AUTOMATIC RETRIEVAL OF BIOGRAPHICAL REFERENCE BOOKS

Cherie B. WEIL: Institute for Computer Research, Committee on Information Science, University of Chicago, Chicago, Illinois

A description of one of the first projected attempts to automate a reference service, that of advising which biographical reference book to use. Two hundred and thirty-four biographical books were categorized as to type of subjects included and contents of the uniform entries they contain. A computer program which selects up to five books most likely to contain answers to biographical questions is described and its test results presented. An evaluation of the system and a discussion of ways to extend the scheme to other forms of reference work are given.

Ideally the reference librarian is the "middleman between the reader and the right book" (1), and this is what the program here described is intended to be. In the past there has been very little interest shown in automating this service, probably because it is neither urgent nor practical in current reference departments. Many developments in automating other areas of libraries have indirectly benefitted reference librarians, and the literature primarily emphasizes this aspect. For instance, where circulation systems have been automated, the location of a particular volume can be quickly ascertained and librarians need not waste time searching. Automation of the ordering phase provides them with information on the processing stage of a new volume. If the contents of the catalog have been put in machine readable form, special bibliographies can be rapidly produced in response to a particular request or as a regular service of selective dissemination. The development of KWIC (Key Word In Context) in-

dexes, which are compiled and printed by computer, has enabled publishers to provide indexes to their books much faster. Computers have also been programmed to make concordances and citation indexes (2). The combination of paper-tape typewriters, computer and a photocomposer has introduced automation into compiling *Index Medicus* (3).

Changes in reference services themselves, however, may make automation of question-answering practical. One trend is toward larger reference collections to be shared by several libraries; some areas have already set up regional reference services. There are also cooperative reference plans whereby several strong libraries agree to specialize in certain fields and cooperate in answering questions referred by the others (4). These trends will mean two things to reference librarians: greater concentration of resources, allowing more specialized books and mechanization; and screening of questions at the local level, letting reference centers concentrate on more complex questions that utilize their specialized books. Thus it seems likely that special reference centers may look increasingly toward mechanizing their services, and retrieval schemes of the type presented here will be important to consider.

BASIC ASSUMPTIONS

The categorizing system was based on two nearly universal generalizations about biographical reference books: 1) They are consistently confined to biographies of persons who have something in common: for example, being alive or dead; or having the same nationality, sex, occupation, religion, race, memberships; or possessing some combination of those attributes. These common characteristics in the people covered by a given book are herein called "exclusive categories." 2) The books generally maintain uniform entries for each subject; that is, they give the same data for each biography. These facts are referred to herein as "specifics" or "specific categories."

Certain assumptions were made about reference work: 1) All biographical reference books fit into the scheme and can be categorized. 2) The more limited a book's scope, the more likely it is to contain the person a user wants to find. In other words, if a user is interested in a Dutch economist, he is more likely to find information in a book limited to Dutch economists than in a general biographical dictionary. The user, however, does not want to miss any source that might be useful. Therefore a general biographical dictionary should be given to him as a last resort, after books on Dutch economists, Dutchmen of all occupations, and economists of all nationalities. 3) Certain requirements, the specifics, have no substitutes. For example, a book lists addresses or it does not, and if a user wants an address, books without them are useless.

There is merit in suggesting to a user which book to use as opposed to giving him the direct answer to his question. Probably the best argument for this assumption is that the volume of names that would have to be

compiled and stored for a direct inquiry system is staggering, only a small number would ever be looked up, and it is impossible to predict which ones would be searched for.

There are advantages to mechanizing this particular task of a reference librarian: good reference librarians should be freed to perform work less easily mechanized; there are not enough reference librarians who have perfect recall of their collections even to knowing which exclusive categories all the books fit into; and no librarian could have complete recall as to the specifics contained in each biographical reference book in the collection.

THE COMPUTER PROGRAM

The program was written in the COMIT language, a non-numerical programming language developed for research in mechanical translation, information retrieval and artificial intelligence. It is a high-level problem-oriented language for symbol manipulation, especially designed for dealing with strings of characters. The program could probably be converted to other list-processing languages (6) for operation at other installations.

The program was run at the University of Chicago Computation Center on an IBM 7094 having the COMIT system on a disk. Questions were submitted and run in large batches.

THE DATA

All biographical reference books in English, with alphabetical ordering of subjects, which are in the reference room of the University of Chicago's Harper Library were included in the data and no other books were included. Since one assumption was that all biographical reference books could be categorized by the scheme, it seemed more useful to prove the system could handle any biographical reference tool than to compile a balanced list of biographical books. There was no difficulty in categorizing the books.

All books are categorized in the following way. First an arbitrary abbreviation for the book is chosen to be its entry in the file; it is referred to as a "constituent." Each book is then described by determining the values of nine subscripts each constituent carries, the subscripts being SEX, LIVING, NAT (nationality), OCCUP (occupation), MIN (minorities), DATE, INDEX, SPEC1 and SPEC2 (specifics).

Values of the first five subscripts—the exclusive categories—are first determined. That is, is the book limited to one sex? Are all the subjects living or dead? Do they all have a certain occupation? Does the book include only certain nationalities? Or is there another restriction; e.g., to alumni of a college, members of the nobility or a religious group? The exclusive categories for a book are determined and coded from a table of abbreviations. SEX, for example, allows three values: restricted to males (M), restricted to females (F), or no restriction (Z). Also a value X must occur

with M or F, indicating there is a restriction. Therefore SEX can have the following combinations: SEX Z, SEX F X, or SEX M X; the values M X and F X are both the opposite of Z.

Next the book's date is determined by asking "At what date did the values on LIVING (yes or no) apply? Or, if the subjects are not restricted to living or dead (LIVING Z), "When was the book up to date?"

Next any indexes to the biographies are noted. All the biographical books list subjects in alphabetical order by surname. Lists of subjects in any other order are considered indexes even if the subjects are actually listed in some other order in the main body and the list that is alphabetic by surname is an index.

Finally, specific categories (SPEC1 and SPEC2) are coded for such facts as birthdate, birthplace, college attended, degrees held, hobbies, illustrations, social clubs, and marital status.

When all categorizing is finished, a data item is punched in this form: DICTPHILBIO/INDEX FIELD X, LIVING N X, OCCUP Z, SEX A, NAT PHILIP ASIAN X, SPEC1 D C DS FL BP L CL CM DG E I Z, DATE 50S X, SPEC2 P PL R MS PD Z, MIN Z +. This represents the *Dictionary of Philippine Biography*, a book limited to dead Filipinos and giving for each entry: dates, career, descendants, field, birthplace, long articles, class in college, degrees, education, picture, parents, publications, references, marital status and physical description. The book has a special index to find subjects by their field of work.

One specific value, that for a long article, requires special mention. Though most biographical reference books provide the same facts about all the subjects in list form, a few provide different facts about different subjects in a narrative form. Such books carry the SPEC1 L, and the other specifics these books are listed as providing are not always given for every subject. For example, a book with a list format may provide the birthplace for every subject when it can be ascertained, but in a book using the narrative form, where often different authors write the articles, birthplace is not necessarily given. Books in narrative form are used less for quick reference; therefore the program provides a note, when a long article is requested, that the card catalog may provide more long articles on the subject.

Ease of file maintenance is one advantage of this system. As data is analyzed in the first place, if a new value for a category is required, such as an occupation which is not in the list, the new value is simply added under OCCUP for that particular book and in the list of abbreviations for future use. It is a little more complicated to make an existing value more specific. For example, to differentiate BOTANIST, CHEM, PHYSICS and ASTRON and still maintain SCIENTIST as a general category embracing them all, another short program is required to retrieve the data to be reclassified.

CODING THE QUESTION

A biographical question can be quickly coded. The nine required subscripts are the same as those for the data books, but only one value for each subscript is necessary. For example, "What are the publications of a living Dutch economist? A current book is desired." is coded as Q / SEX Z (or M), LIVING Y, NAT DUTCH, OCCUP ECON, MIN Z, INDEX Z, DATE 60S, SPEC1 Z, SPEC2 PL +.

OPERATION OF THE PROGRAM

Briefly, the program reads in data and then the first question. It weeds out data items that can never be suitable, discarding all but those items that have the same values as the question has on the subscripts INDEX, SPEC1 and SPEC2. It then weeds out data items that do not have either the same values as the question, or the value Z, on the subscripts OCCUP, NAT, MIN, SEX and LIVING. After each weeding the program checks to determine that there are data items left; if all the books have been weeded out, there are no answers. There is also a provision to allow the user to designate certain titles to be ignored on a particular question in case he has already checked them, for example.

All data items left after weeding are potential answers and could simply be printed out. However, subsequent searches over the remaining items serve the purpose of rearranging them into an order in which they are more likely to produce answers. It was decided that five answers are enough to judge the types of titles chosen yet few enough to avoid very long searches. A shorter list of answers would obviously be cheaper and a longer list more likely to produce a book containing the desired subject.

Ordering proceeds as follows: first values of subscripts SEX, LIVING, MIN, OCCUP, NAT and DATE on the question as originally stated are matched to those of books in the data. The computer is at this stage searching for books that are limited in just those categories in which the question is limited. For example, if the question Q / SEX Z, LIVING Y, MIN Z, NAT DUTCH, OCCUP ECON, INDEX Z, DATE 60S, SPEC1 A, SPEC2 PL + will match only those books published in the 1960's and restricted to living Dutch economists which give publications for all the subjects (or the majority), and the books cannot be restricted to a sex or to any "minority" group. The books found may or may not have additional values on the subscripts; that is, a book may also contain French economists. Such books found on the first search are mostly likely to contain the subject the questioner is looking for.

If there are fewer than five books found which are a perfect match with the question, the program begins to alter the question. To make the least significant possible change in the question, the program changes the value of the subscript judged to be the limiting factor on the fewest books in the data, namely sex. If SEX has a Z as its value (because the questioner did not know the sex or did not prefer a book limited to one

sex) it is changed to X so that a book limited to one sex will not be overlooked. If SEX does not have a Z value (which means it has either M X or F X), it is changed to Z. This means the questioner preferred books limited to one sex but presumably his second choice is books not limited to any sex. Clearly if the question has SEX F X it can never be changed to SEX M X or SEX X, since SEX X will find books in the data classified SEX M X. Anything other than Z changes to Z, and Z only changes to X. After this change is made, another search is conducted and the answers counted. Until there are five books or the data is exhausted, the original question is altered and the cycle continued. Alterations proceed by changing the values of one subscript at a time in the following order: SEX, LIVING, MIN, NAT and OCCUP. Then they are changed two at a time, three at a time, four at a time, and finally all five are changed, so there are thirty-one possible changes.

If at the end of the thirty-second search there are still not five answers and there are more data items, the date restriction on the question is checked. If DATE has a value other than Z, it is changed to Z, which matches all the data items, and the computer prints a note if this is done; the program will then select any book regardless of date. Control returns to search and begins the cycle again, continuing until five answers are found or the data is exhausted.

After searching is finished, the writing routine commences. One at a time the computer takes each answer, writes out its code for possible further reference, and then writes out the complete author, title, copyright date and Library of Congress call number, all of which the computer finds in a list within the program.

RESULTS

To obtain some measure of the program's accuracy, fourteen textbook questions, probably more challenging than the average patron would ask, were submitted to the computer and to a professional librarian who was especially familiar with biographical reference books. (See Figure 1 for sample questions and results.)

The librarian spent a total of an hour and a half, and found answers to eleven out of fourteen questions. On the three she could not answer she felt she had exhausted the resources. In one of the eleven she answered ("How many Americans won the Nobel Prize in medicine between 1930 and 1950?") she found the answer in a source not specifically biographical (*World Almanac*) and therefore not in the computer's data.

No problems occurred in forming the questions for submission to the computer. The program found some reasonable sources in all cases. It found books containing the answer in ten out of fourteen cases, the four answers not found being those three the librarian missed and the one requiring an almanac. In all but one case there were more possibilities than the five books given per answer. Some questions were rerun ignoring

Question: In one source find a list of at least twenty references to biographical information about Dmitri Mendeleef (or Mendelev), Russian chemist (1834-1907).

As submitted to computer: Q / SEX M, LIVING N, OCCUP CHEM, NAT RUSSIAN, MIN Z, SPEC1 Z, SPEC2 R, INDEX Z, DATE Z +

Librarian's results:

B	Phillips, Dictionary of Biographical Reference	- 0 references
A	Encyclopedia Britannica	- 6 references
A	Encyclopedia Americana	- 1 reference
A	Biography Index (1949-64 volumes)	- 14 references

time: 10 minutes

Computer's results:

A	Index to Scientists	- 27 references
A	Biography Index	
C	Drake, Dictionary of American Biography (sounds wrong but it is international)	
B	Phillips, Dictionary of Biographical Reference	
A	Encyclopedia Britannica	

Question: What academic degrees have been earned by Professor Reuben L. Hill, Director of Family Study at the University of Minnesota?

As submitted to computer:

(1) Q/ SEX M, LIVING Y, OCCUP EDUC, NAT AMER, MIN Z, SPEC1 DG, SPEC2 Z, INDEX Z, DATE Z +
(2) IGNORE + AMECONASSN +
IGNORE + AMERSCIENCE +
IGNORE + AMPOLISCI +
IGNORE + DAMERSCHOL +
IGNORE + LEADEDUC +
Q / SEX M, LIVING Y, OCCUP EDUC, NAT AMER, MIN Z, SPEC1 DG, SPEC2 Z, INDEX Z, DATE Z

Librarian's results:

B	Leaders in Education	
A	Who's Who in America	- Answer: BS, PhM, PhD

time: 3 minutes

Computer's results:

(1) D	Handbook of the American Economic Association	
D	American Men of Science	
D	Biographical Directory of the American Political Science Assocation	
D	Directory of American Scholars	
B	Leaders in Education	
(2) B	Who's Who in American Education	
C	Outstanding Young Men of America	
A	Who's Who in America	
B	Who's Who in various areas	
B	National Cyclopedia of American Biography	

Question: Where might I find information about a New England ancestor named Jacob Billings who was born around 1753?

As submitted to the computer: A / SEX M, LIVING N, OCCUP Z, NAT AMER, MIN FF, INDEX Z, DATE Z, SPEC1 Z, SPEC2 Z +

Librarian's results:

D	Handbook of Genealogy - about genealogists not families
A	Compendium of American Genealogy

time: 8 minutes

Computer's results:

A	Compendium of American Genealogy
C	Dictionary of American Biography
C	Who Was Who in America
C	Lamb's Biographical Dictionary of the U. S.
C	Concise Dictionary of American Biography

A = It has the answer or at least part of it
B = Good choice but it does not have answer
C = Reasonable choice but there are better ones
D = Poor choice

Fig. 1. Sample Reference Questions.

the first five answers, and five more titles were retrieved; even then there were more possibilities.

In some cases the program did better than the librarian because she wasted time looking in sources that did not give the specifics sought. For instance, when the question asked for the pronunciation of the surname of Paul and Lucian Hillemaker, French composers, she looked in dictionaries that do not give pronunciation. The computer found the only four possible sources immediately.

In other cases the program came up with rather far-fetched answers a human would have skipped. A question asking for biographies of Franz Rakoczy, an Hungarian hero, retrieved in its second five sources three Jewish encyclopedias and a book on composers! These were not wrong and, in cases where occupation or minority group affiliations were unknown, these might be good sources.

As an answer to the Nobel-prize-winner question the computer retrieved sources on American doctors, Nobel winners and scientists, which are the best choices from the data and would have the answers buried in them. However, what is really required is an index to award winners, and there were none in the data.

The test revealed the necessity for allowing questions to have dummy values; that is, ones not used in the data. For instance there are no books limited to botanists, so OCCUP BOTANIST is not allowed in a question, though OCCUP SCIENTIST is, and CHEM and PHYSICS are included as more specific values under SCIENTIST. Asking for OCCUP SCIENTIST when searching for a botanist avoids getting books devoted to non-scientific occupations but also gets books devoted to chemists and physicists. Since one would want these books if he did not know the scientist was a botanist, that should not be changed. If he asks for OCCUP BOTANIST he wants books devoted to botanists first, then scientists in general.

A short-term solution is to have dummy values to stand for all these other values. For example OCCUP OTHER-SCIENTIST could include all scientific occupations except those specifically listed, and it would retrieve books limited to all scientists but not to specific scientific occupations mentioned in the data. A long-term solution is to use a computer language allowing tree-structured data. Presently this problem does no more than cause extraneous retrievals which the person using the list can easily skip.

DISCUSSION

Advantages of the scheme can be speculated. From the library's point of view its virtues are that it is simple and inexpensive. Original implementation would not require a major block of time to be spent in human indexing or abstracting. Operating costs would be low because it does not require such a large store of information in memory that several tapes must be searched, and because updating the file is simple. When a new

book is added, an experienced person could categorize it in five minutes, punch a new data card and, if required, add to the list of values in the table of abbreviations.

The system could provide useful information to other departments. It could keep tallies for the acquisitions department of how often a book is given as an answer, indicating whether new editions of it or similar books would be good buys.

From the user's point of view the system avoids a major pitfall of some retrieval schemes which retrieve on the basis of ambiguous terms or association chains; that is, missing relevant items. If the user resubmits the same question ignoring already retrieved books each time, he will eventually have a comprehensive list of possible sources in the data that have the index and specifics he requires. A user also wants his information as brief as possible, listed in order of importance and with no extraneous answers (7); this requirement could be met as the program stands by having a human simply cross out any unnecessary titles. Users like to know the reliability of the information (7); this detail could be provided along with the titles.

Users also want speed and convenience. As it stands, this system could be made available to users of the University of Chicago Library tomorrow with no more equipment than is presently in the Computation Center. Time delay in the present implementation could be remedied by using an on-line system.

Users often prefer to be given facts themselves and not just citations (7). A program that gives biographical facts directly has no connection with this scheme or classification system, but the output of this program could be used as a tool by a librarian to find the answer for a patron.

BIBLIOGRAPHIES

The most obvious area to which the retrieval scheme could be extended is that of bibliographies. Like biographies, they are limited in their scopes to certain exclusive categories, and they contain the same specific facts for each entry. Logical exclusive categories could be: NATIONALITY, FORM (with such values as drama, poetry, fiction, maps, etc.), SUBJECT (probably the most frequently used criterion on which to select books for a bibliography), and DATE. Since there is no LIVING with which to connect DATE, DATE here should probably have not just the most recent relevant date but as many values as necessary. For instance DATE 40S 50S 60S would apply to an index that began publication in the 1940's and is current. Then a request for any of those dates would find it.

Possible SPECIFICS include number of pages, the cost, or a facsimile of the title page. ARRANGEMENT would be needed, being different from INDEX in that bibliographies, unlike biographies, cannot be assumed to have the same order (alphabetic by subject's name) plus indexes in other orders. ARRANGEMENT would list as values all the ways the con-

tents of the bibliography could be approached: by subject, author, title, chronology or a combination of these.

DICTIONARIES

Dictionaries also lend themselves well to this type of scheme; one exclusive category, SUBJECT, might even be adequate for dictionaries. Dictionaries' special subjects could be broken down into FIELD (such as chemistry or business) and TYPE (such as slang or geography), if necessary. LANGUAGE would be a specific category, since there are no substitutes for the language required. Other possible SPECIFICS are pronunciation, definition, etymology and illustration.

ATLASES

Atlases are also suited to the scheme. Exclusive categories that seem appropriate are AREA covered, special SUBJECT atlases, and the size of the SCALE. SCALE should probably act as DATE does in the biographical program; that is, if a particular scale is requested, that would be searched for first and, if no answer is found, a note would be given and another search made for any scale. SPECIFICS for atlases could include items like topography, rainfall, winds, cities, highways and major products.

Factual books (those that give the highest mountain, the first four-minute mile, the January 10th price of U. S. Steel, etc.) do not lend themselves to the scheme. Because these books are not uniform as to entries and subject coverage, the list of possible specifics and exclusive categories would be extremely long and the number of searches consequently prohibitive. Also, since such books are far fewer in number than biographical or bibliographical works, the proper one is easier to find by browsing.

CONCLUSION

A scheme for categorizing biographical reference books by their exclusive and specific categories makes it possible to automatically retrieve titles of those which would best answer reference questions. When tested it was found acceptable, with minor refinements, and it is easily adaptable to other reference book forms. Such a system seems a logical direction in which to go when automation of actual reference functions is undertaken.

ACKNOWLEDGMENT

The project under discussion was undertaken in partial fulfillment of requirements for the M. A. degree at the University of Chicago's Graduate Library School. The computer program employed is detailed in the author's thesis (8). The work was partially completed under the auspices of AEC Contract No. AT(11-1)614.

REFERENCES

1. University of Illinois Library School: *The Library as a Community Information Center.* Papers presented at an Institute conducted by the University of Illinois Library School September 29-October 2, 1957 (Champaign, Illinois: University of Illinois Library School, 1959), p. 2.
2. Shera, Jesse: "Automation and the Reference Librarian," *RQ*, III, 6 (July 1964), 3-4.
3. Austin, Charles J.: *Medlars 1963-1967* (Bethesda, National Institutes of Health, 1968).
4. Haas, Warren J.: "Statewide and Regional Reference Service," *Library Trends.* XII, 3 (January 1964), 407-10.
5. Yngve, Victor: *COMIT Programmers' Reference Manual* (Cambridge, Mass.: M. I. T. Press, 1962).
6. Hsu, R. W.: *Characteristics of Four List-Processing Languages* (U. S. Department of Commerce, National Bureau of Standards, Sept. 1963).
7. Goodwin, Harry B.: "Some Thoughts on Improved Technical Information Service," *Readings in Information Retrieval* (New York, Scarecrow Press, 1964), p. 43.
8. Weil, Cherie B.: *Classification and Automatic Retrieval of Biographical Reference Books* (Chicago: University of Chicago Graduate Library School, 1967).