frequency statistics--number of occurrences in the given document, number of documents in the collection using the term, or frequency in the given document relative to frequency in the total collection. Suppose the system recognizes two significant terms. The two weights could be plotted on (Term₁, Term₂) axes to

give a point in Cartesian space. As it is, several terms are extracted, but their weights may still be calculated in the same way and the document described as a point in n -dimensional space, n being the number of "significant" terms extracted throughout the whole collection. This two-dimensional example will serve to discuss the validity of the more complex procedure. (In maximum-depth indexing, see page 133, n could be a hundred thousand or so, exploding the cost of some pattern recognition algorithms.)

The geometry and its rationale are clear enough, but what does it mean? Since two documents will rarely yield the same weights even if the same terms were extracted from them, then every document will be represented by a different point in the n -dimensioned space (A). The document description is an arithmetic vector or profile based on term frequencies and the plotted point is a convenient way of representing it. Casting these frequency profiles into the analytical geometrical mold simplifies examination of the relationships between documents (B). If two documents use about the same terms with about the same frequencies, their two points will be close to each other in the n -dimensioned space, and their underlying contents should be similar.

The field of pattern recognition has developed several similarity measures to calculate how close two vectors are

to each other, e.g. the Mahalanobis distance, Tanimoto's measure, [Tou, 1974, pg 87-88], the Dice, and the Jaccard Coefficients [Salton, 1983, pg 203]. Salton prefers the cosine coefficient function, which in vector notation is:

$$\text{Similarity}(\text{Doc}_P, \text{Doc}_Q) = p'q / ||p|| ||q||$$

where p and q are the n -dimensioned vector profiles of the documents whose similarity is to be measured and $||p||$ is the norm (distance from the origin) of vector p .

Since these functions calculate infinitely varied real numbers, then each document will have a different nearness to a query point. A retrieved document set may be ranked with the nearest document presented first.

If these frequency based document descriptions really indicate content, and if these similarity functions really measure similarity of content, then the entire job of classification may be done mechanically. There exist several cluster seeking algorithms which will automatically define subsets of any given set of clusterable points [Tou, 1974]. Whether points are "clusterable" or not depends upon the data they represent [Stevens, 1970, pg 110], and documents are universally assumed to be classifiable.

Unexpectedly, experiments along this line produce document classifications which are not "groupings that a classifier in his right mind could make." [Stevens, 1970, pg 112, 181]. But if retrievals based on the

classifications succeed, and if humans need not view the classification scheme, then odd categories are acceptable. However, as discussed below, evaluation is a tricky business; and, as far as FT is concerned, users wanting customized classification will want to view their classification schemes.

Therefore, the application of automatic clustering to FT will not be further pursued here. Instead, the fundamental problem of conceptualizing or verifying the correctness of treating documents like vectors will be discussed. Salton's favorite cosine coefficient function will serve as a good example.

He is not exactly measuring the distance between two points but rather the angle between their vectors. In Figure 3 document q is closer to p than to r because its vector makes a smaller angle with p's vector than with r's. The more common similarity function in pattern recognition is the simple, Euclidean distance measure:

$$\text{Similarity}(\text{Doc}_P, \text{Doc}_Q) = ||p - q||$$

under which q will be closer to r than to p because a straight line drawn from q to r is shorter than the one from q to p.

This distance function proclaims that q and r are closer because the weight of Term₁ in Doc_Q is close to

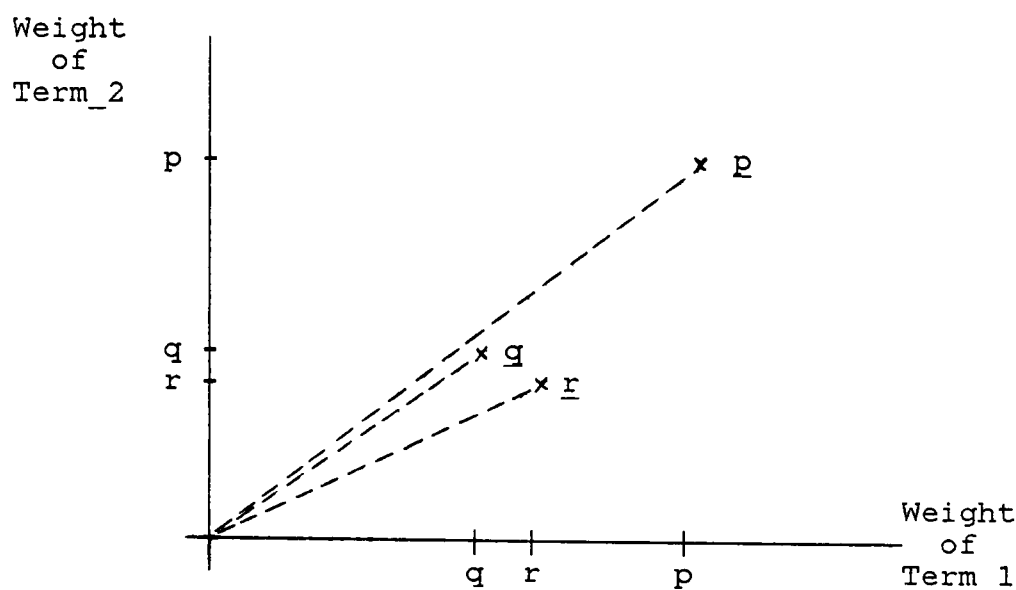


Figure 3 Three arithmetic document descriptions represented in 2-dimensional space

the weight of Term_1 in Doc_R AND the weight of Term_2 in Doc_Q is close to the weight of Term_2 in Doc_R. The closer together are the weights of respective terms in two documents, the sorter will be the straight line drawn between their vectors, the smaller the distance function, and the higher their similarity. This makes good sense.

However, the cosine coefficient function essentially compares the slopes of the vectors. That is, if the fraction $q(1)/q(2)$ --the weight of Term_1 in Doc_Q over the weight of Term_2 also in Doc_Q--is close to the ratio $p(1)/p(2)$ then the contents are close. Why should ratios of weights be more indicative of a document's content than

absolute weights? Theory and conceptualization begin to part company.

The reason Salton favors the cosine coefficient function surfaces after a little study. Though he never makes it clear in his much discussion of calculating weights and comparing them [Salton, 1983, pp 59-75, 120-130, 200-207], his weights are not normalized to eliminate differences in document length. That is, if Term_1 has the same percentage of occurrences in document S and T, but S is longer than T, then $s(1)$ will be greater than $t(1)$. Therefore the raw frequency values plotted on the Term_1 axis of Figure 3 depend on both percentage of occurrence and document length. Therefore, using Salton's weights, the simple distance measure, would vary as a function of document length. For instance, two unequal segments of the same document would exhibit different weights for the same term simply because their lengths differed. This appears counter to the original assumption that frequency indicates content. Therefore Salton prefers the cosine coefficient function which negates differences in document length. It measures angles, as though all documents had the same length. Compare Figure 4

Salton could achieve the same normalization more directly by making weights a percentage of document length to begin with, but is this an intuitively superior

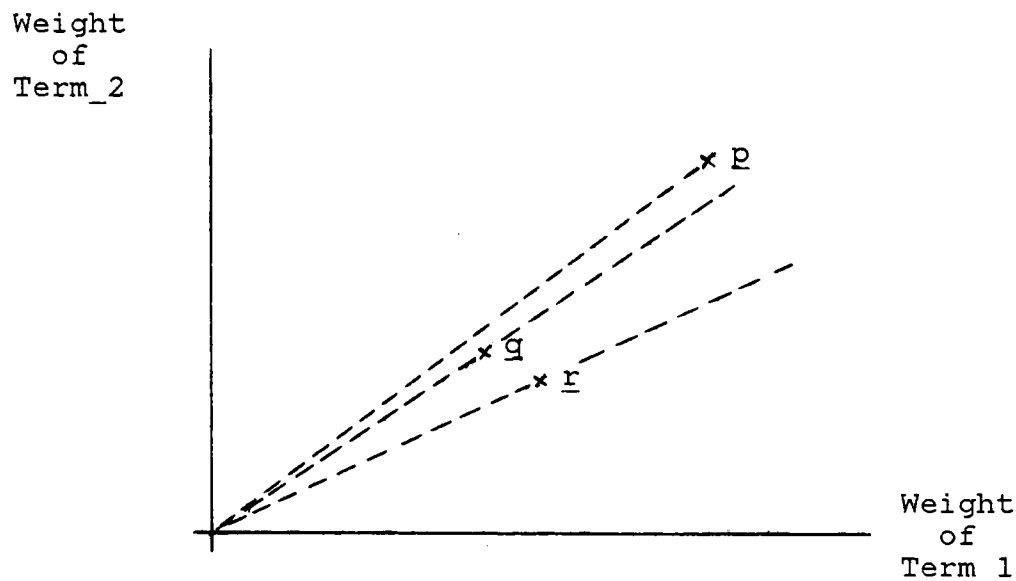


Figure 4 The cosine coefficient normalizes document length.

approach? Is document length really the unit on which to normalize weights? The 2% of 10,000-word Doc_U which uses Term_3, may be more relevant than the 10% of the 500-word Doc_V which uses it. And if document length is not the right unit, then the cosine coefficient function is an inferior measure.

The purpose of the above discussion is to illustrate the difficulty of building on the frequency-can-indicate-content theory. It seems to work [Salton, 1983, pg 267], but is not sufficiently developed to provide guidance through the complexities of natural language processing. Bar-Hillel's contention in 1960 is worth remembering:

The prospects for efficient use of computers for information retrieval have so far been greatly hampered by inept theorizing, by pointless attempts to 'apply' prestige-carrying scientific, and especially mathematical, disciplines . . . and by greatly underestimating the intellectual effort required to arrive at a good abstract, a good index set, and good judgment as to the closeness of topics. [Doyle, 1975, pg 334]

Experimental verification

Also disconcerting is experimentation suggesting that Luhn's auto-abstracting approach may have been wrong all along. Edmundson's study of 1969 compared extracts made manually with those made mechanically. This is not an IR experiment, but rather an experiment to see how often the sentences the computer extracts as most indicative of an article's content coincide with those a human would choose for the same purpose. Aside from the human document processing, Edmundson tried (1) extracting with Luhn's frequency and proximity criterion, (2) extracting by means of linguistic clue words, such as "in summary" or "Therefore," and (3) extracting sentences chosen randomly.

The random selections matched the human selections 25% of the time. Luhn's selections matched only 37% of the time. The clue words method matched 55% of the time. When Luhn's method was combined with the clue word method, the number of sentences chosen in common did not rise but fell below 55% [Doyle, 1975, pg 364].

Theoretical and experimental assurances of the quality of statistical document processing are lacking. Perhaps Lancaster's approach is best:

We should only be concerned with results . . . that are as good as or better than or cheaper than the results achieved by humans" (see page 131)

Try different systems and chose the one which gives the best measurable results.

Problems with evaluation

Several measures of retrieval effectiveness have been used [Salton, 1983, pg 174-184; Doyle, 1975, pg 337]. None will satisfactorily grade an entire system, even the two most popular ones, Recall and Precision. This pair of measures is a function of the number of documents retrieved by a query (Docs_Retrieved), the number of those judged relevant by the user (Rel_Docs_Retrieved), and the total number of documents the user would have judged relevant if he could have looked at all the documents in the system (Tot_Rel_Docs).

$$\text{Recall} = \text{Rel_Docs_Retrieved} / \text{Tot_Rel_Docs}$$

$$\text{Precision} = \text{Rel_Docs_Retrieved} / \text{Docs_Retrieved}.$$

The first problem is discovering Tot_Rel_Docs. For the user to look at all the documents in the collection is impossible unless the collection is unrealistically small. No one else can judge relevance for the user, therefore the

actual Tot_Rel_Docs may never be known until the perfect IR system arises to pluck them all out quickly. In other words, accurate evaluation of IR by these measures will be impossible until the IR is done really well.

Several approximating methods are used. Perhaps the most common in actual experiments is to sum the different Rel_Docs_Retrieved by different techniques and use the non-overlapping total as indicative of the true Tot_Rel_Docs [Sparck Jones and Kay, 1973, pg 19]. This may be good enough to compare one technique with another in the same system, but it gives little help in comparing system to system. Too many variables lie untouched by these measurements.

The second problem concerns judging relevancy. To get a real user with a real information need, run his query in the experimental system, and record his relevancy judgements can only be done once. The experiment cannot be repeated, either with another user because he will have a different query, or with the same user because, having seen a number of relevant documents already, his relevancy judgements will change. In fact, a few days after his search, the initial relevancy judgments of this user may change.

Experimenters have tackled both problems by having independent subject experts judge relevancy of documents to

experimental queries. Sometimes relevancy judgements are solicited for every document in the experimental collection against every query to be used in the experiment. The methodological problems here are best summed up by the "Silent Postulate" which C. N. Mooers suggests lies behind every IR evaluation experiment:

That (real people)
 (real documents)
 (real jobs to do) can somehow be
 eliminated from the experimental study, and
 that (role-playing people)
 (substitute documents)
 (imaginary jobs) can be substituted and
 still give valid experimental results.
 [Stevens, 1970, pg 104]

But encouragingly, some experimenters have discovered that changing relevancy judges, e.g. non-experts for experts, doesn't change the relative effectiveness of compared retrieval methods [Doyle, 1975, pg 356].

But the meaning of relevance is an even larger issue. To help satisfy an information need, a document must be new to the user as well as informative. A document too advanced for the user will not inform, though it may be on-subject [Boyce, 1982, pg 105]. Relevance is a very subjective decision.

Ideally, relevance in experiments would mean an objective correspondence between query and document, whether previously known or now informative to the user or not [Salton, 1983, pg 163]. But such a definition of

relevance begs the question. This correspondence is exactly what the IR system is working to establish. If some standard other than the satisfaction of human information needs is used to measure relevance, then the whole exercise has no point of contact with the real task it is trying to accomplish.

Finally, note that none of these problems of evaluation is unique to computerized IR. Evaluation studies only began in the computer age because the quality of IR a user previously received from a manual library was taken for granted. The quality of manual document processing is just as hard to measure as the mechanical.

The Edmundson study (see page 146) used the selections of human document processors as the standard by which to judge the selections made by computer. Using a human performance as though it were of high quality is wide-spread in IR experimentation. In the Edmundson study this use is probably valid. If the auto-abstracts do not summarize for people in the way humans want these articles summarized, then the machine has failed. But retrieving documents is a different matter. The right machine procedure might retrieve many more documents pleasing to a human user than any number of humans could retrieve manually from a large collection.

. . . the state-of-the-art in evaluation of indexing systems is in fact far more primitive than that of automatic indexing itself. An easy, and an early, solution is not likely. Therefore, today, in appraising machine potentials for assignment indexing we are faced with what is in effect a single criterion: namely, will a given group of human evaluators, whatever their standards and requirements, agree as much with the products of an automatic indexing procedure . . . as they do amongst themselves? [Stevens, 1970, pg 177]

D. R. Swanson articulate the difficulty in the often quoted final paragraph of the report on his 1960 experiments comparing human and machine retrieval:

In terms of these measures, the effectiveness of all information search techniques tested on the model was found to be rather low. . . . Thus, even though machines may never enjoy more than partial success in library indexing, a small suspicion might justifiably be entertained that people are even less promising [Doyle, 1975, pg 341].

There seem to be more problems than solid conclusions in the theory of IR system evaluation. Recall and Precision are relative measures, perhaps more useful for discussion about, than for evaluation of, systems. Progress means increasing the numerator of their respective fractions (see page 147). To improve Recall is to retrieve a larger document set (more Docs_Retrieved) containing a higher percentage of the Tot_Rel_Docs. Improving Precision means that, of the Docs_Retrieved, however many there are, more are relevant. One fixed law does seem to hold: that Recall

and Precision are inversely related. Increasing one inevitably decreases the other.

Most needed is a theoretical basis for perfecting statistical techniques and evaluating them. The theory will most probably come from linguists, for most of the descriptions of language used thus far have come from non-linguists, such as Luhn, Doyle, and Lancaster. Until then, the obvious approach to improvement is attempting to understand and mechanize human methods of document processing.

Basic linguistic research usually begins in opposition to the direction Lancaster intends to go:

. . . if we are going to use automatic processing rather than human processing we should not be concerned in trying to faithfully reproduce the activities now performed by humans. (see page 131)

D. Linguistically

A premise of Sparck Jones and Kay's entire book was the uncertainty that linguistics had significant contributions to make to IR beyond processing natural language queries. Salton [1983, pp 91, 267, 287] repeats the same wait-and-see attitude. But in the search for a solid theoretical basis for document processing, linguistics has begun to contribute by explaining some of the techniques already used.

While documents are processed manually, indexers and users presume upon the deeply hidden mechanisms of their language skills. But when computers, which had to be told everything, came along, information science turned to linguists for explanations of these skills [Sparck Jones and Kay, 1973, pg 3]. Among them, only Chomsky proposed a deterministic (and therefore computable) model of the process [Bott, 1970, pg 224].

A well formed sentence is a surface structure" communicating a meaning more efficiently held in a "deep structure." For instance, the difference between "Jake opened the door" and "The door was opened by Jake" is a purely surface one of active versus passive voice. The deep meaning of the two is identical [Lyons, 1970a, pg 121]. Since a document contains more than enough information for its classification, Luhn and statistical theories are saying that enough of a document's meaning can be estimated by looking at word meanings alone, without concern for structures--surface or deep. Equivalent terms, for instance, do not have to be reduced to a common "deep symbol" to show their equivalence, they need only be listed together in a thesaurus. And larger meaning is the sum of smaller meanings. Both the above sentences contain the terms "Jake," "door," and "opened." Further analysis for IR purposes is unnecessary (see page 138). The sentence, "Jake

did not open the door," has an opposite meaning but the same content in the sense that it would probably be of interest to the same user retrieving the former sentences.

Stop lists

Just as statistical processing sorts "sheep from goat" words (see page 136) using easily computed frequency counts, so other approaches attempt to make the same sort using some easily computed linguistic features of a document. Luhn prefaced his frequency processing with stop list action. Such a list, as used for instance in a KWIC processor (see page 27), is a linguistic filter subtracting syntactic function words which define surface structures. A stop list for the above three sentences would be

by
did
not
the
was.

The assumption is that wherever in text these words occur, they are of interest neither as access points nor as bearers of content. As a stop list grows longer, this assumption becomes less tenable.

Clue words

Contemporary with Luhn's auto-abstracting was Phyllis Baxendale's work on automatic document processing. She

chose as most significant the words in the first and last sentences of each paragraph, having examined 200 paragraphs to learn that the topic sentence appeared first in 85% and last in 7%. In another trial, she chose the four words following each preposition, unless finding punctuation or a second preposition. The same words with approximately the same relative frequencies were chosen by the topic sentence and the prepositional phrase methods, as well as by a simple stop list subtraction from the total vocabulary [Edmundson, 1961, pg 403]. Subsequent sorting suggestions include honoring title or sectional heading words which also appear in text, picking words in the same sentence with "summary," "conclusion," "prove," "it was found that," "the central problem is," etc., and simply taking the words of first and last paragraphs of the whole article [Stevens, 1970, pg 85-86]. FT allows much more extensive investigation of this syntactic clue word method of designating the sheep [Sparck Jones and Kay, 1973, pg 132].

Stemming

Another linguistic process already discussed is stemming (see page 129). English, less than more declined languages such as German or Russian, changes the spelling of some words in accordance with their functions in the surface structures in which they appear. "Opened" versus

"open" in the above example. As in an analysis for deep structure, IR interest also centers on content rather than surface form. Stemming to reduce different surface forms of the same word to a common root improves Recall.

Obviously a computer which could convert any well formed sentence into the deep structure it represents (into "a representation of the ideas it contains," see page 77) could attack several IR problems directly. Stop list, clue word, and stemming algorithms work, but the total linguistic problem is presently beyond any but limited solutions for two reasons. (1) Though the transformational grammars of Chomsky and others promise to show how any given surface structure may be generated from its deep structure, their theories are neither complete nor gaining consensus [Lyons, 1970a, pp 119-139; Sgall, 1976, pg 106]. (2) Even if the process for generating all the well formed sentences from a deep structure were perfected, the reverse process of deriving a deep structure from one of its surface structures explodes in complexity because of ambiguities in surface structures [Sparck Jones and Kay, 1975, pg 97]. To resolve these ambiguities requires knowledge of the properties and histories of the things discussed (the larger-context problem of page 22 and the cognitive data structure of page 78).

Again comes the perversity of natural language processing. Small steps in linguistic computing have been made, but each further step seems to present a multiplicity of new problems. Certainly progress in linguistic processing will be made, but the surrogate memory Q-A system (see page 80) seemingly awaits comprehensive solutions to these problems. However FT, with its lower expectations (see page 81) may be able to take advantage of partial progress in linguistic processing.

IV. PROPOSED LINGUISTIC ADDITIONS TO A STATISTICAL SYSTEM

To process a document is to produce a description from it for later use in retrieval. The system index is an easily searched list of document descriptions. Many methods of processing documents and compiling indexes have been discussed in the first three sections. All except Maximum-depth indexing (see page 133) aim for brief document descriptions (A) so that each document is briefly distinguished from other documents (B) and each term retrieves only those documents giving prominence to that term.

If documents must be characterized in a fairly summary way {shorter, see page 106}, and allowance must be made for requests approaching topics from very different angles {desirable, see page &Luhndesire.}, descriptions must be designed for gross rather than refined matching. The experiments just mentioned {SMART, Western Reserve University, Lancaster's} support this view. [Sparck Jones and Kay, 1973, pg 104].

Sparck Jones and Kay suffer from the limited storage space syndrome. Documents must be characterized in a general way to compensate for the loss of detail in shortened document descriptions. Minor themes of any document are not searchable. Since the document's characterization is summary," the subsequent matching must be "gross rather than refined," sacrificing Precision.

Perhaps the "law" of the inverse relationship between Recall and Precision (see page 151) is really only a function of document description length, a phenomenon of "gross matches" to "summary" descriptions. As long as document descriptions contain only the most prominent ideas in a document, the more precise query can have no effect other than retrieving the fewer documents which give prominence to the query terms. But if indexing is to a greater depth, then the more precise query may dig for documents with relevant passages among their minor emphases. Themes may be retrieved which neither author, abstractor, nor indexer might have considered important at the time the document was processed.

The following improvements, through linguistic techniques applied to a statistical system, will have the effect of giving the interactive user more information about the total collection or of various subsets of it.

A. Deeper Indexing through Phrase Formation

The problem

In the purely mechanical system, to provide more access points (terms) for each document is also to multiply the postings under each term and devastate Precision (see page 133). As the collection grows a user needs longer and

longer queries to achieve the same precision. An unsatisfactory solution is the current practice of maintaining separate databases within a single system. DIALOG, BRS, etc. acquire bibliographic records in separate blocks from different database producers. Duplicate records are not culled. Westlaw and LEXIS divide cases by Federal court district or State. Users must always choose a database, sometimes blindly, as their first step into the system.

The ranking possible in a statistical system (see page &Ranking.) is a workable solution. For instance, the 120,000 document, DIRECT system, at Philips Corporation, The Netherlands, uses automatic indexing which produces document vectors much longer than the vectors resulting from manual, controlled vocabulary indexing. The excess is handled in this case by unlimited ranking and the lengthening of queries with large groups of synonyms and quasi-synonyms.

For example, the single term "PASCAL" would retrieve documents about a man, a computer, a programming language, and a European information system. The additional ORing of "programming, programming language, higher-level language, compiler, software" insures that the abstracts using "PASCAL" as a programming language will be ranked near the top. Others which contain fewer, or none, of the extra

terms are also retrieved, but ranked so low that most users would not stick with the session long enough to view them.

. . . he can easily regulate the wanted recall and precision of his search: if he is interested in 'something about a subject', preferring high precision at lower recall, he scans a few documents at the top and he stops his session at an early moment. If he is interested in high recall, he goes on along the queue of documents less and less associated with his inquiry and stops the search at any later moment. [Stibic, 1980, pg 275]

In one of the system databases, over half the abstracts, 6,288 of them, contained one or more of the words in the following narrative-type query:

Trends of development of computers and automation. Impact of the growing use of computers on the social, cultural, economical and political evolution of nations and on international relations. Are computers, automation and robots dangerous for the privacy, freedom, employment and welfare and for the future of mankind? [Stibic, 1980, pg 276]

Rank is expected to bring to the top documents with the greatest similarity to the most query terms.

But as long as the weights--which make ranking possible--are assigned or used on a percentage-of-document-length basis (see page 144), many documents with near-subject prominent themes must be viewed to find the one with the on-subject minor theme. And the user has the burden of composing lengthy queries. A better solution is to get away from summary" descriptions and extract terms of greater specificity for document descriptions.

The problem aggravated by deeper indexing is that as the number of postings increases, the number of terms in the index rises very little, if at all. Naturally each term in the index becomes less selective in the documents it retrieves.

A solution

Salton [1983, pg 67, 86] describes these extra postings as broad, high-frequency terms with negative discrimination values and excessive document frequencies. He means that they occur in lots of documents and the frequency of occurrence does not differ significantly from document to documents. Therefore, to a statistical system they are noise. To use these terms, Salton recommends combining them with other low-discrimination-value terms to form phrases with more specific meaning than the original terms would have [Salton, 1983, pg 85].

Therefore, the solution to deep indexing in the FT system is to form phrases of the higher frequency terms. In this way, as the number of documents in the system increases, the total number of terms in the index can also increase because new index phrases are being formed.

When postings to, say, Term_1 grow to the point that a phrase should be formed, the system searches its FT for possibilities. The simplest phrase is an adjective-noun

adjacency. More complex phrases may be found using various partial parsing approaches reported in Salton [1983, pp 285-286] and Sparck Jones and Kay [1973, pp 107-108]. The system will search for all its possible phrases using Term_1 and choose the most frequent phrases as new, compound index terms. This search will be $O(k)$ where k is the number of Term_1 postings which triggered the phrase searching.

The postings where Term_1 does not appear in the new phrase are left unchanged, and a pointer is set from the Term_1 entry to the new phrase entry. Postings to the other terms drawn into the new phrases must be reduced. As a result the total number of index terms for the collection has increased while some of them have become more selective. That is, the phrase is more Precise than the intersection of its constituent words. Also, a query on the original Term_1 returns (1) fewer documents than before, plus (2) a notice that more are available, but on a more specific topic.

Cost

Maintaining weights requires a partial reprocessing of the documents containing the phrase. This will be an $O(1)$

operation, where l is some portion of k above, $l < k$. Therefore, the processing for one phrase formation is $O(k) + O(l) < O(2k)$.

The number of phrases versus single terms in the system will depend on the size of the triggering point. As in any processing of natural language, it is difficult to foresee the problems in building such an index; but conceivably the building could iterate, that is, phrases of phrases may be build as well as phrases of single terms.

Note that these natural language index phrases arise as a function of the number of documents added to the system, without reference to queries. Each time a document is added, some fraction, r , of the hundreds of terms in its vocabulary, v , may trigger the phrase forming process. Each of the rv terms invokes an $O(2k)$ process for a total computational complexity per added document of $O(2rvk)$.

That is:

- k - the number of postings to a term, considered excessive for good Précision, which initiates phrase forming. It is a constant set by the system managers.
- v - number of different words used in a single document. Varies with the document and cannot exceed the number of words in the longest document. Probably less than a few thousand.
- r - fraction (probably small) of v whose postings cause certain terms to trigger phrase formation.

" v " is bounded by subject field, author, and, after a point, rather insensitive to document length. When k is small for highest Précision, r will be large. But their

product is nowhere dependent upon the number of documents in the system and can always be adjusted by changing k .

The penalty for a small k is a large index with lots of phrases. Therefore the thesaurus maintenance computation for a single document, added to the complexity of term extraction and weighting, is $O(ck)$ where $c = 2rv$ and is bounded by vocabulary usage and dependent on k . Over the life of the system--all the thesaurus maintenance it will ever have to do--is $O(ckn)$ where n is the number of documents in the collection.

B. Automatic Thesaurus Maintenance

This index is really a post-controlled vocabulary (see page 125), and a viewing of it, as with DIALOG's EXPAND command (see page 58), shows a searcher what terms to use and how many documents each will retrieve. A very valuable additional service would be to establish hierarchical relations within this automatically built vocabulary. Relating single terms is a complex semantic task most efficiently done by humans. An existing thesaurus could be incorporated into the system and automatically updated to suppress terms not actually found in the collection. But the more interesting feature is an automatically maintained hierarchy as the system builds phrases.

How it would look

It should not be difficult to make explicit the uniterm "classification" scheme previously explained (see page 113). Whenever a phrase is formed, e.g. "cigarette production," different approaches may be useful to different users. One may want the subordination:

cigarette - NT consumption, marketing,
production, ...

while another may want:

production - NT candy, car, cigarette,
cloth, ...

depending on what other phrases have been formed from the FT collection. Providing either display at an interactive terminal is simply a matter of which set of phrase pointers, those from "cigarette" or those from "production," are followed. This example shows how easily the simple hierarchial relationship of adjective-noun may be reversed.

Assuming that more complex phrases are build, they may be manipulated using articulated index or PRECIS type algorithms (see page 120) where the previously discarded function words are used to indicate rotation points. One of several phrases formed in a metal working collection could be "the Preparation of Steel plates by Forging." It would produce PRECIS type entries of:

Forging.
Steel plates. Preparation.

Steel plates. Forging.
Preparation.

Preparation. Steel plates. Forging.

and the adjective-noun adjacency phrase, "Steel plates," could be handled at a separate level as before.

Suppose a less informed user has an information need which this document would satisfy. His clearest concept may be "plates," which he enters into the system.

User: EXPAND plates

Sys: Plates - NT iron 11, pressed 50, rolled 15,
scaled 1, steel 235, ...

User: EXPAND plates, steel

which the system, not being linguistically sophisticated, is not ready to accept. Instead,

Sys: WHAT DO YOU MEAN plates, steel?

Plates - NT iron 11, pressed 50, rolled 15,
scaled 1, steel 235, ...

Steel - NT plates 235, rod 179, sheets 17,
wire 433, ...

The user is expected to use the syntactically correct phrase, and is encouraged to continue trying because he is making progress. The second frame has more information about the collection than the first frame, and the words he wants are there, he just has to get them in the right order.

User: EXPAND steel plates

Sys: Steel plates 102

Steel plates.

Forged. 13

Hammered. 14

Rolled 10

Steel plates. Forging.

Cleaning. 11

Preparation. 17

Recovery. 35

Steel plates. Rolling.

Bending. 10

Preparation. 23

etc.

TOTAL DOCUMENTS ON Steel plates. 235

meaning that there are 102 documents using the phrase "steel plates" but not yet incorporated into longer phrases; 13, 14, and 10 documents containing phrases such as, "the forging of steel plates"; and the remainder containing the longer phrases indicated.

Not only does the PRECIS type index show the contents of the system, it also communicates something about the subject matter at hand. What can be done to steel plates? They can be forged, hammered, or rolled. What is accomplished by forging? Cleaning, preparation, etc. Hopefully these clues will be enough for the user to write his desired query:

User: RETRIEVE preparation. steel plates. forging.

If not, further EXPANDs on "forging" or "preparation" should eventually lead to the same RETRIEVE request.

British experience with PRECIS indicates that dropping the

function words does not make entries too cryptic, as long as syntactical ordering is maintained. But for a clearer display the function words could be retained.

The object is to help the user produce his best query. The classification indicated by "NT" (see page 123) is not hierarchial in the traditional sense. "Plates" is as likely to be subsumed under "Steel" as vice versa. The adjective-noun displays only show the user how he may enter the system. It gives him a choice of productive access points from which he may apply his language skills to produce a syntactically correct query.

The PRECIS type display does represent a more traditional type hierarchy, but not because the system understands concepts. For instance, the system doesn't know whether "Rolling" or "Bending" should be the superior term in a classification scheme. Rather, the formed phrases are syntactically correct because extracted from a pool of syntactically correct phrases (the FT). Their syntactic content is maintained by the PRECIS type rotations so that a user may look at the display and form a good query.

This simple combination of uniterm classification and PRECIS rotations will not extract all the good indexing phrases from the FT. There should be in the range of 60% agreement between terms chosen by this type of automatic system and terms assigned to the same documents manually

[Salton, 1983, pg 100]. One way to tune this system would be to compare it with the PRECIS output of the British National Bibliography on identical documents, assuming, of course, that humans are doing a good job (see page 151).

C. Learning from Queries

An additional source of information useful to the system for indexing may be derived from users. What the system needs is syntactically correct indexing phrases reflecting the interests of the user population. The automatic thesaurus maintenance procedure discovers such phrases from authors and assigns them, not to every document which should have them, but to those which use the phrases verbatim.

The system follows any document set retrieved on phrase matching with a good number of additional documents ranked according to statistical similarities between document vectors and a vector formed from the query. If the initial document set does not satisfy the user, he may continue sifting documents presented from the less specific, statistical mode. By displaying the phrase matched documents ahead of the others, the system has increased a user's chances of finding the minor but on-subject passage, i.e. has increased Precision.

Each time a user presents a fresh query written in the articulated index form, the system has available another, presumably legitimate, potential index phrase. The attitude of the system should not be, "I don't have this phrase in my index. Reject it," but "I don't have this. Let's process it so I can add it to my index." Therefore, users are instructed to present detailed, subject-heading type queries using prepositions and any other function words or verbs handled by the partial parsing of the system. A phrase thus presented may be established in the index by finding it in one or more of the collection's documents, without waiting for any triggering mechanism to initiate the search. As before, this document set, empty or not, is followed by retrievals ranked statistically.

D. Automatic Relevancy Feedback

Assigning index terms to documents in which they do not occur (assignment indexing) is the chief weakness of all mechanical systems. More advanced parsing will help; a complete Q-A system will cure it. An FT system has one additional source of information to apply to this problem.

A new measure of relevancy

Since the system presents full text for the user's inspection online, the time a user spends looking at each

page indicates his interest in it. By timing his interest, or eliciting simple evaluations, the system gains an appropriate measure of relevance. Systems which only present titles or abstracts, can, at best, only solicit a user's guess at the relevance of the document referenced. More normally, relevance judgements for such a system are taken at some indefinite time after the user has had a chance to chase down the FT, or has decided, for reasons which may have nothing to do with the performance of the system, that he will not. Though measuring relevancy from the moment a user first lays eyes on the FT may not be the best time, it is certainly a reasonable, uniform, and convenient time to do so. And this way of measuring has the advantage of defining relevancy on the basis of user actions rather than user opinions. The approach is the same as if a reference librarian evaluated his performance by leading a patron to a source and then timed how long the user maintained interest in the book.

Compensating for individual differences in reading speed or evaluation standards is fairly simple. Documents which phrase-match to a user's query and engage his attention for a minimum length of time are assumed to be on-subject, one hundred percent relevant. Measurements taken from the phrase-matched set become a standard for that user's "relevancy span."

Using this measure

In case one, a set of phrase matched documents is followed by a set of statistically ranked documents. The user may or may not know when his viewing crosses into the ranked document set. Relevancy measures taken there are compared with the earlier, standard measures. When a later measure matches or exceeds the standard, that document is assumed to be as relevant to the current query as one from the phrase-matched set. The phrase is added to the document's description in an act of assignment indexing based on a user's rather than on an indexer's evaluation.

Sometimes the assumption may be wrong. Something other than the topic of the current query catches the user's interest. But this error, if reasonably bounded, can benefit others. And who is to say that one user's "diversion" will not be the next user's "inspiration" as his mind catches the former connection and brings his own originality to bear on this "accidental" association. In this way, a community of users may build on each other's insights, like Bush's Memex user building new associations on his past associative trail (see page 44). This is certainly an example of Luhn's "dynamic interpretation at the output phase" (see page 75).

However, if members of this community are competing with each other, one individual may not want another

"spying" on his associations. Adequate security levels can easily be incorporated for privileged users. Small town Westlaw installations have already encountered this situation to an extent. When the terminal resides in the town's Bar Association office, crafty attorneys have been known to follow another man on the terminal, and at his departure execute a RESTORE type command to discover his query. The solution is straight forward. Those conscious of the problem close their sessions with a final, bogus query. Then the next RESTORE produces a red herring [Eppes, 1983].

In case two, the phrase-matched document set is empty and only a statistical document set is presented. Those documents receiving the highest relevance measurements may tentatively have the current query phrase assigned to their document descriptions. When another user specifies the same phrase, now presented in some of the EXPAND displays, this document will be presented first followed by a statistically retrieved set. Relevance feedback from the new user may be treated in the stronger, case one mode of balancing between phrase-matched and statistical sets. However, the tentative nature of the case two index additions must be maintained until confirmed by a certain number of users.

What happens when two users present the same query--as could easily happen in a student academic library--but make different evaluations? Their differences may be cumulatively average in the weight assigned to the new document descriptor, giving a consensus of indexing not previously available in IR. As user assignments accumulate, documents are gradually reclassified in response to current and informed opinions of their meanings.

Correcting the new measure

Missing so far in the discussion has been any mention of a retreat from an automatically assigned phrase or weight adjustment. Suppose the system makes a flat mistake. A phrase is created and assigned to a document description which no one ever uses again. A portion of the weight those terms carried towards useful relevancy is permanently lost, and a note will be given to all users in the neighborhood, "SEE ALSO x", where x is the useless phrase. Or suppose a cuckoo user provides really dumb relevancy feedback and manages to attach a high weight to a term very unrepresentative of the document. Despite subsequent averaging, that weight or that term will linger at some level in the system forever.

Salton [1983, pg 146] corrects his vectors both positively and negatively. A false movement in one direction will eventually be completely corrected by several movements in the direction most users want the document to go. In the method of assignment here presented, only positive corrections are given. Therefore, some delayed stability needs to be built into this system. If the strange phrase gets assigned to the index and never used for a time, it should be removed and the pre-assignment weights restored. If, after the dumb weight alteration occurs, the document receives no approvals and lots of rejections, the original weight should be restored. Gathering the statistics for making these decisions should be easy.

The counting of rejections and approvals would probably be done in any case the sake of those in charge of acquiring documents for the system. Implementing the stability described above is only a matter of determining the cutoff points experimentally. Unlike experiments in relevancy evaluations, this monitoring is strictly objective. Once assigned, has this phrase ever been used by anyone? Did the alteration of weight increase or decrease to a statistically significant degree the approvals received by the document affected? These statistics indicate the true value of the system to its users.

E. Conclusion

The purpose of this thesis has been to show the erratic path of FT development and to enunciate some principles from which it may progress more directly. The clue was given by Doyle [1961] in the aptly titled article, "Semantic Road Maps for Literature Searchers." His idea is often repeated but never improved. IR computers have been primarily used as searching instruments with concentration on optimizing the matching of queries with document descriptions. They ought rather to be used to arrange and label things for human searching.

Doyle cites the open shelving of a supermarket as a positive example. Goods are spread out for easy viewing in a standard classification scheme. A shopper may see at a glance whether he is in the bread or meat section and proceed to look at finer details to make his selection.

Or consider a magazine rack, where the manually prepared covers and titles attract the eye. The contents of any single item may be quickly sampled, just as FT allows the quick sampling of its items. Browsing in a newspaper is a similar process of making relevancy decisions on the basis of titles serving as document descriptions. That the magazine cover or newspaper headline is often misleading does not matter so much because the text is immediately at hand. Errors in the overview may be quickly corrected--with

newspapers or FT--by examination of the finer details, the full text. Adaptable humans quickly learn how to bias the overview in deciding whether to retrieve or not. To have the overview in the first place is much more important than its accuracy. The clear lesson to FT systems is: Make your contents more visible.

Most current computer IR is like trying to browse on a roll of microfilm. Every roll or fiche looks like every other. To make use of microforms the user must identify and locate a particular item before getting any content. Contrast the subject field and collection information communicated by the PRECIS type display (see page 168).

To represent documents and sets of documents, Doyle suggested the display of pictures of networks. Words are to be joined by lines--statistically or topically prominent words shown in bolder type, stronger relationships between words shown in thicker lines. With the growing popularity of graphic terminals, these semantic roadmaps seem attractive.

But words connected by lines are hard to understand. Even though familiarity could make semantic roadmaps a very effective means of estimating document content at a glance, the nature of FT systems requires service to infrequent users. That is, forming the right query from PRECIS type entries and examining the FT for relevancy are tasks only

the person with the information need can do, not intermediaries. And the normal FT user will not need so much information that he will become a frequent user. Words in index or outline form, as described earlier (see page 167), is the more natural way of communicating collection information to the normal user.

We are, after all, dealing with the elemental situation of an author talking to a reader--even if by means of a buffer storage which will grow more and more mechanical. The basic problem is to increase the mental contact between the reader and the information store, so that the reader can proceed unerringly and swiftly to identify and receive the message he is looking for. [Doyle, 1961, pg 553]

LIST OF REFERENCES

LIST OF REFERENCES

A star, "*", marks a few references
which are basic to the field of FT.

- Anthony, L. J. (1983), "European Notes," a regular column,
Online, 7(2), March, pg 86.
- Austin, Derek (1976a), "The Development of PRECIS, and
Introduction to its Syntax," "The Semantics of
PRECIS: Vocabulary Control and the RIN System," and
"Management Aspects of PRECIS and Current Research
and Development," in [Wellisch, pp 3-80].
- Austin, Derek (1976b), "Perspective Paper: Library
Science," in [Walker, 1977] pp 45-56.
- Bar-Hillel, Yehoshua (1964), Language and Information,
Addison-Wesley, Reading, Mass.
- Bauer, Charles K. (1963), "Practical Application of
Automation in a Scientific Information Center--a Case
Study," Society of Automotive Engineers-American
Society of Naval Engineers paper No. 698C, reprinted
in [Sharp, pp 569-611].
- Bernier, C. L. and E. G. Crane (1962), "Correlative Indexes
VIII: Subject-Indexing vs. Word-Indexing," J of
Chemical Documentation, Vol 2, as quoted in [Stevens,
pg 67].
- Boodson, K (1976), "Subject Bibliographies in Information
Work," in [Harrod, pp 271-279].
- Boral, H., D.J. Dewitt, W.K. Wilkinson (1982), "Performance
Evaluation for 4 Associative Disk Designs,"
Information Systems, 7(1), pp 53-64.
- Bott, M. F. (1970), "Computational Linguistics," in [Lyons,
1970b, pp 215-228].
- Boyce, Bert (1982), "Beyond Topicality: a Two Stage View of
Relevance and the Retrieval Process," Information
Processing and Management, 18(3), pp 105-110.
- Boyle, Harry F. (1982), "The Pricing of Information--a
Search-based Approach to Pricing an Online Search
Service," Online Review, 6(6), Dec, pp 517-523.

- Branscomb, Lewis M. (1979), "Information: the Ultimate Frontier," *Science* 203(4376), 12 Jan., pp 143-147.
- Buckland, Lawrence F. (1980), "Book Composition by Authors--System and Equipment Consideration," 5 page reprint available from Inforonics, Inc. 550 Newtown Rd., Littleton, MA 01460.
- * Bush, Vannevar (1945), "As We May Think," *Atlantic Monthly*, 176(1), Jly., pp 101-108, reprinted in [Sharp, pp 19-41].
Still fresh and stimulating after nearly 40 years. Required reading for anyone in the field of FT. Highly cited.
- Business Week Magazine* (1974), "Putting Law Libraries into the Computer," Jan. 26, pg 36.
- Cawkell, A. E. (1980), "Electronic Information Processing and Publishing--Problems and Opportunities," *J of Information Science*, Vol 2, pg 189-192.
- Chamberlain, Steve (1983), manager of Cumberland Ave. Radio Shack, Knoxville, Tenn., in personal conversation, April 9.
- Christie, Bruce (1981), Face to File Communication--A psychological approach to information systems, John Wiley and Sons Ltd.
- Collison, Robert (1971), Abstracts and Abstracting Services, American Bibliographical Center--Clio Press, Santa Barbara, Calif.
- Computer Magazine* (1983), "New Products," 16(1), Jan, pg 91.
- Cushing, Ralph (1963), "Improved Personal Filing System," *Chemical Engineering*, 70(1), also in [Sharp, pp 211-243].
- Deuel, Leo (1965), Testaments of Time--The Search for Lost Manuscripts and Records, Alfred A. Knopf Inc., New York.
- * Doyle, Lauren B. (1975), Information Retrieval and Processing, Wiley-Becker & Hayer Series Book, Melville Publishing Co. Los Angeles, Calif.
Though a third is devoted to microform and computer hardware, the book is very well written with creative insights and excellent sense of history. As

easy to read as Salton is difficult. Age is not that important.

Doyle, Lauren B. (1965), "Is Automatic Classification a Reasonable Application of Statistical Analysis of Text?," J of the ACM, 12(4), Oct. pp 473-489.

Doyle, Lauren B. (1961), "Semantic Road Maps for Literature Searchers," J of the ACM, Vol 8, pp 533-578.

Encyclopedia of Library and Information Science (1975), "Legal and Legislative Information Processing in the United States," Marcel Dekker Inc., New York, Vol 14, pp 183-194.

Edmundson, H. P. and R. E. Wyllys (1961) "Automatic Abstracting and Indexing--Survey and Recommendations," Communications of the ACM, 4(5), May, reprinted in [Sharp, pp 390-412].

Eppes, Bill (1983), slaesman for Westlaw in training session at UT Law Library, March 29.

Faden, B. R. (1959), "Information Retrieval on Automatic Data Processing Equipment," Special Libraries, 50(4), April, reprinted in [Sharp, pp 319-344].

Fenichel, Carol H., Thomas H. Hogan (1981), Online Searching: A Primer, Learned Information Inc., Marlton, N. J.

Fortune Magazine (1980), "Why TV Sets Do More in Columbus, Ohio," Oct. 6, pp 67-73.

Foster, Caxton C. (1976), Content Addressable Parallel Processors, Van Nostrand Reinhold Co., New York, N. Y.

Hamilton, L. Clark (1973), "Automated Law Research in Congress and the other Federal Law Branches," in [May, pp 109-111].

Harrod, Leonard Montague (1978), ed., Indexers on Indexing--A selection of articles published in The Indexer, R. R. Bowker Co. New York N.Y.

Harter, Stephen P. and Kenneth F. Kister (1981), "Online Encyclopedias: the Potential," Library Journal, Sept, pg 1600-2.

- Henkle, M. David (1973), "Open Meeting Discussion," in [May, pp 121-123].
- Hiltz, S. R. and M. Turoff (1978), The Network Nation, Addison-Wesley, Reading, Maine, as quoted in Online Review, 5(4), 1981, pg 300.
- Hogan, Thomas H. (1979), "News Retrieval Services--growing but where are they headed?," Online Review, 3(3), Sept. pg 247.
- Horowitz, Irving Louis (1982), Mary E. Curtis, "The Impact of the New Information Technology on Scientific and Scholarly Publishing," J of Information Science, 4(2-3), May, pp 87-96.
- Inforum--the Newsletter of the New York Times Information Service, (1981), "Announcing the New York Times-Online--The full text of all New York Times stories with full bibliographic and editorial enhancements," July, pp 1-4.
- Jonker, Frederick (1963), "The New 'Termatrex' Line of I.R. Systems--the 'Minimatrex' Line of I.R. Systems," American Documentation, 14(4) Oct., reprinted in [Sharp, pp 244-258].
- Kahen, M., V. Reich, and L. Cohen (1981), "Influence of Online Bibliographic Services on Student Behavior," J of American Society of Information Science, 32(6), Nov., pg 412.
- Knoxville City Ordinances, 5577, July 17, and 5612, Sept. 25, 1973, City Recorder's Office, City County Building, Knoxville, Tenn.
- Kodak (probably 1976), "Kodak Oracle Equipment--microfilmer/retrieval terminal/accessories," brochure A-1924, 3/76, printed in USA, four unnumbered pages.
- Kucera, Henry and W. Nelson Francis (1967), Computational Analysis of Present- Day American English, Brown University Press, Providence, R.I.
- Kutz, Myer (1983), "The Upside of Downloading," Wiley Search Update: User Newsletter, John Wiley & Sons, New York, NY, 1(2), Spring, pg 2.

- Lancaster, F. W. and Julie M. Neway (1982), "The Future of Indexing and Abstracting Services," *J of the American Society for Information Science*, May, pp 183-189.
- Lancaster, F. W. and Herbert Goldher (1981), "The Impact of Online Services on Subscriptions to Printed Publications," *Online Review*, 5(4), Aug, pp 301-311; and two opposing "Letters to the Editor" from Roger K. Smith and A. W. Elias, 5(6), Dec, pp 496-497.
- Lancaster, F. W. (1976), "Perspective Paper: Information Science," in [Walker, pp 19-43].
- Lawlor, Reed C. (1973), "Historical Perspective and Future Prospects," in [May, pp 5-34].
- Lawrence, Gail H. (1983), "Libraries Can Learn from Banks," *Library J*, Jan. 15, pp 105-107.
- LEXIS (1982), pamphlet, Libraries 1982, Mead, printed in U.S.A., April, 16 unnumbered pages.
- LEXIS (1980), pamphlet, A Primer, Mead, printed in U.S.A., 16 pages.
- Luhn, H. P. (1958), "The Automatic Creation of Literature Abstracts," *IBM J of Research and Development*, 2(2), April, pg 159, also in [Sharp, pp 375-389].
- Luhn, H. P. (1957), "A Stastical Approach to Mechanized Encoding and Searching of Literary Information," *IBM J of Research and Development*, 1(4), pg 309, Oct.
- Markee, Katherine M. (1981), "Economics of Online Retrieval," *Online Review*, 5(6), Dec.
- May, Ronald A (1975), ed., Sense and Systems in Automated Law Research{--proceedings of Second National Confernece on Automated Law Research in Monterey, Calif., Oct, 1973}, American Bar Association Press.
- Martin, Thomas H. (1973), "The Features Legal Researchers Need In an Interactive Retrieval System--as derived from an analysis of eleven interactive systems," in [May, pp 53-64].
- McGonigal, Richard M. (1973), "Computerized Legal Research: One Law Firm's User Experience," in [May, pp 113-119].

- McMurdo, Geroge (1980), "The Interface Between Computerized Retrieval Systems and Micrographic Retrieval Systems," J of Information Science, 1(6), March, pp 345-349.
- Meadow, Charles T. (1979), "Information Science and Scientists in 2001," J of Information Science, 1(4), Oct., pp 217-222.
- Nash, Mary (1982), "Videotex Versus Online Bibliographic Databases: a Canadian Prospective," Online Review, 6(4), Aug., pp 291-296.
- Nash, Mary M. (1979), "The Globe & Mail Database--a Canadian First," Online Review, 3(4), Dec., pp 367-371.
- Nau, Dana S. (1983), "Expert Computer Systems," Computer 16(2), Feb, pp63-85.
- News, a regular feature of Online Review, date and page as indicated at citation.
- Online Review (1982a), "News," 6(4), pg 344; (1981) "News" and Ad, Vol 5, pp 91, 429, 501; (1980) Ad and "Letters to the Editor," Vol 4, pp 139, 180.
- Perry, James W. (1960), "Information Retrieval: What Will It Do For The Chemical Engineer?," Chemical Engineering, 67(26), reprinted in [Sharp, pp 286-313].
- Posner, Raphael (1972), "The Index of the Encyclopaedia Judaica," The Indexer, 8(2), and in [Harrod, pp 199-209].
- Powers, Ron, (1977), The Newscasters, St. Martin's Press, New York.
- Quinn, Karen Takle (1980), "STAIRS Search Strategy: Ideas and Opinions," Online Review, 4(2), June, pp 163-168.
- Ranganathan, S. R. (1965), The Colon Classification, Rutgers University Press. 1965.
- Rochell, Carlton (1982), "Telematics--2001 AD," Library Journal, Oct., pp 1809-1815.
- * Salton, Gerard and Michael J. McGill (1983), Introduction to Modern Information Retrieval, McGraw-Hill Book Co.

Obtuse, standard IR text. Good for organizing many diverse topics under one subject, but tending to over complicate some explanations while over simplifying others.

- Sauvain, R. W. (1970), Structural Communication in a Personal Information and Retrieval System, PhD Dissertation, U of Michigan, quoted by H. D. Burton (1981), "FAMULUS Revisited: Ten Years of Personal Information Science," J of American Society for Information Science, 32(4), Nov.
- Schabas, Ann H. (1976), "Machine Searching of UK MARC on Title, Library of Congress Subject Headings, and PRECIS for Selective Dissemination of Information," in [Wellisch, pp 149-153].
- Seiler, Lauren H. and Joseph Raben (1981), "The Electronic Journal," Society, Sept./Oct., pp 76-83.
- Sgall, Petr (1976), "Perspective Paper: Linguistics," in [Walker, pp 101-126].
- Shaftesley, John M. (1964), "The 'Jewish Chronicle' Index, 1841-," The Indexer, 4(1), and in [Harrod, pp 177-186].
- Sharp, Harold S. (1964), ed., Readings in Information Retrieval, The Scarecrow press, New York and London.
- Sigel, Efrem (1980), ed., Videotext--The Coming Revolution in Home/Office Information Retrieval, Knowledge Industry Publications, Inc., White Plains, NY.
 Rambling yet profound discussion of why linguistics has not contributed to IR. Elucidates many subtleties of IR. Thorough citations.
- * Stevens, Mary Elizabeth (1970) Automatic Indexing: a State-of-the-Art Report, National Bureau of Standards Monograph 91, Washington, DC
 A fundamental survey to which all writers eventually still return. Differs from the 1965 edition only in the addition of two appendices: a 1966 (Moscow) conference paper by Stevens and a 24 page bibliography compiled in 1969. Neither edition is available in the generally comprehensive IS collection of the UT library.
- Stibic, Valado (1980), "Influence of Unlimited Ranking on Practical Online Search Strategy," Online Review, 4(3), September, pp 273-279.

- Storey, I. N. J. (1982), "The Future of Videotext Systems--a personal view," J of Information Science, 4(2-3), May, pp 111-115.
- Swanson, Don R. (1981), "Review of The Role of the Library in an Electronic Society--Proceedings of the 1979 Clinic on Library Applications of Data Processing, edited by F. Wilfrid Lancaster," Library Quarterly, 51(3), July, pp 316-319.
- Tanenbaum, Andrew S. (1981), Computer Networks, Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Time Magazine (1982), "A Terminal in Every Home?," Sept. 13, pg 65.
- Tou, Julius T. and Rafael C. Gonzalez (1974), Pattern Recognition Principles, Addison-Wesley Publishing Co., Reading, Mass., 2nd printing.
- Townley, Helen M. (1969), "Computers and Indexes," The Indexer, 6(3), also in [Harrod, pp 347-352].
- Trubkin, Loren (1980), "Migration from Print to Online Use," Online Review, 4(1), March, pp 5-12.
- U.S. News & World Report (1981), "Electronic Newspapers: Will They Be Here Soon?," May 11, pp 61-62.
- Vickery, B. C. (1968), Faceted Classification--a guide to construction and use of special schemes, Aslib, London.
- Voedisch, Bill (1983), Director of Sales, Westlaw division of West Publishing Co. Private phone conversation with one of Mr. Voedisch's assistants in St. Paul, Minnisota, April 8.
- Walker, Donald E., Hans Karlgren, Martin Kay (1977), eds., Natural Language in Information Science--Perspectives and Directions for Research, Results of a workshop on Linguistics and Information Science, Biskops-Arno, Sweeden, 3-5 May , 1976, Skriptor AB, Stockholm, Sweeden.
- Wellisch, Hans H. (1977), ed., The PRECIS Index System--principles, applications, and prospects, Proceedings of the International PRECIS Workshop, University of Maryland, Oct. 15-17, 1976, H. W. Wilson Co., New York.

Westlaw (1982), "WT Training Course," online, copyrighted,
West Publishing Co., St. Paul, Minnesota.

Wyllys, Ronald E. (1975), "Measuring Scientific Prose with
Rank-Frequency ('Zipf') Curves: a New Use for an Old
Phenomenon," Proceedings of the American Society of
Information Science, Vol 12, Information Revolution,
pp 30-31.

APPENDIX

INDEX

Included are (1) minor headings, e.g. "CompuServe" page 10, which don't appear in Table of Contents, (2) locations in text of all citations to List of References, and (3) a selected list of concepts used to help the author keep track of where things are without any attempt to be exhaustive.

- | | |
|--|---|
| <p>"</p> <p>"secondary"
publishers 38</p> <p>A</p> <p>A beginning 133</p> <p>A new measure of
relevancy 171</p> <p>A new possibility 73</p> <p>A solution 162</p> <p>Anthony 8, 20</p> <p>AT&T 43</p> <p>Austin 25</p> <p>Austin, 56, 121, 123,
124</p> <p>B</p> <p>Back to basics 138</p> <p>Bar-Hillel 130, 136, 145</p> <p>Bauer, 27</p> <p>Bernier, 128</p> <p>Boodson, 90</p> <p>Boolean operators 3</p> <p>Boolean search
strategy 55</p> <p>Boolean searches 14</p> <p>Boral 24</p> <p>Bott, 127, 153</p> <p>Boyce, 149</p> <p>Boyle, 49</p> <p>BRS 33</p> <p>Buckland, 38</p> <p>Bush 70, 76</p> <p>Bush, 45</p> | <p>Business 15</p> <p>C</p> <p>Caldwell 40</p> <p>CEEFAQ 8, 71</p> <p>Chamberlain 14</p> <p>Christie, 89</p> <p>cloister of the invisible
college 96</p> <p>Clue words 154</p> <p>Collison, 88, 90, 108,
109, 110, 112</p> <p>CompuServe 10, 42, 66,
71</p> <p>Computer, 40</p> <p>Congress 17</p> <p>Content Addressable
Memory (CAM 23</p> <p>controlled vocabulary 71</p> <p>Copyright 50</p> <p>Correcting the new
measure 175</p> <p>Cost 24, 48, 163</p> <p>Cushing, 116</p> <p>D</p> <p>database producer 2, 16,
17</p> <p>database producing 14</p> <p>Deuel, 96</p> <p>DIALOG 33, 52, 57, 59,
62, 64, 66, 69, 71</p> <p>Document descriptions as
summaries 106</p> |
|--|---|

double the amount of
storage needed 106
Doyle 27, 33, 34, 107,
133, 177
Doyle, 18, 22, 37, 45,
85, 112, 114, 115, 118,
125, 126, 128, 129, 132,
137, 138, 146, 149, 151,
179
Doyle, 1975, pg 337] 147

E

Edmundson, 137, 155
Encyclopedia 19
Eppes, 174
Experimental
verification 146

F

Faceted
classification 116
Faden 31
Fenichel 33, 37
Finichel, 41
For example 30, 40, 54,
56, 71, 78, 79, 103,
114, 120, 123
For instance 117
Fortune 10
Fortune, 42
Foster 23
French telephone
directories 20

G

Getting the right
word 56
Goodwin, 83
Gore 72
Gore, 90, 99
Gregory 18

H

Hafner 77, 78, 79, 81,
82
Hall 59
Hall, 53, 58
Hamilton 18
Harter, 42
Henkle, 55
Hiltz 62
Hogan 5, 15
Horowitz, 48, 97, 98,
101
How it would look 166

I

IBM 14, 38, 53
Indexing and
classification 110
Info Globe 2, 71
Information Bank 5, 16,
17
Inforum 5
Interactive computing 32

J

Jonker 29

K

Kahen, 90
Knoxville 42
Kodak, 45
Kucera, 136
Kutz, 47
KWIC 126
KWOC 52

L

Lancaster 72, 118
Lancaster, 50, 77, 91,
96, 110, 111, 116, 125,
131
Lawlor 14
Lawrence, 68

LEXIS 14, 15, 16, 53,
54, 71

Lockheed 26

Luhn 58, 70, 71

Luhn, 75, 135

Lyons 156

Lyons, 153

M

Markee, 95

Martin, 60

McGonigal 15

McMurdo 70

Meadow, 99

Memex 70, 71, 76, 84,
100, 173

Migration from print to
electronic
technology 50

N

Nash 3

Nash, 43, 67

Nau 78

New York Times 5, 14, 16

News 19, 33

News, 15, 20, 33

Newsletters 19

NEXIS 16, 71

note that 26, 38, 55,
72, 94, 115, 122, 133,
150, 164

O

Online 5

Other advantages of
FT 72

P

parallel processing 24

Parenthetically 132

Pattern Recognition 61

Peek-a-boo 52

Perry 83

perversity of natural
language 137

perversity of natural
language computing 129

perversity of natural
language processing 157

Posner 111

post-coordinate 34, 56

Powers, 98

PRECIS 120

predicting the
future 119

Prestel 6, 14, 20, 67,
71

Problems with a
controlled
vocabulary 118

Problems with
evaluation 147

Properties of Q-A 77

proximity operator 72

Punched cards 29

Q

Q-A is more ambitious
than FT 80

QL Systems 3, 4

Quinn, 60

R

Reasonable expectations
for current IR 82

RECON 33

Relative frequency 136

retrospective search 80

Rochell 43

Rochell, 93, 95

Route Four 88

Route One 85

Route Three 88

Route Two 87

S

Salton 59, 77, 78, 79,
 147, 152, 162, 176
 Salton, 40, 69, 74, 79,
 106, 107, 109, 129, 135,
 141, 144, 145, 150, 162,
 169
 Salton, 1983, pg
 86-7] 133
 Sauvain, 80
 Schabas, 75
 SDI 18, 80
 Search Helper 63, 69
 Searching without
 document
 descriptions 102
 Seiler, 90
 Sgall 156
 Shaftesley, 1964] 111
 Sharp, 29, 35
 Sigel 8, 10
 Sparck 24, 25, 32, 103,
 108, 128, 131, 132, 136,
 139, 148, 153, 155, 156,
 158
 Sparck Jones and Kay 81,
 163
 Sparck Jones and Kay,
 1973, pg 50] 107
 Stability of Routes 89
 STAIRS 18, 53, 58, 59
 Stemming 155
 stemming algorithm 129
 Stevens, 104, 107, 126,
 128, 130, 132, 136, 141,
 142, 149, 151, 155
 Stibic, 161
 Stop lists 154
 Storey 8
 SYS file locations
 CH1P1 1
 CH1P2 35
 CH2 68
 CH3P1 102
 CH3P2 124
 CH3P3 147
 CH4 157
 List 180

T

Tanenbalm, 43
 The abstract 108
 The larger-context
 problem 22
 The problem 159
 The problem of cost 95
 The problem of
 quantity 97
 The problems of Q-A 77
 The thesaurus 123
 This author 56, 73
 Time 20
 Tou, 141
 Townley, 121
 Trubkin, 50
 truncation 3
 Two kinds of
 indexing 130
 Two routes to FT 45
 Two types of FT
 system 71

U

U.S. 12
 U.S. News & World
 Report 10, 43
 Understanding an
 empirical approach 139
 Uniterm
 "classification" 113
 Uniterms 28
 University of
 Pittsburgh 14
 Using the right word 58
 Using this measure 173

V

vendor 2, 16
 vendors 19, 33, 91
 Vickery, 117
 Voedisch 4

W

Was Luhn right 136
 Washington Post 16

Wellisch, 122
Westlaw 4, 16, 17, 60
Wyllys, 135

5

50 million book
titles 90

VITA

This thesis was written in the house where Hal Harrison grew up. Born in 1939, he graduated from UT with a Bachelor's in Philosophy in 1961. After being drafted into the Army and serving in Kaiserslautern, Germany, he lived for 10 years in Hong Kong, working as a newspaper reporter and filmmaker. Family illness prompted a return home with his lovely Chinese wife and two children.

He gained teacher certification from Carson Newman College but then worked in local television as a news cameraman. He took several quarters of language at UT as a part time student before entering the CS graduate program in 1981. He is now moving to a job with IBM in Boca Raton, Florida, leaving again the house where he grew up in East Knoxville.