An Error Analysis For Functions of Qualitative Attributes With Application To Information Retrieval

By
D.M. Jackson

Technical Report
70-56
March 1970

Department of Computer Scienc
Cornell University
Ithaca, New York 14580

Presented at the THIRD INTERNATIONAL SYMPOSIUM ON COMPUTER AND INFORMATION SCIENCE (COINS-69). Florida December 18-20, 1969

ACKNOWLEDGEMENTS

I am indebted to Dr. R. M. Needham of the University Mathematical Laboratory, Cambridge, Englsnd, for his encouragement in the planning of this work. This work was supported, in part, by Grant Number GN-543.1 from the Office of Scientific Information Service of the National Science Foundation and, in part, by the Office of Scientific and Technical Information of the United Kingdom.

SUMMARY

The use of overlapping, non-hierarchical classifications in information retrieval is considered. It is assumed that the population of objects to be classified is such that only a subset of the classes satisfying the classificatory criterion may be found. The effect of this assumption on the measurement of classification stability is considered. As a step towards the determination of stability, a general technique is presented for deriving the expectation of a statistical function of the similarities between the objects of the population. It is assumed that the objects are describable in terms of two-state attributes which are susceptible independently and equiprobably to error with an assignable probability. Two commonly encountered similarity functions are treated in detail. The techniques disclosed are applicable, in principle, to classification algorithms, whether hierarchical or non-hierarchical, which utilize a similarity matrix giving the similarities between pairs of objects described by two-state independent attributes.

TABLE OF CONTENTS

Acknowledgments.................................. i

Summary.. ii

1 Preliminary considerations................. 1

2 The measurement of similarity.............. 6

3 Constraints on classification algorithms... 12

4 Errors in the assignment of attributes to objects...................... 14

5 Error analysis............................. 21

6 Conclusion................................. 32

References..................................... 34-35

An Error Analysis for Functions of Qualitative Attributes with Application to Information Retrieval.

D. M. Jackson

1 Preliminary considerations.

Type of classifications which will be considered are unusual in that classes are intentionally allowed to overlap. The role of overlap between classes of terms as a possible precision device in information retrieval is clear since the different senses or contextual uses of a term or keyword may be identified by determining the term's affiliations to different classes. In many applications, however, it is a requirement that the objects to be classified should be assigned to disjoint classes, and the clustering and hierarchical methods for achieving this comprise the main bulk of classification techniques. In plant taxonomy, for example, where the objectives are somewhat different since categorical differences between the objects to be clsssified are of primary importance, Williams and Dale [(1)] are sceptical about the use of overlapping classifications. Techniques appropriate to these applications are given by Sokal and Sneath.[(2)] Nevertheless techniques for constructing overlapping classifications have been applied by Parker-Rhodes[(3),(4)] and myself[4] to mycological data containing ecological and meteorological descriptors, with some measure of success.

In information retrieval we are interested in forming classes of terms in such a way that they may be used to achieve an increase in retrieval performance. There is, however, no a priori reason for claiming that a particular classification, in which the members of the classes are related in some specific way, is better than any other classification of the terms. The value of a classification is determined empirically by evaluating the effect on retrieval performance when the classification is incorporated into the system. Thus, in constructing classifications for information retrieval, the Platonic concept of error in the assignment of terms to classes is entirely absent. That is, there is no prior classification which automatic techniques are intended to reconstruct in detail; and there is therefore no basis for the probabilistic assignment of terms to classes to give the extent that properties of particular terms in the vocabulary are inferable from a knowledge of the classes to which each belongs. Sutcliffe[5,6] has attempted to replace the Platonic concept of error by comparing statistically a large number of classifications of the same population to determine the typical and atypical allocation of objects to classes. The stringent conditions under which observations on the objects must be made so that atypical class allocation may be adopted as an indication of error make this approach inapplicable in information retrieval.

Certain classifications suggest themselves as devices for improving retrieval performance, and it is quite consistent to assert that these improve retrieval performance provided that such an assertion is phrased in terms of an hypothesis which is susceptible to testing. It may be supposed, for example, that a classification of the term vocabulary into groups of synonyms can be used to improve the performance of a retrieval system beyond that obtained with other classifications. Gotlieb and Kumar(7) have proposed another classification based on the a priori (pre-assigned) semantic relationships between terms rather than on the relative frequency with which terms occur together in documents. The assertion that a particular classification on retrieval performance, and the validity of the assertion for document collections in general may be determined provided sufficiently many document collections are examined. However, the success with which terms have been arranged into groups of synonyms must be determined by inspection of the classes and an intuitive assessment of the constitution of each class. As a relation between terms, synonymy is quite strong and, indeed, methods have been proposed for isolating synonyms automatically by Needham[8] and Lewis.[9] Topical relatedness is, however, less definite and the answer to the question whether two terms are topically related (that is: the two terms may relate to the same topic area), as opposed to synonymously related, is less obvious. In probabilistic classifications, the probability of the assignment of terms to a class gives the reliability of the information conveyed by asserting that the term belongs to the class. Topical relatedness is not sufficiently definite a relationship in the first place to permit the calculation of probabilities with any degree of accuracy. Accordingly, a classification in which the assignment of terms to classes is non-probabilistic is more appropriate for dealing with topical relatedness.

The classification which will be considered consist of non-hierarchically arranged classes of terms which show a marked tendency to occur together in the documents of the collection. The classes formed by this criterion will hopefully specify topic areas within the field covered by the collection, and accordingly terms which appear together in the same class will be said to be topically related. The criterion on which classes are formed is a weak one in that classes are to be composed of terms which tend to occur together in the documents of the collection. The criterion is therefore less restrictive than that for cliques[1] in which the classificatory criterion must be exhibited between every pair of objects belonging to a clique. A further requirement of the classification is that the classes be self-generated[2] in the following sense. Classes are defined as sets of terms which satisfiy the classificatory criterion, and the complete term vocabulary must be explored to find such classes. It is inadmissible to specify a set of possible topic classes ab initio, and subsequently to distribute the terms to the most appropriate of these since, first, these

---

1 A clique is a maximal complete subgraph of a graph in which the presence of an arc between a pair of nodes indicates that the nodes are related. The degree of relatedness may be specified by the length of the arc.

2 Williams and Dale (op. cit. p. 38) use the term "self-structuring"

classes must be treated as unknown; secondly, significant yet intuitively obscure classes would not be located; and thirdly, the number of classes required is unknown. Such procedures, for example Sebestyen's[10], might however be adopted as an economical means of updating an existing classification to avoid complete reclassification of the term vocabulary as the document collection expands.

2 The measurement of similarity.

Automatic classification is concerned with assigning objects to classes according to the information about each of the objects. The objects to be classified are described or defined in terms of measurements or observations made upon them, and the information is characteristically conveyed by specifying the qualitative or quantitative attributes which the objects possess. Qualitative attributes are sometimes called "qualities", and may be two state, indicating the presence or absence of the quality, or may be multistate. The attribute "leaf-shape", for example, in which linear, ovovoid and pandurate must be distinguished, is three-state. The attribute "leaf-length", possibly averaged over a suitable sample, is quantitative.

A great deal of discussion in taxonomic literature has been devoted to the problems arising from the interdependence of attributes. Suppose, for example, that the binary attributes "black" and "white" were used in describing the objects of a population. Then, provided that these attributes were relevant to the population, the value of one of the attributes would determine the value of the other. A less obvious example is given by the use of "carnivorous" and "ungual". Here the dependence is not a logical one. Sokal and Sneath[3] are of the opinion that dependencies between attributes, which are known without recourse to analysis, must be eliminated before the objects are classified. William and Dale[4], while admitting that such dependencies may dominate the analysis, consider that their use in classification is a matter of taste - the sole criterion is the interest or otherwise of the user. Thus, since the dependence between "carnivorous" and "ungual" has been established, subsequent users

---

3 op. cit. p. 66.

4 op. cit. p. 42.

can afford to be more sophisticated in their choice of initial attributes. The problem corresponds in information retrieval to the presence of terms which co-occur frequently compared with the frequency of co-occurrence of either term. The acceptability of the dependencies must be determined by their effect on retrieval performance. There is evidence that they have a deleterious effect. The dependence, however, is not a logical one.

Suppose that x and y are two objects of a population U and suppose that d(x,y) is the similarity between x and y. Then the following conditions are imposed:

a) $d(x,y) = d(y,x)$ for all x,y in U. (symmetry)

b) $d(x,y) \geq 0$ for all x,y in U

c) $d(x,y) = 0$ if $x = y$ for x,y in U (indistinguishability) of identical objects)

d) $d(x,y) = 0$ only if $x = y$ for x,y in U (distinguishability of non-identical objects)

e) $d(x,z) \leq d(x,y) + d(y,z)$ for all x,y,z in U (triangle inequality)

A distance function which satisfied (a)-(e) is called Euclidean. A function in which (e) is not everywhere true is called a quasi-metric. A function in which (d) is not everywhere true is called pseudo-metric. (or semi-metric). Functions have been examined for which (e) is replaced by the more restrictive condition:

e') $d(x,z) \leq \max(d(x,y),d(y,z))$ for all x,y,z in U. Functions which satisfy (e') are called ultra-metric, and their use in constructing taxonomic hierarchies has been discussed by Johnson[11] and by Jardin.[12] A number of functions satisfying some or all of (a)-(e) have been proposed for information retrieval and for wider applications. Typical among the functions used in ecology and anthropology, for example, are those of Pearson,[13] Mahalanobis[14] and Rao.[15] These functions are intended for use when the descriptions of the objects to be classified themselves result from the statistical determination of the mean and variance of a character or atrribute over a sample of all instances of the object in the population. The counterpart in information retrieval is the determination of the "distance" or similarity between index terms, taking account of the variability in the assignment of terms to documents by a sufficiently large number of indexers independently indexing the collection. The virtue of the uniterm indexing scheme rests in its simplicity and cheapness to apply in practice. The possibility of repeated indexing of the same document collection would grossly add to the cost of obtaining the initial index descriptions, and therefore cannot be considered.

Observations on the terms of the vocabulary consist of the names (call numbers) of those documents to which the terms have been assigned. Since the documents of the collection are indexed using uniterms, each document may be characterized by a vector whose components, one or zero, indicate whether or not a particular term has been used in indexing the document. The complete collection may therefore be represented by a rectangular array, D, the i-th row of which gives the index description of the i-th. document. $D_{ij}=0$ indicates that document i has not been indexed by term j, while $D_{ij}=1$ indicates that it has. Since, in associative retrieval, it is the terms which are to be classified by their co-occurrence in documents, the terms are the objects of the universe (the term vocabulary) while the documents are two-state or binary attributes. D is called the document array or the document matrix. Co-occurrence of terms within documents may be determined from the matrix D quite simply. Consider two terms t_j and t_k. Then the frequency of co-occurrence of the terms is given by $|D_{.j} \wedge D_{.k}|$ where $D_{.j}$ denotes the j-th. column of the document array D. The relative frequency of co-occurrence of terms within documents may now be defined as:

$$s_1(t_j,t_k) = |D_{.j} \wedge D_{.k}| / |D_{.j} \vee D_{.k}|$$

This is a quasi-metric and varies between o, then t_j and t_k never occur together, and 1, when t_j and t_k always occur together. The function, which arises naturally in information retrieval, is ascribed to Tanimoto.[16] It is biassed in favour of positive occurrence of terms in documents. The common non-occurrence of terms in a document does not contribute towards the similarity of the two terms. It has already been seen in Section 1, that the similarity between pairs of terms is based on actual occurrence of terms in documents. A symmetrical form of the function may, however, be defined. A function with similar properties is the correlation coefficient, defined by:

$$s_2(t_j,t_k) = \sqrt{\frac{|D_{.j} \wedge D_{.k}|^2}{|D_{.j}| \cdot |D_{.k}|}}$$

It is clear that other functions may be defined for calculating the similarity between pairs of terms based on their frequency of co-occurrence within documents. Account may be taken, for example, of the variation in the numbers of terms assigned to documents. The functions defined above, however, have the substantial a practical advantage of making the manual inspection of classes, member by member, a relatively straightforward matter.

3 Constraints on classification algorithms.

The following are constraints generally placed on classification algorithms:

a) The classification must be well-defined. That is, application of the algorithm must supply a single result.

b) The algorithm must be unaffected by a permutation of the names of the terms. That is, the algorithm must be independent of the labelling of the terms.

c) The algorithm must be independent of scale. That is, the algorithm must be unaffected by multiplication of the similarity matrix by a positive, non-zero constant.

Independence of scale is an obvious requirement since the scale of the similarity function is essentially arbitrary.

The stronger condition of invariance under a monotonic transformation of the similarity function has also been suggested, and this leads to methods in which distances or similarities between objects are largely irrelevant. The methods rely on the inequalities between the similarities of pairs of objects. Similarity functions induce a partial ordering on the set of objects to be classified, and a classification is constructed such that a partial ordering derived from it agrees as closely as possible with the partial ordering induced on the data. The measurement of the agreement between the partial orderings is given by Benzécri.[17] The methods are accordingly insensitive to monotonic transformations of the similarity function, with the result that a number of similarity functions given by Sokal and Sneath may be regarded as equivalent. A review of these methods is given by de la Véga.[18] Independence of labelling requires that a classification should not be dependent on the order in which the terms are treated, and is demonstrable only in a classification in which all classes which satisfy the classificatory criterion have been enumerated. For the majority of classification algorithms, however, computational limitations prevent the discovery of all of the classes. A permutation of the names of the terms would result in a classification which, although sharing a substantial number of classes, would differ in a number of classes. The failure of (a) and (b) are related. Relabelling enables more classes to be found, and examination of all permutations corresponds to inspecting all subsets of the terms to determine which satisfy the classificatory criterion. It is clearly impracticable to examine all subsets since there are too many of them. Provided a large enough number of classes is found, the failure of (a) and (b) is unimportant.

4 Errors in the assignment of attributes to objects.

In any practical application of classification techniques, errors are liable to occur during the collection or compilation of the information about the objects to be classified, and it is therefore appropriate to examine the effect of such errors on classifications subsequently produced. We shall be concerned only with sets of objects which have been described by the set of qualities which they possess. The possession of a quality is not quantified; its presence or absence only is recorded. The attributes and the objects are given numerical codes, typically integers running from 1 to m for the attributes, where r is the number of attributes used for describing the data collection, and running from 1 to n for the objects, where n is the number of objects in the population. The occurrence of the number j in the list of attributes belonging to object i indicates that object i possesses attribute j, while the absence of j indicates that object i does not possess attribute j. The following remarks describe the types of error which may occur.

a) In the large collections of information to which automatic classification procedures must be applicable, there is a certain probability of finding "transcription" errors. These errors are purely clerical in origin and arise, for example, when information is transcribed into a form suitable for processing by computer. Some of the errors may be detectable if conventions are imposed on the presentation of the data. If the attributes are required to be presented in increasing order of their code numbers, then an attribute which is out of order in the sequence of attribute code numbers relating to a given object signals an error somewhere in the sequence. Stronger conventions may be imposed on object codes if it is assumed, as a precaution, that objects which do not have a description in terms of the available attributes, have been removed. Accordingly, object codes must appear as a sequence of successive integers in increasing order of magnitude, and their failure to do so is indicative of an error. The codes themselves may be tested to ensure that they are well-formed and are in the appropriate ranges, namely 1 to m for the attributes, and 1 to n for the objects.

b) The description of an object may be incomplete. There may be no way of arriving at a set of attributes adequate to describe the population of objects until some crude form of classification has been done. For example, information consisting of the "significant" and "relevant" symptoms exhibited by a set of patients may have been collected by a number of doctors. Observations may be one doctor on his patients may not have been made by another doctor on his patients, although the conditions of the patients may have been clinically similar. If information of this type, collected over a period of time, is to be classified, retrospective judgements may enable some, but probably not all, of the original information to be retrieved with some degree of accuracy.

c) The data may be heterogeneous, in the sense that a number of attributes may not be relevant to a number of the objects. Since, however, the attributes considered are binary, "not-relevant" must be compounded with "absent". The situation is clarified if three-valued attributes are used, the additional value corresponding to "not-relevant". Alternatively, if the number of attributes affected is small, they may be removed completely from the descriptions, unavoidably incurring, however, the removal of all their useful references to a proportion of the objects. The loss of these useful references may be a greater disadvantage than the confusion incurred by compounding "not-relevant" with "absent".

d) An object may be mistakenly described by an attribute which it does not possess.

Each of these types of error may clearly occur in documentary data. (a) may occur whenever a set of documents is prepared for machine use. (b) may occur when the collection expands to include material not envisaged at the time the original documents were indexed. For example, in computer literature, the disrinction now drawn between "parallel" and "series" computation may not have been made originally, although in retrospect, the distinction would have been a relevant one. Errors of type (c) obtrude, but their effect on classification may be eliminsted by a suitable choice of similarity function. Provided that only the positive assignment of a pair of terms to s document contributes towards the similarity between the terms, the distinction between "not-relevant" and "absent" is immaterial. The fact that a document was not indexed by a particular pair of terms does not contribute to the similarity between the terms. The similarity function must therefore not be symmetrical in "1" and "0". Neither of the similarity functions defined in Section 2 are symmetrical in "1" or "0". Uniterm indexing is susceptible to errors of type (d) since no control is exercised over the term vocabulary. Techniques which reduce the effects of this type of error on subsequent classification are of particular interest.

It will be assumed throughout that errors occur independently, in that an error does not necessarily imply the existence of another. Consider, for example, the transcription errors which may arise in a binary sequence of n digits in which information is carried by the position of a digit in the sequence as well as by its value. The complete omission of a digit from the sequence implies that the subsequent digits are all shifted up by one place. Thus errors will be propagated in each subsequent digit position, after and including the omitted digit, whose value differs from the immediate neighbour to the right. Although, an indexed document may be represented as a sequence of binary digits, this representation is used only when the document is processed within the computer. It will be assumed that the documents are coded initially in a form in which positional information of a term code is not utilized so that propagated or systematic errors may be reduced.

In each of these cases, it is hoped that a classification of the material may be achieved in spite of the rrrors the material contains. Obviously, the original information is not recoverable and the best which can be done is to ensure that the erroneous parts of the data do not exert a significant effect on the classification as a whole. Accordingly, a condition, in addition to those of Section 3, must be imposed on the classificstion algorithm as follows. Classifications must be _stable_ with respect to small changes in the msterial to be classified. That is, small changes in the material to be classified lead only to small changes in the classification.[5] For this condition to be appropriate, it must be demonstrated that small changes both in the data and in the classification are measurable quantities.

Small changes in a document array may be measured quite simply. Suppose that two document arrays are given, representing differing index descriptions of the same document collection. It is known that document i of one

---

5 A connection may be seen between _stability_ and _continuity_. Suppose that a classification is produced by a trans formation of the data. Small changes in the dats necessarily lead to small changes in the classification, provided that the transformation is continuous. The classification is therefore stable. It is assumed here, however, that the clsssification algorithm cannt be regarded as a continuous transformation.

array may be identified with document i of the other array. The difference between the document descriptions may be discovered by inspecting row i in each array. Over the whole collection, a measure of the difference between the document descriptions may be discovered by inspecting row i in each array. Over the whole collection, a measure of the difference between the two arrays may be obtained by regarding the arrays as preference tables and computing Kendall's[19] Coefficient of Agreement.

Changes in the classifications are less easily measurable, although a classification of the type described in Section 1, like the document collection, may be represented by means of a binary array. In a classification array C, the element C_{ij} of the i-th. row indicate the presence or absence of term j in class i. Suppose that classification arrays C and C' are produced from two document arrays, D and D', which represent differing index descriptions of the same document collection. The classes of C and C' are defined extensionally and the names which are given to each class are entirely arbitrary. It is therefore not appropriate, although it was for the document arrays, to identify the i-th. row of C (that is, class C_i) with the i-th. row of C' (that is, class C'_i) in the sense that the differences between D and D' cause C_i to devolve into the class C_i'. Nor is it possible to show that C and C' are conformable by determining a permutation p of the integers, such that C_i may be identified, in its extension, approximately with $C_{p(i)}$, since it cannot be assumed that C or C' exhaust the classes which may be found in D or D'. The problem of measuring small differences between pairs of classifications has been treated elsewhere by myself,[20] and further details will not be given here.

5 Error analysis.

Suppose that there are M terms in the vocabulary and N documents in the collection D. The i-th. row of the binary array D gives the terms j which appear in the index description of document D_i. If $D_{ij}=1$, then term j appears in document i, while if $D_{ij}=0$, term j does not appear in document i. The array D accordingly consists of NM bits. Errors of the type listed in Section 4 (a)-(d) all have the effect of changing a '0' into a '1' or a '1' into a '0' for a number of the D_{ij}'s. The errors are assumed independent and equiprobable, as stated in Section 4. The effect of the errors is to change the array D into the array D', Let r be the probability of finding an error in a particular D_{ij}. Thus rMN is the number of errors in the array. The number of errors which will be tolerated may be decided in advance, and from this a value of r may be obtained. Consider now two arbitrary terms t_j and t_k represented by the binary vectors A and B, both of length N, and suppose that these vectors are scrambled by errors into the vectors A' and B', the vectorial representations of the same terms in D'. In terms of the document array D, A and B are given by $D_{.j}$ and $D_{.k}$ respectively, where $D_{.j}$ is the j-th column of D. Thus (A,B) is transformed into (A',B') by a <u>scrambling</u>, S. Let f and f_S be the values of similarity between A and B before and after A and B have been transformed into A' and B'. Suppose that $\emptyset$ is the function of f and f_S which is to be investigated and that $\emptyset^*$ is the statistical estimate for $\emptyset$. For example, $\emptyset$ may be the mean square deviation. Let S* be the set of all possible scramblings for the pairs of vectors of length N. Then:

a) $$\emptyset^* = \sum_{S \text{ in } S^*} P(S)\emptyset(f_S, f)$$

where S denotes a particular scrambling in S*, and P(S) is the probability of the particular scrambling S. Consider now a particular document i. Let A_i and B_i be the i-th. components of A and B. The following combinations may occur for (A_i,B_i), namely: (0,0), (1,0), (0,1) and (1,1). For simplicity, denote (0,0) by a, (1,0) by b, (0,1) by c, and (1,1) by d. When A and B change into A' and B', any one of a,b,c and d may be changed into any other. a,b,c and d will be called <u>categories</u>, and changes from any one of the categories in (A,B) to any other in (A',B') will be called a <u>transition</u>. Let S_{lm} be the number of l's which are converted from m's. Then the following relations hold among the S_{lm}:

$$\sum_l S_{lm} = n_m$$ where n_m is the number of m's in (A,B).

$$\sum_m S_{lm} = m_l$$ where m_l is the number of l's in (A',B').

$$\sum_m n_m = \sum_l m_l$$ = N, the number of documents in the collection.

A distinction must be drawn between the problem treated here and the large class of problems called Matching Problems. Barton [(21)] gives a statement of the problem, for simplicity with reference to packs of cards although the theory is applicable to any set of objects, as follows: 'To fix ideas we consider a "target pack" of N cards laid out in a row and a "matching pack" of the same number of cards laid out one by one beside the target forming N <u>pairs</u> of which r are pairs of like cards (or matches).' The problem is to find the distribution of r. There are a number of conditions which may complicate the simplest case in which there are N distinct cards. For example, there may be k sorts of cards with a_i cards of the i-th sort: or there may be K packs of cards between which simultaneous matches are to be found. These problems differ from the one treated here becsause, continuing the card playing terminology they are concerned with matches between two packs of cards where one of the packs has been laid out in random order. Instead the problem treated here deals with the number of matches for two given arrangements under the condition that there is a probability that the cards will be misread. thereby affecting the counting of matches. Levene[22] encounters a matching problem in studies of Medelian inheritence and considers the effect of misclassifying the objects, between which matches are to be made on the distribution of the number of matches. Although Levene s problem, with its treatment of misclassification, is closer to the problem treated here it is still remote from it for the reason given above.

Let P(l m) be the probability of the transition $m \to l$. There are n_a s a's which must be distributed among the categories a.b c.d in such a way that there are S_{aa} in a, S_{ba} in b, S_{ca} in c and S_{da} in d. Let T_a be the probability of a particular one of these. Then:

$$T_a = \frac{n_a!}{S_{aa}!\,S_{ba}!\,S_{ca}!\,S_{da}!} \cdot P(a,a)^{S_{aa}}.P(b,a)^{S_{ba}}.P(c,a)^{S_{ca}}.P(d,a)^{S_{da}}$$

$$= \frac{n_a!}{\prod_l S_{la}!} \prod_l P(l,a)^{S_{la}}$$

Now $P(S) = \prod_m T_m$

so substituting the above expression for T_a, we obtain:

$$\text{b)} \quad P(S) = \frac{\prod_m n_m!}{\prod_{l,m} S_{lm}!} \cdot \prod_{l,m} P(l,m)^{S_{lm}}$$

Now, from (a).

$$\phi^* = \sum_{S \text{ in } S^*} P(S).\phi(f_S,f)$$

$$= \sum_{\substack{m_i \\ \sum_i m_i = N}} \; \sum_{\substack{S \\ \sum_i S_{ij} = n_j \\ \sum_j S_{ij} = m_i}} P(S).\phi(f_S,f)$$

by splitting the sum into first a sum over all scramblings which reach a particular m_a, m_b, m_c, m_d and then a sum over all permitted values of these.

Therefore we may write:

$$\text{c)} \quad \phi^* = \prod_m n_m! \cdot \sum_m \phi(f_S,f).R$$

where $R = \sum\limits_{\substack{S \\ \sum_i s_{ij}=n_j \\ \sum_j s_{ij}=m_i}} \left(\dfrac{\prod\limits_{l,m} P(l,m)^{s_{lm}}}{\prod\limits_{l,m} s_{lm}!} \right)$

Consider now the polynomial $R^* = \left[\sum\limits_{l,m} p_l \cdot q_m \cdot P(l,m) \right]^N$

This is a polynomial of degree N in $p_l \cdot q_m$. Then by multinomial expansion:

$$R = \frac{1}{N!} \, \underline{\text{coeff. of}} \prod_l p_l^{m_l} \cdot \prod_m q_m^{n_m} \; \underline{\text{in}} \; R^*$$

$$= \frac{1}{N!} \, \underline{\text{coeff. of}} \prod_m q_m^{n_m} \text{ in } \underline{\text{coeff. of}} \prod_l p_l^{m_l}$$

$$\underline{\text{in}} \left[\sum_l p_l \sum_m q_m \quad P(l,m) \right]^N$$

$$= \underline{\text{coeff. of}} \prod_m q_m^{n_m} \; \underline{\text{in}} \; \frac{1}{\prod_l m_l!} \prod_l \left(\sum_m q_m \cdot P(l,m) \right)^{m_l}$$

by multinomial expansion

$= h(\underline{n},\underline{n})$ say.

Therefore, from (b):

$$\text{d)} \qquad \emptyset^* = \prod_m n_m! \sum_{\underline{m}} \emptyset(f_S, f) . h(\underline{m}, \underline{n})$$

$$\text{where } h(\underline{m}, n) = \underline{\text{coeff. of} \prod_m q_m^{n_m}} \ \underline{\text{in}} \ \frac{1}{\prod_l m_l!} . \prod_l \left(\sum_m q_m . P(l, m) \right)^{m_l}$$

The formula will now be applied to the similarity function $s_1(t_j, t_k)$, defined in Section 2, to obtain an expression for $\emptyset^*$ in terms of the error probability r. Let a, b, c, d denote, as before, the categories (0,0).(1,0), (0,1) and (1,1). Since $s_1(t_j, t_k)$ is symmetrical in t_j and t_k, there is no distinction between (0,1) and (1,0) and accordingly these will be replaced by the category $x = b \vee c$. The determination of $\emptyset^*$ may, therefore, proceed using only the three categories a, x and d. Suppose that, before scrambling, the number of occurrences of each category is n_a, n_x and n_d respectively, and that after scrambling, the number of occurrences of the same categories is m_a, m_x and m_d. Then the values of the similarity between t_j and t_k before and after scrambling are given respectively by;

a') $f = n_d/(N - n_a)$ and $f_S = m_d/(N - m_a)$

Now r is the probability that there is an error in a particular digit of the document array D. Then the transition probabilities are given, for each pair of categories, by:

b') $[P(l,m)] = \begin{bmatrix} s^2 & rs & r^2 \\ 2rs & r^2+s^2 & 2rs \\ r^2 & rs & s^2 \end{bmatrix}$

where $s = 1-r$. Thus, from the expression for $h(\underline{m},\underline{n})$ given in (d) and substituting the expressions for $P(l,m)$ given in (b'), we have:

c') $h(\underline{m},\underline{n}) = s^{2N} \cdot \sum_I C_I \cdot t^{N + I(a,d) + I(d,a) + I(d,d))} \cdot (1+t^2)^{I(x,x)}$

where the sum is taken over all $I(l,m)$ such that $\sum_m I(l,m) = m_l$ and $\sum_l I(l,m) = n_m$. Also, in the above:

d') $C_I = 2^{m_x - I(x,x)} / \prod_{l,m} I(l,m)!$ and $t = r/s$

Thus $\emptyset^*$ may be expressed as a polynomial in t, which may be truncated once a sufficient number of terms has been collected, depending upon the magnitude of t. The expression for $\emptyset^*$, for the particular similarity function chosen, however, is unmanageable and no further reduction of $\emptyset^*$ to simpler terms has been achieved.

The analysis may, however, be carried further for functions $\emptyset$ which are expressible as polynomials in m_a, m_b, m_c and m_d. Suppose that $\emptyset$ is expressible in this form.

Then:

e) $\emptyset(f_S, f) = \sum_{\underline{i}} a_{\underline{i}} \cdot m^{\underline{i}}$

where $a_{\underline{i}}.m^{\underline{i}} = a_{i_x,i_y}\ldots\ldots . m_x^{i_x} . m_y^{i_y}\ \ldots\ldots$

Now, from (a):

$$\phi^* = \sum_{S \text{ in } S^*} P(S).\phi(f_S,f)$$ where S^* is the set of all possible scramblings. Therefore:

f) $\phi^* = \sum_{\underline{i}} a_{\underline{i}}.B_{\underline{i}}$ where $B_{\underline{i}} = \sum_S m_{\underline{i}}.P(S)$

Now $m^{\underline{i}} = \prod_l m_l^{i_l} = \prod_l \left(\sum_m S_{lm}\right)^{i_l}$

so that:

$$B_{\underline{i}} = \sum_S \prod_l \left(\sum_m S_{lm}\right)^{i_l} .P(S)$$

P(S) is given in terms of the P(l,m) by (b). We define the operator D_{pq} by:

g) $D_{pq} \equiv p(p,q).\ \dfrac{d}{dP(p,q)}$

Then differentiating P(S) formally with respect to P(p,q), we have: $D_{pq}P(S) = S_{pq}.P(S)$. Therefore:

$$B_{\underline{i}} = \prod_l \left(\sum_m D_{lm}\right)^{i_l} . \sum_S P(S)$$

$$\prod_l \left(\sum_m D_{lm}\right)^{i_l} \sum_S \prod_m \left(\frac{n_m!}{\prod_l S_{lm}!} . \prod_l P(l,m)^{S_{lm}}\right)$$

so:

h) $B_{\underline{i}} = \prod_l \left(\sum_m D_{lm}\right)^{i_l} \prod_m \left(\sum_l P(l,m)\right)^{n_m}$

Consider now, in the notation of Section 2, the similarity function $s_3(t_j, t_k)$, defined by:

$$s_3(t_j, t_k) = (|D_{.j} \wedge D_{.k}| + |\bar{D}_{.j} \wedge \bar{D}_{.k}|)/N$$

where t_j and t_k are two arbitrary terms. The similarity function is symmetrical in t_j and t_k and is also symmetrical in 0 and 1. There is, therefore no distinction between the categories (1,0) and (0,1) and these may accordingly be replaced by the category $x = b \vee c$. In addition, there is no distinction between a and d so these may be replaced by the single category $y = a \vee d$. Suppose that f and f_S are the values of $s_3(t_j, t_k)$ before and after scrambling. Then $f = n_y/N$ and $f_S = m_y/N$ where N is the number of documents in the collection. Suppose that $\emptyset$ is the mean square deviation of f due to the presence of errors. Therefore, $\emptyset = (f_S - f)^2$ whence:

$$\emptyset = (n_y/N)^2 + (-2n_y/N^2)m_y + (1/N^2)m_y^2$$

Thus $\emptyset$ is expressible as a polynomial of degree two in m_y. Now, from (e), $\emptyset = \sum_{\underline{i}} a_{\underline{i}} \cdot m^{\underline{i}}$, whence, identifying coefficients in $s_3(t_j, t_k)$:

$a_{00} = (n_y/N^2$, $a_{01} = -2n_y/N^2$, and $a_{02} = 1/N^2$. All other $a_{\underline{i}}$ are zero. Thus, since from (f), $\emptyset^* = \sum_{\underline{i}} a_{\underline{i}} \cdot B_{\underline{i}}$, then:

$\emptyset = a_{00}B_{00} + a_{01}B_{01} + a_{02}B_{02}$ where the B's are given, from (g) and (h), by:

$$B_{i_x \cdot i_y} = \prod_l \left(\sum_m D_{lm}\right)^{i_l} \prod_m \left(\sum_l P(l,m)\right)^{n_m}$$

$$= (D_{xx}+D_{xy})^{i_x} (D_{yx}+D_{yy})^{i_y}.(P(x,x)+P(y,x))^{n_x}. (P(x,y)+P(y,y))^{n_y}$$

the transition probabilities $P(l,m)$ are given by:

$P(x,x) = P(y,y) = r^2+s^2$, $P(x,y) = P(y,x) = 2rs$. Now B_{00} corresponds to $i_x=0$ and $i_y=0$. Therefore:

$$B_{00} = (P(x,x)+P(y,x))^{n_x}.(P(x,y)+P(y,y))^{n_y} = (r + s)^{2N} = 1$$

B_{01} corresponds to $i_x = 0$ and $i_y = 1$, so:

$$B_{01} = (D_{yx}+D_{yy}).(P(x,x)+P(y,x))^{n_x}.(P(x,y)+P(y,y))^{n_y}$$

$$= 2rs.n_x + (r^2 +s^2).n_y$$

Finally, B_{02} corresponds to $i_x=0$ and $i_y=2$, so:

$$B_{02} = (D_{yx}D_{yx}+2D_{yx}D_{yy}+D_{yy}D_{yy}).(P(x,x)+P(y,x))^{n_x}.(P(x,y)+P(y,y))$$

$$= 2Nrs(r^2 + s^2) + (n_y + 2(n_x - n_y)rs)^2$$

Therefore $\emptyset^* = 4r^2s^2(n_x-n_y)^2/N^2 - 4r^2s^2/N^2 + 2rs/N$ for a given pair of terms t_j, t_k. Thus for any pair of terms in the vocabulary, $\emptyset^*$ must lie between K and $K+4r^2s^2$ where $K = (2rs/N)(1 - 2rs/N)$. $\emptyset^*$ is zero for this similarity function, as expected, when $r=0$ or $r=1$, corresponding respectively to the cases where there are no errors at all, and where there are errors in every term assignment.

The error analysis of the similarity function $s_1(t_j,t_k)$, that is the Tanimoto function, will now be reconsidered. Suppose that f and f_S are the values of $s_1(t_j,t_k)$ before and after scrambling. Then, as before,

f and f_S are given by $f = n_d/(N-n_a)$ and $f_S = m_d/(N-m_a)$.

Again, the mean square deviation will be considered.

Thus $\emptyset = (f - f_S)^2$

$$= \left(\frac{n_d}{N - n_a} - \frac{m_d}{N} \cdot \left(1 - \frac{m_a}{N} \right)^{-1} \right)^2$$

Providing m_a/N is sufficiently small, the binomial series expansion of $\emptyset$ may be truncated and expressed as a polynomial in m_a and m_d. With this provision, the value of $\emptyset^*$ may be obtained using the second of the two general expressions for $\emptyset^*$ given above. However, if m_a is to be small compared with N then, assuming r to be small as well, the proportion of documents of the collection to which may term is assigned must be large. This is unlikely to be true in practice. It is more likely that the number of documents to which a particular term has been assigned will vary considerably from term to term. It is therefore unfeasible to treat the Tanimoto function in this particular way, at least for the type of data which is considered here.

6 Conclusion

The methods derived above for obtaining an estimate $\emptyset^*$ for a statistical function, $\emptyset$, of the similarity between pairs of terms, when the assignment of terms to documents is susceptible to errors with probability r, have been shown to be applicable in certain circumstances. No estimate, however, has been obtained for the asymmetric Tanimoto function for the reasons alraady given. However, an estimate has been obtained for the symmetrical Tanimoto function, and therefore all similarity functions expressible as polynomials in m_x and m_y, since the coefficients B_i do not depend on the particular function under examination.

The analysis given is applicable outside the field of information retrieval which has been considered here. The analysis is applicable, in principle, to classification techniques, whether hierarchical or non-hierarchical, which utilize a similarity matrix giving the similarities between the objects of the population, provided that the objects are described in terms of mutually independent binary attributea and provided that these attributes are susceptible to errors independently and equiprobably, with an assignable probability r.

References.

1 Williams, W. T., and Dale, M. B. Fundamental problems in Numerical Taxonomy. Advances in Botanical Research, 2, 1965, pp. 35-68.

2 Sokal, R. R., and Sneath, Ph H. A. Principles of Numerical Taxonomy. Freeman, 1963, 359 pp.

3 Parker-Rhodes, A. F. Statistical aspects of fungus forays. Transactions of the British Mycological Society, 38, 1955, pp. 283-290.

4 Parker-Rhodes, A. F., and Jackson, D. M. Automatic classification in the ecology of the Higher Fungi. In: Numerical Taxonomy (Ed. Cole, A. J.), Academic Press. 1969, 320 pp.

5. Sutcliffe, J. P. A probability model for errors of classification. I. General considerations. Psychometrika, 30, 1965, pp. 73-96.

6. Sutcliffe, J. P. A probability model for errors of clsssification. II. Particular cases. Psychometrika, 30, 1965, pp. 129-155.

7 Gotlieb, C. C., and Kumar, S. Semantic clustering of index terms. Journal of ACM, 15, 1968, pp. 493-513.

8 Needham, K. Synonymy and semantic classification. Ph.D. Thesis, University of Cambridge, England, 1964.

9 Lewis, P. A. W., Baxendale, P. B., and Bennett, J. L. Statistical discriminstion of the synonymy/antonymy relationship between words. Journal of ACM, 14, 1967, pp. 20-44.

10 Sebestyen, G. S. Recognition of membership in classes. IRE Transsctions on Information Theory, 1961, pp. 44-50.

11 Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 12, 1967, pp. 241-254.

12 Jardine, C. J., Jardine, N. and Sibson, R. The structure and construction of taxonomic hierarchies. Mathematical Biosciences, 1, 1967, pp. 173-179.

13 Pearson, K. On coefficients of racial likeness. Biometrika, 18, 1926, pp. 105-117.

14 Mahalanobis, P. C. On the generalized distance in statistics. Proceedings of the National Inatitute of Science of India, 12, 1936, pp. 49-55.

15 Rao, C. R. Advanced statisticsl methods in biometric research. John Wiley and Sons, New York, 1952.

16 Tanimoto, T. T. An elementary theory of classification and prediction. IBM report, York Town Heights, N.Y., 1958.

17 Benzéin , J. P. Sur les algorithmes de classification. 1965. 8p. Texte multigraphié.

18 de la Véga ,W. F. Techniques de classification automatique utilisant un indice de ressemblance. Revue française de Sociologie, 8, 1967, pp. 506-520.

19 Kendall, M. G. The Advanced Theory of Statistics. Vol. 1, C. Griffin and Co. Ltd., London, 1943.

20 Jackson, D. M. Comparison of classifications. In: Numerical Taxonomy (Ed. Cole. A. J.), Academic Press. 1969, 320 pp.

21 Barton, D. E. The matching distributions: Poisson limiting forms and derived methods of approximation. Journal of the Royal Statistical Society, Series B, 20, 1958 , pp. 73-92.

22 Levene, H. On a matching problem arising in genetics. Annals of Mathematical Statistics, 20, 1949, pp. 91-94.