Janda, Kenneth, ***TRIAL: A Computer Technique for Retrieving Information from Abstracts of Literature*** , Behavioral Science, 11:6 (1966:Nov.) p.480

COMPUTERS IN BEHAVIORAL SCIENCE

Publication of this department is partially supported by a grant from the National Science Foundation to its editor, Steven G. Vandenberg.

Information retrieval has been discussed in this journal by Stone and others in October, 1962 and October, 1965; by Olsti in October, 1964; by Iker and Harway in April, 1965. Here is another description of a working system: a trial of TRIAL.

TRIAL: A COMPUTER TECHNIQUE FOR RETRIEVING INFORMATION FROM ABSTRACTS OF LITERATURE

by Kenneth Janda and William H. Tetzlaff

Northwestern University

TRIAL is an acronym formed from "Technique for Retrieving Information from Abstracts of Literature." The TRIAL system consists of two basic computer programs written in MAP programming language for the IBM 709/7090/7094 computers operating under the IBSYS monitor (Tetzlaff and Janda, 1966). (At the time of this writing, the programs are being revised and rewritten for the CDC 3400.) The EDIT program in TRIAL loads and edits abstracts on magnetic tape; the SEARCH program searches magnetic tape and retrieves information according to specified combinations of keywords. Although TRIAL was originally designed to retrieve research findings and propositions abstracted from books and articles in political science (Milbrath and Janda, 1964), the system is sufficiently flexible for other applications in information retrieval.

PREPARATION OF INPUT

TRIAL was originally devised to cope with problems encountered in collecting, using, and managing several hundred propositions on political participation, culled from literature and recorded on 5 x 8 index cards. (See Milbrath, 1965, for the result of this inventory.) The sheer size of this inventory hampered attempts at exploitation. A method was needed to find specific propositions of interest and to reproduce or otherwise make them available for research. A further requirement was to present these propositions in the *context* of the original study, enabling the researcher to evaluate information he had retrieved. A system was designed to achieve these objectives through a combination of (1) computer programming and (2) careful preparation of the abstracts serving as input to the computer. The preparation of computer input will be discussed first.

Each TRIAL abstract is organized into several parts, each part being recorded on cards punched with corresponding "class" numbers. Authors are recorded on class 1 cards; titles on class 2 cards; and sources (publications) on class 3 cards. Class 4 cards carry a summary that briefly describes the study in terms of the following headings: the problem, research design, conclusions, and suggestions for research. Every proposition or finding in the abstract is represented by both a statement and an elaboration (punched on class 7 and 8 cards respectively). The "statement" rephrases the proposition in basic key words intended to facilitate search and retrieval operations. The "elaboration" quotes original wording and provides additional context for understanding and evaluating the proposition. The elaboration might include operational or conceptual definitions and empirical evidence supporting the statement.

Janda, Kenneth, ***TRIAL: A Computer Technique for Retrieving Information from Abstracts of Literature*** , Behavioral Science, 11:6 (1966:Nov.) p.480

The abstracts are organized in this manner to retrieve information in the context of the original study. Standard TRIAL operating procedure is to instruct the computer to search for specified combinations of keywords occurring in "statements" contained in hundreds of abstracts recorded on tape. For every statement that satisfies the search, the computer prints out (1) the complete citation for the abstract, (2) a summary of the study, (3) the statement itself, and (4) its accompanying elaboration. This information provides a context within which the researcher can evaluate the statements turned up in his search.

Several steps are involved in abstracting articles for input to the TRIAL system. The original material is first read through in order to identify propositions and to take notes for the summary. The article is then reread and the abstract typed on erasable paper with typewriter margins adjusted for a 60-space line. This corresponds exactly to the TRIAL input format, which uses 60 columns of the card. Words at the end of a line are not hyphenated and agreed upon substitutions are made for characters not available on computer printers. This procedure produces typed copy which keypunch operators can reproduce exactly—one punchcard per line.

The text of the abstract is punched in the first 60 columns of the cards. "Class" numbers are punched in column 75. As indicated below, the program uses the class numbers to tell the computer what to search—authors, titles, sources, summaries, statements, or elaborations.

The keypunch operator also assigns a number i in columns 73–74 to every card in the ith statement and to every card in its accompanying elaboration. This results in identifying all cards relevant to the first proposition with 01 punched in columns 73–74, all cards for the second proposition with 02 in columns 73–74, and so on. After all the cards for an abstract have been punched and corrected for errors, they are numbered sequentially in four columns (76–79) beginning with 0001. Column 80 is punched 0. In effect, this produces a sequencing by *tens* instead of *units*, which facilitates the editing of abstracts once they have been read from punchcards onto magnetic tape. Sample input to the TRIAL system is shown in Figure 1, which reproduces a tabulator printout of some propositions and findings abstracted from Campbell (1962).

COMPUTER PROCESSING

Magnetic tape is used instead of punchcards for the search and retrieval operations because it is a more efficient and convenient input medium for computers. Once the information has been put on tape, however, it is removed from direct observation and manipulation. While revisions in punchcard data are made simply by altering individual cards, revisions cannot be made so directly when information is represented by magnetized spots on a plastic ribbon.

The EDIT program in TRIAL provides the necessary flexibility in correcting and updating abstracts after they have been put on tape. This program permits inserting, replacing, or deleting whole abstracts or individual lines without reference to the original cards. The program refers to the information on tape by means of the sequence numbers punched in columns 76–80. Because the cards are numbered in tens instead of units when read onto tape, as many as nine new lines may be inserted between cards 00040 and 00050 without disturbing their sequence on tape. EDIT inserts new lines from new cards and automatically shifts abstracts on the tape to accommodate changes.

A search of the tape containing abstracts is initiated when a card identifying the search is placed after the SEARCH program. The information on this card labels the printout for identification purposes. If a search were to be made for propositions about the effect of education on likelihood of voting, the card might be punched, "SEARCH FOR EDUCATION AND POLITICAL PARTICIPATION." All instruction cards in TRIAL are virtually free of format restrictions, and this label might begin anywhere on the card. The program simply instructs the machine to begin scanning in column 1 for a nonblank column and to regard the contents of that column and the following 58 columns as

Janda, Kenneth, ***TRIAL: A Computer Technique for Retrieving Information from Abstracts of Literature*** , Behavioral Science, 11:6 (1966:Nov.) p.480


```
CAMPBELL A                                                              100010
THE PASSIVE CITIZEN.                                                    200020
ACTA SOCIOLOGICA, 6 (1962) 9-21.                                        300030
     PROBLEM-- INTERESTED IN THE PSYCHOLOGICAL FACTORS                  400040
INVOLVED IN PASSIVE OR ACTIVE POLITICAL BEHAVIOR.  ''THE                400050
BULK OF THIS PAPER IS DEVOTED TO A CONSIDERATION OF THOSE               400060
CHARACTERISTICS OF THE PERSON AND HIS ENVIRONMENT WHICH MAY             400070
BE THOUGHT TO UNDERLIE POLITICAL PASSIVITY.'' (P.10)                    400080
CONSIDERS PERSONALITY TRAITS, BASIC PREDISPOSITIONS, AND                400090
SHORT-TERM ATTITUDES.                                                   400100
     RESEARCH DESIGN-- NO SPECIFIC RESEARCH DONE FOR THIS               400110
PARTICULAR ANALYSIS.  CITES MANY DIFFERENT ARTICLES AND                 400120
BOOKS ON THE SUBJECT OF POLITICAL BEHAVIOR.                             400130
     CONCLUSIONS-- TOO LITTLE IS KNOWN OF THE BASIC                     400140
PERSONALITY NEEDS OF THE ELECTORATE FOR THE POLITICAL                   400150
SCIENTIST TO MAKE MUCH PROGRESS IN THIS AREA.  THE MOST                 400160
IMPORTANT BASIC PREDISPOSITIONS ARE SOCIAL DETACHMENT AND               400170
POLITICAL ALIENATION.  SHORT-TERM ATTITUDES ARE CONSTANTLY              400180
CHANGING.                                                               400190
''...THE RELATIONSHIP BETWEEN OUR MEASURE OF PERSONAL                01700200
EFFECTIVENESS AND THE PERSON'S EXPRESSION OF DEGREE OF               01700210
INVOLVEMENT IN POLITICS IS MUCH HIGHER AMONG PEOPLE OF               01700220
LIMITED FORMAL EDUCATION THAN IT IS AMONG THOSE OF COLLEGE           01700230
TRAINING.'' (P.12)                                                   01700240
     CITES ARTICLE BY E. DOUVAN AND A.M. WALKER, THE                 01800250
SENSE OF EFFECTIVENESS IN PUBLIC AFFAIRS, =PSYCHOLOGICAL             01800260
MONOGRAPHS=, 70, NO. 22, 1956-- 1-19.                                01800270
''...POLITICAL DETACHMENT AND IRREGULARITY ARE                       02700280
DISPROPORTIONATELY HEAVY AMONG INDIVIDUALS AND GROUPS                02700290
ISOLATED FROM THE LARGER SOCIETY.'' (P.13)  (ANOMIA)                 02700300
     CITES WORK OF W. KORNHAUSER, =THE POLITICS OF MASS              02800310
SOCIETY=.  GLENCOE, ILLINOIS-- FREE PRESS, 1959.                     02800320
     ''...IN URBAN POPULATIONS THE DEGREE OF POLITICAL               02800330
INVOLVEMENT IS HIGHLY RELATED TO THE ACTUAL NUMBER OF GROUP          02800340
MEMBERSHIPS WHICH A PERSON REPORTS.'' (P.13)  ALSO FOUND             02800350
VERY LOW LEVELS OF POLITICAL INVOLVEMENT AMONG U.S. FARMER.          02800360
ALIENATION SCALE INDICATES THAT ALIENATION ''.../IS/ ABOUT           03700370
AS FREQUENT IN ONE RELIGIOUS, RACIAL, OR ECONOMIC GROUP              03700380
AS ANOTHER.'' (P.14)  HOWEVER, ALIENATION DOES RELATE TO             03700390
POLITICAL INTEREST AND INVOLVEMENT,, THOSE WHO ARE MOST              03700400
INACTIVE ARE FREQUENTLY ALIENATED.                                   03700410
     CITES D.E. STOKES, POPULAR EVALUATIONS OF GOVERNMENT--          03800420
AN EMPIRICAL ASSESSMENT IN HARLAN CLEVELAND AND HAROLD D.            03800430
LASSWELL (EDS.),  =ETHICS AND BIGNESS=, NEW YORK, HARPER,            03800440
1962.                                                                03800450
''IN THE UNITED STATES WHERE NEITHER CLASS NOR RELIGION IS           04700460
THE BASIC DIMENSION OF DIFFERENCE BETWEEN THE TWO PARTIES,           04700470
STRENGTH OF IDENTIFICATION WITH THE PARTIES THEMSELVES IS            04700480
ASSOCIATED WITH POLITICAL INTEREST, THE MOST STRONGLY                04700490
PARTISAN INDIVIDUALS BEING THE MOST INVOLVED AND ACTIVE              04700500
PARTICIPANTS.'' (P.15)                                               04700510
THE INTEREST GENERATED BY AN ELECTION WILL DEPEND ON HOW             05700520
LARGE THE DISCREPANCY IS BETWEEN WHAT THE ELECTORATE WANTS           05700530
```

FIG. 1.

an identifying label. The machine considers the identification label terminated when it encounters a comma. Information that extends beyond a comma, or is more than 59 characters in length, is ignored by the machine.

The comma serves not only to end the identification label but to instruct the computer to look for the card classes to be searched. Assume, for example, that a search is to be made for findings or propositions relevant to the relationship between "education" and "political participation." Although the program permits searching any or all classes of input cards, the search in this example probably would be made only within the "statement" cards, class 7. The computer would be instructed to ignore the author, title, source, summary, and elaboration cards and to search the statement cards only.

The researcher must specify the keywords

Janda, Kenneth, ***TRIAL: A Computer Technique for Retrieving Information from Abstracts of Literature*** , Behavioral Science, 11:6 (1966:Nov.) p.480

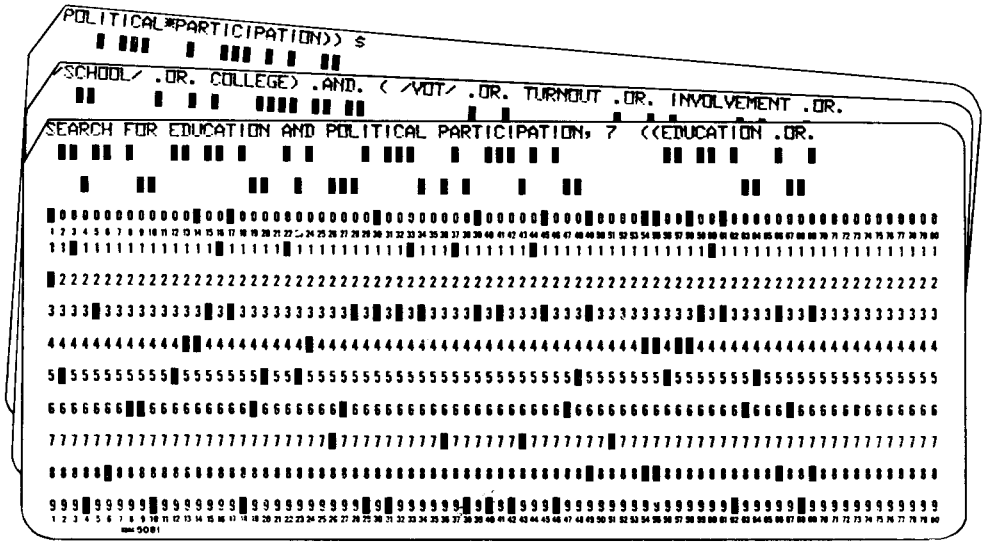

FIG. 2.

he wants the computer to use in its search. In the example above, the words "education" and "political participation" would certainly be used, but the search might also include "college," "school," "schooling," "vote," "voting," "turnout," and "involvement." All these possibilities can be incorporated in a TRIAL search by means of special word commands and logical operators.

The search command is communicated to the computer by specifying within parentheses certain keywords and logical connections that must exist between the keywords for a statement to qualify for retrieval. After reading the code number or numbers of the card classes to be searched, the machine senses that it is receiving a search command when it reads a left parenthesis. Because any given command may contain more than one combination of keywords arranged in "nests" of parentheses, the machine continues to read until the number of right parentheses read equals the number of left ones. It will then evaluate the logic of the command, working outward from the innermost set of parentheses.

Merely enclosing a word within parentheses will involve that word in the searching process. Enclosing any word between slashes (for example, "/SCHOOL/") will define the first six characters of that word as the root word and will cause the machine to retrieve any word containing the same combination of characters. The command "/SCHOOL/" therefore will retrieve "schooling" as well as "school." Joining any pair of words with an asterisk (such as, "POLITICAL*PARTICIPATION") will define the words as a phrase and will cause the machine to search for exactly the same phrase.

The real power of the search command lies in the use of standard logical operators expressed on the punchcard as follows: .NOT., .OR., and .AND. Perhaps the use of these operators can best be conveyed by constructing a sample command for the education and voting turnout search. Figure 2 shows the command punched on cards.

This command will cause the machine to select only those statements containing *both* one or more keywords specified within the first nest of parentheses *and* one or more keywords specified within the second nest. If the machine finds any statement of a proposition or finding which satisfies this logical combination of keywords, it will print out the statement and its supporting elaboration after printing the citation of the article and a summary of the study. A portion of the computer output produced in response to the above command is reproduced in Figure 3. The dollar sign after the last right parenthesis on the punchcards in Figure 2 starts the machine looking for

Janda, Kenneth, ***TRIAL: A Computer Technique for Retrieving Information from Abstracts of Literature*** , Behavioral Science, 11:6 (1966:Nov.) p.480


```
NORTHWESTERN UNIVERSITY TRIAL SEARCH     08/24/64     PAGE       27


CAMPBELL A

 THE PASSIVE CITIZEN.

 ACTA SOCIOLOGICA, 6 (1962) 9-21.

        PROBLEM-- INTERESTED IN THE PSYCHOLOGICAL FACTORS
    INVOLVED IN PASSIVE OR ACTIVE POLITICAL BEHAVIOR.  ''THE
    BULK OF THIS PAPER IS DEVOTED TO A CONSIDERATION OF THOSE
    CHARACTERISTICS OF THE PERSON AND HIS ENVIRONMENT WHICH MAY
    BE THOUGHT TO UNDERLIE POLITICAL PASSIVITY.'' (P.10)
    CONSIDERS PERSONALITY TRAITS, BASIC PREDISPOSITIONS, AND
    SHORT-TERM ATTITUDES.
        RESEARCH DESIGN-- NO SPECIFIC RESEARCH DONE FOR THIS
    PARTICULAR ANALYSIS.  CITES MANY DIFFERENT ARTICLES AND
    BOOKS ON THE SUBJECT OF POLITICAL BEHAVIOR.
        CONCLUSIONS-- TOO LITTLE IS KNOWN OF THE BASIC
    PERSONALITY NEEDS OF THE ELECTORATE FOR THE POLITICAL
    SCIENTIST TO MAKE MUCH PROGRESS IN THIS AREA.  THE MOST
    IMPORTANT BASIC PREDISPOSITIONS ARE SOCIAL DETACHMENT AND
    POLITICAL ALIENATION.  SHORT-TERM ATTITUDES ARE CONSTANTLY
    CHANGING.


STATEMENT OF PROPOSITION...

  ''...THE RELATIONSHIP BETWEEN OUR MEASURE OF PERSONAL
  EFFECTIVENESS AND THE PERSON'S EXPRESSION OF DEGREE OF
  INVOLVEMENT IN POLITICS IS MUCH HIGHER AMONG PEOPLE OF
  LIMITED FORMAL EDUCATION THAN IT IS AMONG THOSE OF COLLEGE
  TRAINING.'' (P.12)

          CITES ARTICLE BY E. DOUVAN AND A.M. WALKER, THE
      SENSE OF EFFECTIVENESS IN PUBLIC AFFAIRS, =PSYCHOLOGICAL
      MONOGRAPHS=, 70, NO. 22, 1956-- 1-19.


STATEMENT OF PROPOSITION...

  ''PERHAPS THE SUREST SINGLE PREDICTOR OF POLITICAL
  INVOLVEMENT IS NUMBER OF YEARS OF FORMAL EDUCATION.''
  (P.20)

          ''EDUCATIONAL LEVELS ARE RISING AND WILL RISE MUCH
      FURTHER.  IN TIME THIS TREND WILL CHANGE THE CHARACTER OF
      THE ELECTORATE, INCREASING ITS CAPACITY TO COMPREHEND
      POLITICAL AFFAIRS AND ITS CONCERN WITH THEM.'' (P.21)
```

FIG. 3.

another identification label for the next search command. Many different searches can be made during one run on the computer.

COMPUTER PROGRAMS

The EDIT program in TRIAL is used to load abstracts from punched cards onto a master tape, called the new master tape. EDIT is also used to correct and update a previously created master tape. When such revisions are made, the existing master tape is referred to as the old master tape, and the new master is the one produced as a result of the changes.

Each deck of cards containing an abstract to be loaded on tape with the EDIT program is preceded by a card punched with $ABSTRACT in columns 1 through 9, and an abstract-identification number in columns 13 through 18. Each deck is followed by a card punched $END in columns 1 through 4. A series of abstracts to be loaded on a

tape, then, will consist of decks of punched cards sandwiched between $ABSTRACT and $END cards.

Five types of editing operations are available for the creation of a new master tape. Control cards indicating the operation to be used contain a dollar sign in column 1 and the name of the operation beginning in column 2. Where applicable, the identifying number of an abstract is punched beginning in column 13. The available editing operations are:

$INSERT—causes one or more abstracts following it to be written immediately onto the new master tape.

$AFTER—enters new abstracts from punchcards immediately after the abstract on the old master tape that is specified in column 13.

$REPLACE—an abstract on the old master tape specified in column 13 will be replaced by an abstract on cards.

$DELETE—delete the abstract specified in column 13 from the old master tape.

$ALTER—makes changes or "alters" individual lines within the body of a specified abstract. Alterations of individual lines within an abstract are referenced by the sequence numbers of the original punchcards.

The SEARCH program in TRIAL is used to retrieve information contained in the abstracts stored on magnetic tape. Data to the program consists of one or more search "commands" punched on as many cards as needed, in columns 1 through 80. Blank columns are ignored by the program and may be used to improve readability. The search command instructs the computer (1) what identifying information is to be printed as a "heading" at the top of each output page, (2) what classes of cards are to be searched, and (3) what keywords and logical relations among keywords will satisfy the search.

The SEARCH program will accept up to 50 search commands on one run. These commands must be separated from each other by dollar signs. Output from separate searches will print out one after another with page numbering restored to 1 for each search command. Less computer time is used in making several searches simultaneously than in making single searches on individual runs.

The operation of the search program is divided into three distinct parts:

First, the program reads and saves in memory all of the search instructions. As they are read they are checked for correctness, and error messages will be printed to indicate mistakes. If any errors are found, the program will terminate at the end of this phase.

Next is the test for the first search instruction. This phase utilizes the 709's ability to read, write, and compute simultaneously. The program uses one data channel to read ahead in the abstract master file. The program simultaneously scans abstracts to find those which satisfy the search instructions. While the first two operations are taking place, the second data channel is being used to write on the standard output tape those abstracts and parts of abstracts that satisfy the first search. The second data channel is also used to write onto an intermediate tape those abstracts and parts of abstracts that will satisfy subsequent searches. This tape will be read during phase three.

The intermediate tape prepared in phase two will expedite subsequent searches, since it contains only information that is actually needed—and thus is a fraction of the size of the abstract master file. The intermediate tape will be searched and rewound until all of the searches are done.

EVALUATION

The TRIAL system has been tested out in a preliminary fashion with literature on political participation. The computer routines have performed with complete satisfaction in handling input to the system. The important unanswered questions about the TRIAL system lie not in programming procedures but in preparing the abstracts for the computer.

Abstracting research literature is a time-consuming enterprise that requires a well-defined conceptual framework geared to a specific research problem. The abstracts

that were prepared to test out the computer programs are not suitable for a broader test of the system's utility. A test of using a computer to retrieve information from abstracts prepared for a definite research application is now in progress for literature on comparative political parties (Janda, 1964). A complete evaluation of this approach will be published at the conclusion of the research.

A listing and description of the program has been deposited as Document Number 8842 with the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C.

Advance payment of $41.25 for photoprints, or $11.00 for 35mm microfilm, should be made in ordering from Chief, Photoduplication Service, Library of Congress.

REFERENCES

Campbell, A. The passive citizen. *Acta Sociologica*, 1962, 6, 9–21.

Janda, K., A methodological approach to the comparative study of political parties. Paper delivered at the Comparative Politics Seminar, The University of Michigan, November, 1964.

Milbrath, L. *Political participation*. Chicago: Rand McNally, 1965.

Milbrath, L. & Janda, K. Computer applications to abstraction, storage, and recovery of propositions from political science literature. Paper delivered at the Annual Meeting of the American Political Science Association, Chicago, 1964.

Tetzlaff, W., & Janda, K. The TRIAL information retrieval system for the IBM 709/90/94 *Behav. Sci.* 1966, 11, 407.

The method of the physical sciences is based upon the induction which leads us to expect the recurrence of a phenomenon when the circumstances which give rise to it are repeated. If all the circumstances could be simultaneously reproduced, this principle could be fearlessly applied; but this never happens; some of the circumstances will always be missing. Are we absolutely certain that they are unimportant? Evidently not! It may be probable, but it cannot be rigorously certain. Hence the importance of the role that is played in the physical sciences by the law of probability.

POINCARÉ