Matching Inquiries to an Index

By M. A. Wright

A method is given for measuring degree of matching between inquiries and index entries which allows for the effect of mistakes and of omissions of characteristics.

Introduction

An index contains a set of entries each of which in turn contains a set of descriptions. In this paper entries in an index are referred to as *records* and descriptions as *characteristics*. The characteristics used in any index can be classified into several groups, e.g. *a*, *b*, *c*, etc. A record (y) may thus contain

$$y_a, y_b, y_c, y_d, \text{ etc.}$$

where y_a is the characteristic of type *a* which applies to record y. For example, in a telephone directory the characteristic type *a* may refer to surnames and y_{an} to a particular surname. A typical record y is usually incomplete and may contain, for example,

$$y_a, y_b, y_d, y_f$$

information appertaining to y_c and y_e not being available.

Inquiries to an index can be described, similarly, in terms of characteristics, e.g. an inquiry (x) may contain

$$x_a, x_b, x_c, x_d, \text{ etc.,}$$

although a typical inquiry is incomplete and may contain only

$$x_a, x_c, x_d.$$

If all the information in an index is complete and correct and an inquiry is completely and correctly quoted, the problem of finding the record which matches an inquiry is reduced to finding the record or records in which

$$y_a \equiv x_a,\ y_b \equiv x_b,\ y_c \equiv x_c, \text{ etc.}$$

However, in practice, neither index nor inquiries are complete and both contain errors; this paper deals with a method of measuring the degree of match between an inquiry and records so that a decision can be taken to determine which record provides the best match to an inquiry, and also to determine whether such a record matches the inquiry sufficiently well such that it is likely to be the one required. It is usually practical to compare (i.e. attempt to match) records with an inquiry in only a simple way dependent on the sequence of records or sequence of characteristics, and this paper also gives an example of an elementary use of heuristics as described earlier (Wright, 1960). This technique allows a change of sequence to be made such that a wider range of search and comparison is practical.

Glantz (1957) has described a method of measuring the match between an inquiry and a record which involves a transformation of input (i.e. inquiries), and a "recognition operation." The transform may be a simple one to one type or complex, amounting even to the use of a classification system. The recognition operation includes comparison, evaluation of the results of comparison to form a score, and a decision as to whether or not the record is selected. The total score is calculated from the results (subscores) of the comparison of individual characteristics and is formulated as

$$\sum_{n=1}^{Q} r_n \qquad (1)$$

where $r_n = 1$ if $x_n \equiv y_n$ and $r_n = 0$ if $x_n \not\equiv y_n$, and Q is the number of characteristics. The record is selected if the degree of error $\dfrac{Q - W}{Q}$ is less than a threshold which is preset to suit the particular application.

Another method of measuring match involves an algorithm such as the record being selected if

$$N = r_1 \,.\, r_2(r_3 \,.\, r_4 \vee r_5) = 1, \qquad (2)$$

where $r_n = 1$ if $x_n \equiv y_n$ and $r_n = 0$ if $x_n \not\equiv y_n$.

This algorithm has the effect of attaching great importance to some characteristics [e.g. in the above formula (2) to r_1 and r_2] but much less importance to others, whereas Glantz' method attaches equal importance to all characteristics (if the transform is simple).

Probability theory has recently been applied to measuring similarity in library indexing by Maron *et al.* (1959). This technique is based on the uncertainty involved in designating characteristics to both library documents and requests and, by use of weighting functions, it enables characteristics to be designated as partly applicable. The measure of similarity is a logarithmic function of the weights and the *a priori* probabilities of the use of the characteristics.

Another way of measuring match in the absence of errors involves measuring the information supplied by an inquiry toward isolating its counterpart record. If errors are admitted then the effect of the probability of error must also be included in the measurement. This method is, of course, employed in communication theory to measure information gain, and this paper is concerned with its application to measuring the match between an inquiry and records in an index.

Information gain has been defined in communication theory (Woodward, 1953; Shannon, 1948) as

$$\log P_2 - \log P_1$$

where P_1 is the initial probability and P_2 the final probability of, for example, a record being the one wanted by an inquiry.

Method of Measuring Match

Before receiving information about an inquiry x, the probability that it refers to a record y is defined as P_{xy} (the *a priori* probability). If $x_a \equiv y_a$ the probability that the inquiry refers to a record containing y_a is increased (to first order) to $1 - P_a(e)$, where $P_a(e)$ is the combined conditional probability of error when x_a is quoted or y_a inspected. Thus the information gain by use of the characteristic x_a is

$$I' = -\log P_{xy} + \log P_{sa} + \log [1 - P_a(e)] \quad (3)$$

where P_{sa} represents the uncertainty within the set (s_a) of records which contain y_a.

If $x_a \not\equiv y_a$ then the record y may still be the one referred to by inquiry x because there may have been an error. This circumstance represents a reduction of information gain because the final uncertainty that record y is the wanted one is higher than the initial uncertainty before characteristic x_a was applied. In this instance

$$I'' = -\log P_{xy} + \log (P_{aex} \times P_{ax} \times P_{sa} + P_{aey} \times P_{ay} \times P_{sa}) \quad (4)$$

where P_{aex} and P_{aey} are the probabilities of an error occurring in an inquiry and a record;
P_{ax} and P_{ay} are the probabilities of an inquiry and a record, when in error, misquoting a specified characteristic; and
P_{sa} is the uncertainty of a record, within one group of records quoting the same characteristics, being the wanted one.

It will be noted that the information gain when, for example, $x_a = y_a$ may be dependent on whether $x_b = y_b$ and on whether this fact has been applied. For example, if $x_b = y_b$ the application of $x_a = y_a$ shows a gain

$$I_1' = -\log P_{xs} + \log P_{xs}' + \log [1 - P_a(e)] \quad (5)$$

where P_{xs}' represents the uncertainty within the set (s') of records which contain y_a and y_b. The differences between equations (3) and (5) will be noted.

The total information gain by application of inquiry x to record y can be written as

$$I_t = \Sigma[q_r S_r I' + S_r(1 - q_r)]I''] \quad (6)$$

where $q_r = 1$ and $S_r = 1$ when $x_r = y_r$;
$q_r = 0$ and $S_r = 1$ when $t \neq y_r$;
$S_r = 0$ when x_r or y_r are not quoted.

An important feature of the above formula is that it makes provision for all three states of inquiry and record characteristics, i.e. matching, mismatching, and absence.

In the special case where the distributions of characteristics and error are independent of each other, then the information gains represented in (3) and (5) are identical. Furthermore, if the distribution of inquiries to records is uniform, if there are N records in the index and n_a containing y_a, and if $x_a \equiv y_a$, then

$$I' = \log \frac{N}{n_a} + \log [1 - P_a(e)].$$

Furthermore, if

$$P_a(e) \ll 1, \text{ then } I' \doteq \log \frac{N}{n_a}. \quad (7)$$

Also, under these circumstances

$$P_{ax} \times P_{sa} = P_{ay} \times P_{sa} = P_{xy}.$$

So if $x_a \not\equiv y_a$,

$$I'' = \log (P_{aey} + P_{aex}) = \log P_a(e). \quad (8)$$

The total information gain is

$$I_T = \Sigma\left[q_r S_r \log \frac{N}{n_r} + S_r(1 - q_r) \log P_n(e)\right]. \quad (9)$$

If the distributions of characteristics are not independent, then the exact calculation of information gain involves the use of conditional probabilities. It is possible that some characteristics may be independent of each other, but other characteristics may be dependent on a combination of other characteristics. However, it is unlikely that all the characteristics in an index are completely independent because such an index would have to contain records with all combinations of characteristics.

If information gain is calculated on the assumption that all characteristics are independent, whereas they are not, the calculated gain will be greater than the true gain. The maximum true gain is $\log N$, where there are N records, and it is possible to use $\log N$ as a threshold for determining whether or not a record is likely to be the one wanted by an inquiry.

The Application of the Measure of Match to Names

If the characteristics a, b, c, etc., represent names in a telephone directory, e.g. "a" represents surname, "b" christian name, etc., the measure of match can be easily calculated using formula (9) when the name quoted in the inquiry is identical to the name in the record, when both names are totally different, or when the name is not quoted in either record or inquiry. However, when there is partial agreement between the quoted names, the calculation is more difficult. This difficulty can be overcome by classifying groups of like names together as previously suggested (Wright, 1960). If the individual names are $a_1, a_2 \ldots a_p \ldots a_q \ldots$ and these are grouped under $A_1, A_2 \ldots A_r \ldots$, then if

$$x_{a_p} \not\equiv y_{a_q}$$

but a_p and a_q are both classified under A_r, then the information gain can be calculated from

$$I = \Sigma\left[q_r s_r \left\{\log \frac{N}{n_r} + \log [1 - P(e)]\right\} + S_r(1 - q_r) \log P_a(e)\right]. \quad (10)$$

It will be noted that formula (9) is not used because, in such circumstances, $P_a(e)$ is likely to be high.

If it is impractical to use a classification system then it is possible to measure the degree of match on the basis of identity between the individual letters or figures of names. Such a measure is, however, likely to be very approximate unless conditional probabilities of error are used, and this is impractical where classification systems are impractical.

When comparison of individual figures or letters is used to determine a measure of partial match, the comparison is likely to be a slow process because it is practical to compare only strings of characters in a fixed sequence quickly on a computer. However, it is possible to modify inquiries by *ad hoc* methods in the hope that any error in sequence will be corrected, and a technique of generating "secondary inquiries" by so-called heuristic methods has been previously mentioned (Wright, 1960). This technique can be used in conjunction with information-gain measuring techniques, provided that a proper allowance is made for the probability of error.

Description of a Demonstration DEUCE Computer Program

The program was designed to demonstrate the use of classification, heuristic, and information-gain techniques as applied to finding records in an index.

The index used in the demonstration is similar to a small section of the MPNI index and contains about 350 records, each of which contains surname, two christian names, birth date, NI number and, where applicable, married woman's maiden name. About thirty different surnames appear in the demonstration index and all are somewhat similar, e.g. they include Leaver, Lever, Levi, etc., and could be classified in single group. The distribution of these surnames is approximately the same as occurs in the London telephone directory. Only about thirty christian names are used, and the distribution of them and birth dates is assumed to be random.

A preliminary analysis of the index was made to prepare dictionaries of surnames with known variants of spelling, and of christian names with known alternative names and known spelling variants. The value of $\log \frac{N}{n}$ was calculated for each surname and stored in each record, and the values of $\log \frac{N}{n}$ were calculated for initials, christian names, and for each pair of figures in the birth date. Also the probabilities of occurrence of various errors were estimated.

The first step in the program is to transform an inquiry into a standard layout and code.

Blanks which occur on the left-hand side of the surname are excluded and the surname is compared with the surname dictionary. If the surname is identical to an entry then the surname code number is extracted from the dictionary and substituted in the inquiry. If the surname quoted in the inquiry is not identical to a dictionary entry, a classification algorithm is applied to test whether or not the quoted surname has the same structure as the names in the group. The process is a crude extension of the Soundex code (Wright, 1960) which takes into account some errors which arise in writing and reading. The flow diagram of the algorithm for the first letter is shown in Fig. 1. Similar algorithms

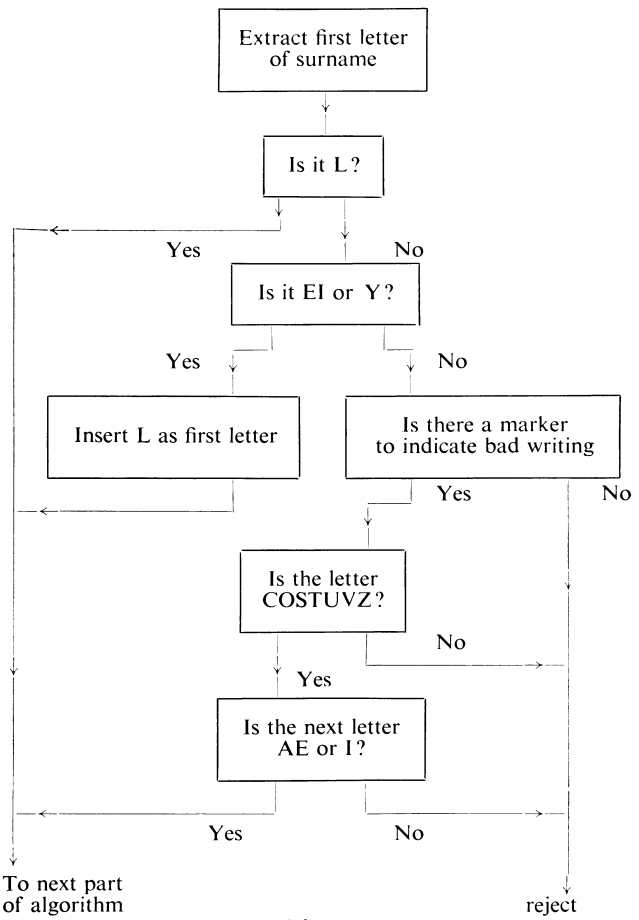

Fig. 1

are used for the remaining letters until a second consonant is reached. If the surname does not have the same structure as the names in the group of surnames contained in the demonstration index, then married women's maiden names are substituted for the quoted surname, and the dictionary test is repeated. If this name is not identical to one of the names in the dictionary, the christian names are substituted for surname and the dictionary test is again repeated. If the test still fails the inquiry is rejected. Otherwise it is accepted and a mark is set to indicate the type of acceptance.

The christian names, if quoted, are transformed and compared with a dictionary of standard christian names. This dictionary contains all known spellings of names in the index in code form, and the transform is again somewhat similar to the Soundex code but it retains more of the original information. For example, CK and CH are transformed to C and only the first three

different letters are used, e.g. Mike, Michael, Mick are all transformed to MIC and coded in the single dictionary entry as M20. The names Elizabeth, Liz, etc., are transformed to LIS and Betty is transformed to BET. Since Betty is a derivative of Elizabeth both names are coded in the dictionary under the same code number, e.g. E17. If only initials are quoted they are retained in their existing form.

The next stage of processing is the comparison with the records in the index, the names of which are coded as described above. Each inquiry is compared with all the records in the group, and information gain is used as a measure of match of the inquiry to record. The record with the highest gain I_M is stored while comparisons with the index proceed. After processing all records, the gain I_M is compared with a threshold gain $I_H = \log N$. If $I_M > I_H$ the record is accepted as a good match and is punched out. If $I_M < I_H$, up to seven secondary inquiries are manufactured and each is processed against the index in turn. The secondary inquiries have initials or christian names interchanged and numbers in birth date interchanged if such interchanges are valid.

After each secondary inquiry has been processed against the index, the record which attains the highest gain to any of the secondary inquiries or the original is punched out together with a code to indicate whether $I_M > I_H$, i.e. whether the match is a "good one."

The time taken on a DEUCE computer for the complete search for one inquiry is about $1\frac{1}{4}$ minutes if secondary inquiries are not used, or 10 minutes if they are. Both figures are inclusive of time spent on tape rewind. It would be practical to combine the first searches for several inquiries in the hope that some of the inquiries would be correctly quoted. Such inquiries could be answered in the first processing and the more complicated program used only for inquiries which were not quoted correctly.

Acknowledgements

This work has been carried out as part of the research programme of the National Physical Laboratory, and is published by permission of the Director.

The help provided by Mr. A. M. Day in preparing the DEUCE computer programs is acknowledged.

References

GLANTZ, H. P. (1957). "On the Recognition of Information with a Digital Computer," *Journal ACM*, Vol. 4, p. 178.

MARON *et al.* (1959). "Probabilistic Indexing: Technical Memo. 3 June 1959," Data Systems Project Office, *Ramo–Wooldridge Corporation.*

SHANNON, C. E. (1948). "Mathematical Theory of Communication," Vol. 27, *Bell System Technical Journal*, pp. 379–423 and 623–58.

WOODWARD, P. M. (1953). *Probability and Information Theory with Applications to Radar*, Pergamon Press.

WRIGHT, M. A. (1960). "Mechanizing a Large Index," *The Computer Journal*, Vol. 3, p. 76.

Book Review

Analogue Computation in Engineering Design, by A. E. ROGERS and T. W. CONNOLLY, 1960; 450 pages. (New York: *McGraw-Hill Book Co. Inc.;* London: *McGraw-Hill Publishing Company Ltd.*, 124s. 0d.).

A book from the McGraw-Hill Series in Information Processing and Computers. It claims (Chapter 2) to "answer questions of where and how one applies the computer, rather than how one builds the computer itself." In fact, the scope is considerably wider than this, including a concisely informative chapter on analogue equipment (D.C. voltage analogue, using the PACE equipment for the majority of the examples) and chapters on the particular set of mathematical concepts with which the analogue computer user needs to be most familiar. The rate at which information is presented in these chapters is necessarily high in order to progress from the assumed initial knowledge of elementary differential equations to the relatively sophisticated concepts involved in, for example, assessing the effects of noise on a system. The greatest value in these chapters lies in the fact that they collect in one book the techniques used most frequently in a form in which they can be applied directly. An excellent reading list at the end of each chapter refers to the standard works should more rigorous treatments be required. The second half of the book discusses in detail the simulation and analysis of a wide range of the systems for which the analogue computer is well suited, including aircraft and missile problems, nuclear reactor engineering and mechanical design problems. In each case the problem is introduced and, where appropriate, the governing equations are derived so that the computer set-up can be seen to follow logically. The bibliography (arranged chronologically) appears to be largely a random selection from the mass of analogue computer papers, but it has a wide pass-band.

The book will probably prove to be of most use as a day-to-day reference for users of analogue computers, and as such is a very welcome addition to the literature.

G. J. HERRING.