

Measuring the degree of similarity between objects in text retrieval systems

DAVID ELLIS,
JONATHAN FURNER-HINES and
PETER WILLETT

Department of Information Studies,
University of Sheffield

Describes the use of a variety of similarity coefficients in the measurement of the degree of similarity between objects that contain textual information, such as documents, paragraphs, index terms or queries. The work is intended as a preliminary to future investigation of the calculations involved in measuring the degree of similarity between structured objects that may be represented in graph-theoretic forms. Discusses the role of similarity coefficients in text retrieval in terms of: document-query similarity; document-document similarity; co-citation analysis; term-term similarity; and the similarity between sets of judgements, such as relevance judgements. Describes several methods for expressing the formulae used to define similarity coefficients and compares their attributes. Concludes with details of the characteristics of similarity coefficients: equivalence and monotonicity; consideration of negative matches; geometric analyses; and the meaning of correlation coefficients.

INTRODUCTION

Our aim in this paper is to describe the use of a variety of similarity coefficients in the measurement of the degree of similarity between objects that contain textual information, such as documents, paragraphs, index terms or queries. This work is intended as a preliminary to future investigation of the calculations involved in measuring the degree of similarity between structured objects of a certain type, *viz.* those that may be represented in graph-theoretic form. An example of such a structure is the set of links connecting the nodes in a hypertext database, and we intend to go on to study the similarity between hypertext databases that share a common set of nodes, but whose link-sets have been manually inserted by different people (Ellis, Furner-Hines and Willett, 1993a & b).

The measurement of the degree of similarity between any two objects involves three steps (Sokal and Sneath, 1963). First, a set should be selected of those *attributes* of the objects whose values are to be used to characterize the objects for the purposes of their comparison. Secondly (and optionally), a *weighting scheme* may be implemented that emphasizes certain attributes according to any differences in their relative significance. We shall not consider weighting schemes in this paper. Thirdly, a value for a *measure* that represents the degree of similarity between the objects may be derived from analysis of their various attribute-values. In the rest of this section, we discuss the use of vectors to represent, in the first step, the objects that we wish to compare; and we introduce the use of arithmetic coefficients to measure, in the third step, the degree of similarity between vector representations of objects. We go on to summarize the role of similarity coefficients in text retrieval; then to analyse the composition of a variety of similarity coefficients; and finally we draw a number of conclusions about the validity of using certain coefficients in work of this kind.

Representing objects by vectors

Any particular thing that *has* properties may be called an *object*. Any object that contains information, or that may be used as a source of information, may be called an *information object*. Any particular thing that *is* a property of an object may be called an *attribute-value*. Just as each object is an instance of a class of objects, each attribute-value of an object is an instance of a general type of attribute-value. We may say that each attribute-value is an instance of an

attribute. Whether or not we consider a particular thing to be an object or an attribute-value depends on our subjective interests: if we are interested in the properties of that thing, then we can call it an object; on the other hand, we might be more interested in that thing as a property of another thing, and consider it as an attribute-value.

An object may be presented or described by listing its set of attribute-values. All the similarity coefficients that we shall consider require each of the sets of attribute-values by which objects are characterized to be expressed in the form of a vector. In this way, an object O_k may be characterized by the vector \vec{X}_k such that

$$\vec{X}_k = \{x_{1k}, x_{2k}, x_{3k}, \dots, x_{nk}\},$$

where there are n attributes, and x_{jk} is the value of attribute A_j for object O_k .

The use of similarity coefficients

If the objects to be compared are each characterized solely by a single attribute whose values take the form of single numbers, it is a simple task to calculate the numerical difference between two such values, and thus to derive a value for a primitive measure of similarity: the smaller the difference, the more similar are the objects. If, however, the objects are each characterized by an attribute whose values take the form of sequences of numbers, or by more than one attribute, then one of a number of more complex arithmetic functions must be used in order to calculate a value for a measure of the agreement between the sets of attribute-values for a pair of objects. Since the late nineteenth century functions of this kind, known as *similarity coefficients*, have been described in all fields in which *classification* (the ordering of objects into groups or sets on the basis of relationships of contiguity or similarity that exist between them) is an important endeavour – notably in data analysis (Anderberg, 1973; Gordon, 1981), numerical taxonomy (i.e., the classification of biological species) (Sneath and Sokal, 1973; Sokal and Sneath, 1963), ecology (Clifford and Stephenson, 1975), and both chemical (Johnson, 1989; Willett, 1987) and textual (Salton and McGill, 1983; van Rijsbergen, 1979) information retrieval. Lack of communication between these fields, however, has resulted in much duplication of effort.

THE ROLE OF SIMILARITY COEFFICIENTS IN TEXT RETRIEVAL

The vector method of representing objects has a long history of application in text retrieval (Becker and Hayes, 1963; Salton, 1968; Tanimoto, 1958), and the purposes for which measures of similarity have been used in this field are various. We can classify these uses according to the nature of the objects whose vector representations are compared.

First, we may calculate a value for the degree of

similarity between any *document* in a database and a *query* (i.e., the representation of a user's particular information need); secondly, we may calculate a value for the degree of similarity between each of any pair of documents in a database. In both of these cases, each document O_k and each query or second document O_l is characterized by a vector \vec{X}_k or \vec{X}_l of n elements (where n is the number of different terms in the vocabulary used to index the documents), each of whose elements x_{jk} or x_{jl} is a value representing either:

- the *presence or absence* of term A_j in the set of terms that is used to represent document O_k or query O_l , in which case the data are in binary form; or
- the *weight* of term A_j in the set of terms that is used to represent document O_k or query O_l , in which case the data are in non-binary form.

(Various schemes exist for *term weighting* (Salton and Buckley, 1988; Sparck Jones, 1973; Yu, Lam and Salton, 1982): the formula that each uses to calculate weights is made up of some combination of functions based on term frequency. These functions include the number of occurrences of a particular term in a particular document or query, the number of occurrences of a particular term in the document collection as a whole or some subset of it, the number of occurrences of all terms in a particular document or query, and the number of documents in the collection in which a particular term is included.)

A third instance of vector comparison in text retrieval differs from those described above, in that the vectors which are compared represent index terms rather than documents or queries. Each index term O_k is characterized by a vector \vec{X}_k , each of whose elements x_{jk} is a value representing either:

- the *presence or absence* of document A_j in the set of documents that are indexed by term O_k , in which case the data are in binary form; or
- the *frequency* with which document A_j includes term O_k , in which case the data are in non-binary form.

In a final set of cases, vectors are compared in order to calculate values for the degree of similarity between sets of judgements that are made in the course of the retrieval process, either by humans or by systems.

We shall deal with each of these cases in turn.

Document–query similarity

A query vector may be compared with each document vector in order to calculate values for the degree of similarity between the query and the documents. This operation stands at the heart of the *vector processing model* of document retrieval (Salton, 1971; Salton and McGill, 1983), often known (perhaps somewhat misleadingly) as the *vector space model* (Salton, Wong and Yang, 1975;

Salton, 1991). This belongs to a class of retrieval models that describe the techniques of *best-match* searching (also known as *ranked-output* or *similarity* searching) (Belkin and Croft, 1987). Other members of this class include the *probabilistic* models of Maron and Kuhns (1960), Robertson and Sparck Jones (1976), and Robertson, Maron and Cooper (1982), *inter alia*.

According to the usual formulation, best-match retrieval techniques allow the documents in a collection to be *ranked* according to their *relevance* (or, more accurately, *probability of relevance*) to the user, and any number of documents may be retrieved by taking that number from the top of the ranking order. In the case of the vector processing model, documents are ranked in order of relevance on the basis of the values calculated for the degree of similarity between query and document vectors. (This is in marked contrast to mechanisms based on the *Boolean* (or *exact-match*) retrieval model, which return unordered lists of those documents that satisfy the logical requirements of queries made up of terms and Boolean operators.) Any mechanism of the vector processing type is based on a *ranking algorithm* composed of five elements (McGill, Koll and Noreault, 1979; Sager and Lockemann, 1976):

- a method of representing documents;
- a method of representing queries;
- a weighting scheme for terms used in document representations
- a weighting scheme for terms used in query representations;
- a formula for a measure of similarity.

It will be seen that the specification of such an algorithm is essentially equivalent to that of the three-step model, described in the first section, of the measurement of the degree of similarity between any two objects.

Wong *et al.* (Bollmann and Wong, 1987; Wang, Wong and Yao, 1992; Wong and Yao, 1990; Wong, Yao and Bollmann, 1988; Wong *et al.*, 1991) develop the notion of *user preference* in place of that of *relevance*, defining a relational system consisting of (i) a set of documents, and (ii) a binary relation of preference, given by \succ , such that $C_k \succ C_l$ means 'the user prefers document C_k to document C_l (with respect to an individual query)'. They suggest that the aim of any ranking algorithm is to map this relational system onto another consisting of (i) the set of real numbers, and (ii) the binary relation $>$, meaning 'is greater than'. It is this mapping that is generated by the use of similarity coefficients.

An equivalent operation to that described above may be carried out in cases where the query is put to a single full-text document (or to a set of such documents), rather than to a set of references to documents, with the aim of identifying the individual paragraphs within the document(s) that are most likely to be of relevance to the

user. The query vector is compared with the vectors representing individual paragraphs, and the paragraphs can then be ranked on the basis of the results (Al-Hawamdeh and Willett, 1989; Salton and Buckley, 1991; Salton, Buckley and Allan, 1992; Tenopir, 1988). This type of operation may also be used as a means for initial access to full-text documents structured in the form of hypertext networks, where the nodes of the network correspond to the paragraphs of the document(s). Values are calculated which represent the similarity between the query and each paragraph, and the paragraph that produces the highest value is displayed to the user. The user may then retrieve other nodes by browsing - i.e., by following predefined links between nodes, rather than by constructing further queries (Al-Hawamdeh, Smith and Willett, 1991; Frisse, 1988; Smeaton, 1992).

Document-document similarity

Two document vectors may be compared in order to calculate a value for the degree of similarity between the documents they represent; the documents may then be grouped into clusters on the basis of these values using one of a variety of automatic *document clustering* techniques (van Rijsbergen, 1979; Willett, 1988). In the subsequent retrieval process, a query may be compared firstly with a single representative of each cluster of documents, rather than with every individual document, and then with only those individual documents whose cluster representatives exhibit a high degree of similarity with the query, thus improving the efficiency (and possibly the effectiveness (Jardine and van Rijsbergen, 1971)) of the process. Here it is assumed that documents which are clustered are likely to relate to the same topic, and therefore to have similar probabilities of relevance to the user.

An equivalent operation may be carried out in full-text databases, where paragraphs rather than documents are the objects to be clustered. Values calculated for the degree of similarity between one paragraph and another may also be used in the automatic definition of the links between nodes in a hypertext network. Two nodes may be linked according to the presence or absence of a relationship between the paragraphs they represent, and such a relationship may be identified by analysis of the paragraphs' similarity (Bernstein, 1990; Crouch, Crouch and Andreas, 1989). This method of creating hypertext links automatically need not be followed *before* the user attempts to retrieve information from the network, as it is necessary to do when defining links manually, but can be performed dynamically *during* the search (Andersen, Nielsen and Rasmussen, 1989; Li, Davis and Hall, 1993).

Co-citation analysis

A document may also be characterized by the set of documents that cite it, rather than by a set of index terms. In this case, each element x_{jk} of the vector \bar{X}_k represents the

presence/absence of a citation to document \mathcal{O}_k in document A_j . Values for the degree of similarity between vectors of this type may then be calculated, each value representing the number of times two documents are jointly cited by documents published at a later date. *Document co-citation analysis* involves several operations: graphs may be constructed in which nodes representing documents are linked if the similarity value is greater than a particular threshold value; clustering techniques may be used to group sets of frequently co-cited documents together; and such clusters may be represented in diagrammatic form using multi-dimensional scaling methods, in a process known as *cluster mapping* (Garfield, 1979; Small, 1973; Small and Sweeney, 1985). It is possible to identify from cluster maps *core* and *scatter* – i.e., documents that are, respectively, more central and more peripheral to the collection or field under analysis.

Objects other than documents may be compared in a similar way. White and McCain (1989) discuss several types of co-citation analysis in the course of their comprehensive review of bibliometrics. In *author co-citation analysis*, each element of an author's vector represents the presence/absence of a citation to that author in another document. By comparing these vectors, cluster maps of authors may be derived whose two dimensions are often respectively interpreted as reflecting differences in subject matter and differences in style, approach or treatment (White, 1990; White and Griffith, 1981). In studies of *journal networks*, each element of a journal's vector represents the presence/absence of a citation in that journal to another journal (note the difference here – this type of work is more properly known as journal *cross-citation analysis*). By comparing these vectors, journals may be clustered, either within or between disciplines (Leydesdorff, 1987). Finally, in *co-citation context analysis* it is documents again that are represented by vectors, but each element of a document's vector represents the presence/absence of a citation to that document in a particular *section* or *paragraph* of another document (Small, 1982).

Term-term similarity

Vectors may be compared in order to calculate a value for the degree of similarity between the index terms they represent. Such values represent measurements of the statistical co-occurrence of words in the documents of the collection. The terms may then be grouped into clusters on the basis of these values in a process known variously as automatic *term clustering*, *term association* or *keyword classification* (Jones and Curtice, 1967; Lesk, 1969; Sparck Jones, 1971). In this way, a type of *thesaurus* may be constructed automatically, in which terms are classified on the assumption that words which tend to co-occur in the same documents are words that relate to the same topic. In the subsequent retrieval process, the attributes whose

values describe queries and documents may be term clusters rather than individuals. Each value in the resultant vectors represents the presence or absence in the query or document of any of the terms in a particular cluster, improving the recall and possibly the effectiveness of the search: the method makes possible the retrieval of documents that are relevant to the query even though they are not indexed by the terms used in the query. The use of co-occurrence data in the clustering of terms is central to probabilistic models of information retrieval that incorporate term dependency (Van Rijsbergen, 1977).

In another context, the vector characterizing an index term may be made up of a set of elements, each of which represents the presence or absence in that term of a particular *digram* or *trigram* (respectively, a pair or a triple of adjacent characters). The comparison of pairs of such vectors in order to calculate a value for the degree of similarity between them is a method used for *spelling correction* and *term conflation* (Angell, Freund and Willett, 1983; Freund and Willett, 1982). It should be noted, however, that many approaches to the calculation of word-word similarities, such as those of *phonetic coding* (Rogers and Willett, 1991) and *dynamic programming* (Robertson and Willett, 1992), are not based on the comparison of vectors.

The similarity between sets of judgements

Inter-indexer consistency

Comparison of vectors is the method most commonly used to derive values that represent the extent to which agreement exists on the terms to be used to index a document – in other words, to measure the degree of consistency in the work of different indexers (Rolling, 1981). The exact composition of the vectors compared, however, has varied from one study of inter-indexer consistency to another (Leonard, 1977).

If the object \mathcal{O}_k represented by the vector \vec{X}_k is considered to be a *document* as indexed by a particular indexer, then each element x_{jk} of the vector represents the presence/absence, or perhaps the weight, of a particular term in the set of index terms assigned to that document by that indexer (see, for example, Hooper (1965)). One such vector may be compared with a second, the latter representing the set of terms assigned to the same document by a different indexer. Values representing the degree of similarity between the two vectors may be calculated for each document in a collection, and a measure of consistency in the work of the two indexers may be derived by taking the average of these similarity values.

If, on the other hand, the object represented by a vector is considered to be a *term* as assigned by a particular indexer, then each element of the vector represents the presence/absence of a particular document in the set of

documents to which that term has been assigned by that indexer (see, for example, King and Bryant (1971)). One such vector may be compared with a second, the latter representing the set of documents to which the same term has been assigned by a different indexer. Values representing the degree of similarity between the two vectors may be calculated for each term in a vocabulary, and a measure of consistency in the work of the two indexers may be derived by taking the average of these similarity values.

Alternative methods exist for the calculation of values of inter-indexer consistency that do not involve the comparison of vectors: Cooper (1969), for instance, suggests that a measurement of the consistency in the assignment of a particular term by a number of indexers may be derived from the ratio of (i) the absolute difference between the number of indexers who assign that term to a particular document and the number of indexers who do not; to (ii) the total number of indexers. Methods for the calculation of values representing the consistency in the work of more than two indexers, whether they involve vector comparison or not, range from the simple (such as Cooper's method) to the complex (see, for example, Iivonen (1990)).

Studies of inter-indexer consistency were especially popular in the 1960s. Their principal conclusions were that recorded levels of consistency display marked variation, and that high levels are rarely achieved. Although interest has waned in subsequent decades (Leonard (1977) and Markey (1984) provide historical surveys), the results of this body of work have commonly been considered significant because of the assumption that they are predictive of the levels of retrieval effectiveness that may be attained by the systems studied (Leonard, 1975). With a view to softening some prominent voices of dissent (e.g., Cooper, 1969), recent work has attempted to clarify the relationship between inter-indexer consistency and retrieval effectiveness (Gomez, Lochbaum and Landauer, 1990).

Relevance judgements

A comparison of another type may be made between vectors that each represent the decisions made by a different judge about the relevance of the individual documents in a collection to the same query. In this case, each element x_{jk} of the vector \bar{X}_k represents a judgement by the judge C_k about document A_j , either indicating whether the document is relevant or non-relevant in respect of a particular query (in which case the data are binary), or indicating the weight of that document's relevance to the query (in which case the data are non-binary). Values for the degree of similarity between pairs of vectors of this type can be used as measurements of the level of inter-judge consistency (Lesk and Salton, 1968). Alternatively, when a vector representing the judgements of a human is

compared with one that represents the 'judgements' of a retrieval system (i.e., that indicates either the retrieval/non-retrieval of each document, or the weight of relevance of each document, as computed by the system), values for the degree of similarity between the vectors can provide a single measure of system performance that combines the two conventional measures of recall and precision (Cleverdon, Mills and Keen, 1966; Gebhardt, 1975; Heine, 1973; van Rijsbergen, 1979).¹

Relevance judgements are widely used in the calculation of two standard measures of the performance of text retrieval systems. We may divide the documents in a collection into four sets according to (i) the judgements of a human as to whether they should be retrieved or not retrieved in response to a particular query (i.e., their relevance or non-relevance), and (ii) their retrieval or non-retrieval by the system. We may define a, b, c, d as the number of documents in each set as follows: a is the number of documents that are judged by the human to be relevant, and are retrieved by the system; b is the number of documents that are not judged by the human to be relevant, but are retrieved by the system; c is the number of documents that are judged by the human to be relevant, but are not retrieved by the system; and d is the number of documents that are not judged by the human to be relevant, and are not retrieved by the system. We may calculate *recall* by dividing a by $a + c$, and *precision* by dividing a by $a + b$. The conclusions of studies of inter-judge consistency – that recorded levels of consistency display marked variation, and that high levels of consistency are rarely achieved – might appear to invalidate these methods of evaluating system performance. There is very little empirical evidence, however, to support the hypothesis that significant differences in the performance of different retrieval systems may be attributed to differences in the relevance judgements used in experiments (Burgin, 1992; Lesk and Salton, 1968).

It has often been noted that the judgement of relevance is dependent not just on the perceived relationship between a document and a query, but on many other external variables, such as the method and order in which documents are represented, the expertise or experience of the judge, the definition of relevance used, and the context of an individual document within the collection (Cuadra and Katter, 1967; Eisenberg, 1988; Rees and Schultz, 1967). To clarify the relationship between variation in relevance judgements and variation in system performance, further experiments are necessary in which variation in these other factors is controlled (Burgin, 1992).

THE COMPOSITION OF SIMILARITY COEFFICIENTS

Methods of expressing the formulae of similarity coefficients

A particular coefficient, or the particular way in which its

formula is expressed, may be suitable for use only with certain types of data. In this context, *binary* (or *two-state*) data may be distinguished from *non-binary* (or *multistate*) data; *quantitative* non-binary data, which can be ordered, may be distinguished from *qualitative* non-binary data, which cannot; and quantitative, non-binary data may be ordered on a *continuous* or on a *discrete* scale, or may be *ranked* (see Figure 1). We shall limit our discussion in this paper to coefficients that are intended for use with binary data or non-ranked, quantitative, non-binary data.

Several methods of expressing the formulae of similarity coefficients have been used in the literature.

Formulae for use with non-binary data

The most common methods belong to the class of those which, like the notation introduced in the previous section, are based on the representation of sets of attribute-values as *vectors*. Certain others, however, are specifically used to express more clearly the formulae of similarity coefficients when the data under analysis is in binary form.

Formulae based on the 2 × 2 contingency table

The first of these that we shall consider is based on the simplest manifestation of an analytical tool known as the *p × p contingency table*, where *p* is equal to the number of

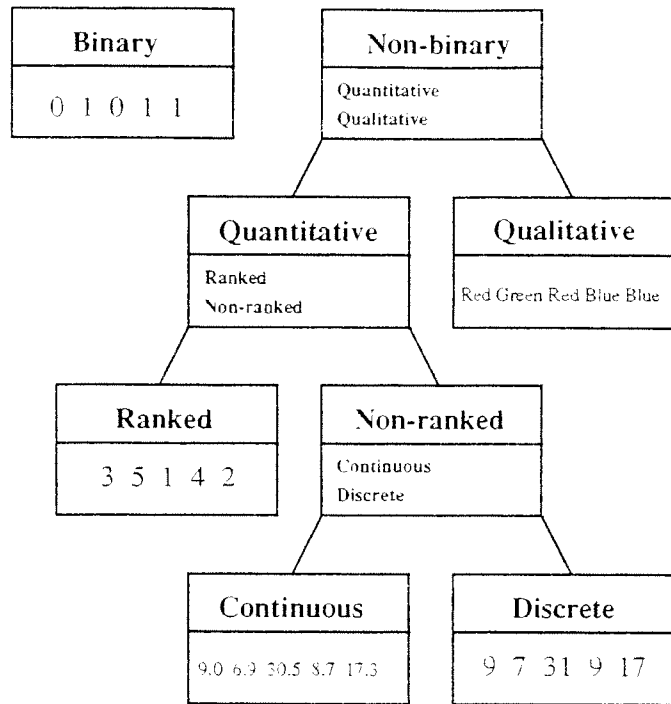


Figure 1 Examples of data types – five-element vectors

possible values (or ranges of values) that the attributes of an object can take.² Contingency tables allow us to compare the distribution of the attribute-values that represent pairs of objects.³ The simplest case is where attribute-values may be either 0 or 1, i.e. where the data is binary and hence $p = 2$. In the 2×2 table, there are four elements, as shown in Figure 2.

	l_1 ($x_{i1} = 1$)	l_2 ($x_{i1} = 0$)
k_1 ($x_{j1} = 1$)	<i>a</i>	<i>b</i>
k_2 ($x_{j1} = 0$)	<i>c</i>	<i>d</i>

Figure 2 The 2 × 2 contingency table

We may express each of the values in Figure 2 by a letter *a*, *b*, *c*, *d* as follows: *a* is equal to the number of attributes whose value both in \bar{X}_k and in \bar{X}_l is 1; *b* is equal to the number of attributes whose value in \bar{X}_k is 1 and in \bar{X}_l is 0; *c* is equal to the number of attributes whose value in \bar{X}_k is 0 and in \bar{X}_l is 1; and *d* is equal to the number of attributes whose value both in \bar{X}_k and in \bar{X}_l is 0. The sum of all these values ($a + b + c + d$) is equal to the number of attributes, *n*, of each object.

Formulae based on set theory

A second alternative is to use *set-theoretic* notation. We may define X_k as the set of all elements x_{jk} in vector \bar{X}_k whose value is 1, and similarly X_l as the set of all elements x_{jl} in vector \bar{X}_l whose value is 1. We may define $|X_k|$, the *cardinality* of X_k , as the number of elements in set X_k , $|X_l|$ as the number of elements in set X_l , $|X_k \cap X_l|$ as the number of elements common to both X_k and X_l , and $|X_k \cup X_l|$ as the number of elements in X_k or X_l (or both). $|X_k \cap X_l|$ in this notation is equivalent to *a* in the notation described above, and $|X_k \cup X_l|$ is equivalent to $a + b + c + d = n$. $|X_k|$ is equivalent to $a + b$, and $|X_l|$ is equivalent to $a + c$.

A classification of similarity coefficients

From the vectors \bar{X}_k and \bar{X}_l , we can derive two basic functions:

1. the *difference sum*, i.e. the sum of the *absolute differences* between the corresponding elements in the two vectors, given by

$$\sum_{j=1}^n |x_{jk} - x_{jl}|;$$

2. and the *inner product* (ordinary vector product, dot product, or scalar product), i.e. the sum of the *products* of corresponding elements, given by

$$\sum_{j=1}^n x_{jk} \cdot x_{jl}.$$

When the data under consideration are in binary form, the difference sum reduces to $b + c$, or $|\bar{X}_k| + |\bar{X}_l| - 2|\bar{X}_k \cap \bar{X}_l|$, while the inner product reduces to a , or $|\bar{X}_k \cap \bar{X}_l|$.

It is possible for either of these two functions to be used on their own as a crude measure of the similarity between two vectors; their values, however, vary in direct proportion with the value of n , the number of pairs of attributes being compared (and hence the number of distances being summed). Moreover, where the attribute-values under consideration are in non-binary form, these functions' values also vary in accordance with the range, or scale, of the attribute-values. In other words, the functions do not have an upper limit because they have not been normalized. Values of the difference sum may vary from 0 (indicating complete similarity) to an indefinitely large number; similarly, values of the inner product may vary from 0 (indicating no similarity) to another indefinitely large number. The characteristic by which most similarity coefficients may be distinguished is the composition of the factor by which they require the difference sum or the inner product to be multiplied. This normalization ensures that the values of the coefficients remain within a specific range, such as that bounded by 0 and 1, or by -1 and +1.

Following Sneath and Sokal (1973; Sokal and Sneath, 1963) (who give much further information on the derivation and formulation of all types of coefficient) we may identify three classes of similarity coefficient. *Distance coefficients* are based on the difference sum, and their values vary in *inverse* proportion with degree of similarity, so that greater similarity is indicated by lower values. *Association coefficients* are based on the inner product, and their values vary in *direct* proportion with degree of similarity, so that greater similarity is indicated by higher values. *Correlation coefficients* are based on a third, more complex function: the sum of the product of the differences between each attribute-value and the mean of all attribute-

values for each of the two vectors. Values of these generally vary from +1 (indicating that any change in the attributes of one object would be accompanied by an identical change in the attributes of the other) to -1 (indicating that any change in one would be accompanied by an equal and opposite change in the other).

Before describing a number of members of each of these classes, and considering their application to both non-binary and binary data, we shall discuss a few of the vector functions that are often used in the composition of normalizing factors. A simple factor is $1/n$, which produces the *mean Manhattan* metric (mean difference sum) when used to multiply the difference sum, and the coefficient of *Russell/Rao* when used to multiply the inner product (see below for formal definitions of these coefficients). The use of such a factor keeps values within appropriate bounds when data is binary, but more complex factors are needed to achieve the same objective for non-binary data. Functions that are commonly used in the composition of such factors are:

1.
$$\sum_{j=1}^n x_{jk}$$

and

$$\sum_{j=1}^n x_{jl};$$

respectively, the sum of all elements in the vector \bar{X}_k and the sum of all elements in the vector \bar{X}_l , which for binary data both reduce to the number of non-zero elements in the appropriate vector, i.e., $a + b$ or $|\bar{X}_k|$, or $a + c$ or $|\bar{X}_l|$;

2.
$$\sum_{j=1}^n (x_{jk})^2$$

and

$$\sum_{j=1}^n (x_{jl})^2;$$

respectively, the sum of the squares of all elements in the vector \bar{X}_k and the sum of the squares of all elements in vector \bar{X}_l , which for binary data again both reduce to the number of non-zero elements in the appropriate vector, as above;

3.
$$\sqrt{\sum_{j=1}^n (x_{jk})^2}$$

and

$$\sqrt{\sum_{j=1}^n (x_{jl})^2};$$

respectively, the Euclidean lengths of the vectors \bar{X}_k and \bar{X}_l in n -dimensional space (see below for further discussion), which for binary data reduce to the square root of the number of non-zero elements in the appropriate vector.

The discussion that follows, of examples of each type of coefficient, draws on the information contained in two surveys (Hubálek, 1982; McGill, Koll and Noreault, 1979). These provide comprehensive listings of coefficients used in the fields of biology and information retrieval, respectively. Table 1 lists the common and alternative names of the coefficients we discuss, together with the identifiers assigned to them both in this paper and in those of Hubálek and McGill *et al.*

Table 1 Identification of similarity coefficients

ID	Common name	Other names	Hubálek ID	McGill ID
D1	Mean Manhattan	Mean difference sum	-	57
D2	Mean Euclidean	Average distance	-	59
D3	Mean squared Euclidean	-	-	-
D4	Mean Canberra	-	-	63 ^a
D5	Divergence	-	-	64 ^b
D6	Bray/Curtis	Non-metric	-	20, 65
<hr/>				
A1	Jaccard	Tanimoto, Doyle, Hooper, Parker-Rhodes/Needham	A ₄	36, 37 / 33, 34, 35
A2	Dice	Czekanowski, Sørensen	A ₅	15
A3	Russell/Rao	-	A ₁₄	3 / 21
A4	Sokal/Sneath (1)	-	A ₆	17
A5	Kulczyński (1)	-	A ₃	16
A6	Simple matching	-	A ₂₀	7, 8, 9
A7	Hamann	-	A ₂₄	14, 66
A8	Sokal/Sneath (2)	-	A ₂₂	-
A9	Rogers/Tanimoto	-	A ₂₃	67
A10	Sokal/Sneath (3)	-	A ₁₉	-
A11	Baroni-Urbani/Buser	-	A ₃₂	-
A12	Ochiai	Cosine	A ₁₁	1 / 19
A13	Kulczyński (2)	-	A ₇	18
A14	Forbes	Kochen/Wong	A ₄₀	24
A15	Fossum	-	-	26
A16	Simpson	Overlap, Asymmetric	A ₂	27
<hr/>				
C1	Pearson	Phi	A ₃₀	4 / 5
C2	Yule	Maron/Kuhns	A ₃₆	12
C3	McConnaughey	-	A ₁₀	-
C4	Stiles	-	-	53 ^c
C5	Dennis	-	-	13

D_n indicates a distance coefficient, A_n an association coefficient, and C_n a correlation coefficient

^a Formula for the Canberra metric itself, not its mean

^b Formula for the square of the divergence coefficient

^c Formula does not take a logarithmic value

Distance coefficients

Distance coefficients measure the dissimilarity of pairs of objects rather than their similarity. Their values may be

equated with measurements of the *geometric* distances between pairs of vectors, each of whose n elements are plotted on a separate dimension in a space of n dimensions. Although we cannot visualize the geometry of a space of more than three dimensions (a *hyperspace*), it can be shown algebraically that most of the geometric theorems which apply to three-dimensional space apply also to n -dimensional space, and hence that it is valid to calculate geometric distances between objects in hyperspace.

For this validity to be preserved, distance coefficients should have the properties of *metrics* – i.e., functions that satisfy four specific axioms over any set of objects. For a similarity coefficient S calculated from the attribute-values in a pair of vectors \bar{X}_k and \bar{X}_l , these axioms may be expressed in the following forms:

1. $S(\bar{X}_k, \bar{X}_l) \geq 0$, and $S(\bar{X}_k, \bar{X}_k) = S(\bar{X}_l, \bar{X}_l) = 0$;
2. $S(\bar{X}_k, \bar{X}_l) = S(\bar{X}_l, \bar{X}_k)$;
3. $S(\bar{X}_k, \bar{X}_l) \leq S(\bar{X}_k, \bar{X}_m) + S(\bar{X}_m, \bar{X}_l)$; and
4. if $\bar{X}_k \neq \bar{X}_l$, then $S(\bar{X}_k, \bar{X}_l) > 0$.

Functions that satisfy only the first three of these axioms are *pseudometric*, while functions that do not satisfy the third axiom (the axiom of *triangle inequality*) are non-metric.

It should be noted, however, that the validity of this interpretation of the meaning of distance, though widely accepted in the information retrieval literature, does not strictly extend to applications where the dimensions under consideration are not *orthogonal* to each other – in other words, where the attribute-values of an object are not *independent* of each other. The indexing of a document by one term is not, in practice, an operation that is independent of its indexing by another term, and the very idea of term-term association undermines any suggestion otherwise. It is therefore unclear whether certain axioms to do with linear independence that should necessarily be obeyed by the elements of a vector space are in fact obeyed by the elements of a text retrieval system (Raghavan and Wong, 1986; Wong and Raghavan, 1984). We shall proceed by describing the form of a number of distance coefficients, noting whether they are metric or not, with the caveat that it may well be inappropriate to emphasize their metric qualities in the particular context of text retrieval.

A series of metric distance coefficients known as the *Minkowski metrics* may be calculated using the general formula (Jardine and Sibson, 1971)

$$\Delta_\mu = \left(\sum_{j=1}^n |x_{jk} - x_{jl}|^{\frac{1}{\mu}} \right)^\mu$$

Δ_1 is called the *Manhattan* or *city block* metric, or the *difference sum* as defined in the previous section, and is given by

$$\Delta_1 = \sum_{j=1}^n |x_{jk} - x_{jl}|;$$

$\Delta_{\frac{1}{2}}$ is called the *Euclidean distance* or *taxonomic distance* metric, and is given by

$$\Delta_{\frac{1}{2}} = \sqrt{\sum_{j=1}^n (x_{jk} - x_{jl})^2}.$$

The *squared Euclidean distance*, $\Delta_{\frac{1}{2}}^2$, is given by

$$\Delta_{\frac{1}{2}}^2 = \sum_{j=1}^n (x_{jk} - x_{jl})^2.$$

The earliest use of geometric distances in the formulae of similarity coefficients is often attributed to Heincke, writing at the end of the nineteenth century (Heincke, 1898). We may define five of the various normalized forms that have been used in subsequent years:

1. the *mean Manhattan metric* or *mean difference sum*, given by $\bar{\Delta}_1$ or

$$S_{D1} = \frac{\sum_{j=1}^n |x_{jk} - x_{jl}|}{n};$$

2. the *mean Euclidean distance* or *average distance*, given by $\bar{\Delta}_{\frac{1}{2}}$ or

$$S_{D2} = \frac{\sqrt{\sum_{j=1}^n (x_{jk} - x_{jl})^2}}{n};$$

3. the *mean squared Euclidean distance* (Sokal, 1961), given by $\bar{\Delta}_{\frac{1}{2}}^2$ or

$$S_{D3} = \frac{\sum_{j=1}^n (x_{jk} - x_{jl})^2}{n};$$

4. the *mean Canberra metric* (Lance and Williams, 1966), given by

$$S_{D4} = \frac{\sum_{j=1}^n \left(\frac{|x_{jk} - x_{jl}|}{x_{jk} + x_{jl}} \right)}{n};$$

5. and the *coefficient of divergence* (Clark, 1952), given by

$$S_{D5} = \sqrt[n]{\frac{\sum_{j=1}^n \left(\frac{|x_{jk} - x_{jl}|}{x_{jk} + x_{jl}} \right)^2}{n}}.$$

The first three of these normalize the variation in the Manhattan, Euclidean and squared Euclidean distances, respectively, that occurs due to the variable number of attributes. The final two additionally normalize the

variation in the Manhattan metric and in the Euclidean distance, respectively, that occurs due to the variable range of attribute-values: these *scale-invariant* metrics reflect proportional rather than absolute differences between values, but the values must all be positive for their use to be valid (Sneath and Sokal, 1973). A family of related scale-invariant coefficients known as *distortion* measures is defined in the section on the meaning of correlation coefficients later in this article. The Euclidean distance and its variants obey Pythagoras' theorem as well as the four axioms listed above, and as such are to be preferred as measures of geometric distance over the Manhattan metric and its variants. The Canberra metric was originally known as the 'non-metric' coefficient, but its metric properties have since been confirmed (Lance and Williams, 1967). A related coefficient, which is non-metric, is the *Bray/Curtis* coefficient (Motyka, Dobrzanski and Zawazki, 1950), given by

$$S_{D6} = \frac{\sum_{j=1}^n |x_{jk} - x_{jl}|}{\sum_{j=1}^n (x_{jk} + x_{jl})}.$$

In the forms in which they are expressed above, these distance coefficients may be used in cases where the x values are *non-binary* frequencies, weights, probabilities or distances. It can be shown that both the mean squared Euclidean distance and the mean Canberra metric will always have the same value as the mean Manhattan metric when the x values are in binary form, denoting the presence or absence of attributes in an object rather than their weight.

For binary data, alternative contingency-table and set-theoretic forms of these coefficients may be used. The mean Manhattan metric, mean squared Euclidean distance and mean Canberra metric, for example, may all be expressed by

$$S_{D1}/S_{D3}/S_{D4} = \frac{b + c}{n}$$

or

$$S_{D1}/S_{D3}/S_{D4} = \frac{|X_k| + |X_l| - 2|X_k \cap X_l|}{n}$$

The formulae for binary data of other coefficients are presented in Table 2.

It should be noted that the literature in this field is characterized by a marked inconsistency of terminology. Functions like distance coefficients that exhibit an inverse relationship with degree of similarity are often referred to as measures of *dissimilarity* rather than similarity. Some authors reserve the term 'similarity coefficients' to refer

Table 2 Formulae of similarity coefficients

ID	Common name	Non-binary		Binary	
		Formula	Range	Formula	Range
D1	Mean Manhattan	$\frac{\sum x_k - x_j }{n}$	∞ to 0	$\frac{b+c}{n}$	1 to 0
D2	Mean Euclidean	$\frac{\sqrt{\sum x_k - x_j ^2}}{n}$	∞ to 0	$\frac{\sqrt{b+c}}{n}$	1 to 0
D3	Mean squared Euclidean	$\frac{\sum x_k - x_j ^2}{n}$	∞ to 0	$\frac{b+c}{n}$	1 to 0
D4	Mean Canberra	$\frac{\sum \left(\frac{ x_k - x_j }{x_k + x_j} \right)}{n}$	1 to 0	$\frac{b+c}{n}$	1 to 0
D5	Divergence	$\sqrt{\frac{\sum \left(\frac{ x_k - x_j }{x_k + x_j} \right)^2}{n}}$	1 to 0	$\frac{\sqrt{b+c}}{\sqrt{n}}$	1 to 0
D6	Bray/Curtis	$\frac{\sum x_k - x_j }{\sum (x_k + x_j)}$	1 to 0	$\frac{b+c}{2a+b+c}$	1 to 0
A1	Jaccard	$\frac{\sum (x_k \cdot x_j)}{\sum (x_k)^2 - \sum (x_j)^2 - \sum (x_k \cdot x_j)}$	0 to 1	$\frac{a}{a+b+c}$	0 to 1
A2	Dice	$\frac{2\sum (x_k \cdot x_j)}{\sum (x_k)^2 + \sum (x_j)^2}$	0 to 1	$\frac{2a}{2a+b+c}$	0 to 1
A3	Russell/Rao	$\frac{\sum (x_k \cdot x_j)}{n}$	0 to ∞	$\frac{a}{n}$	0 to 1
A4	Sokal/Sneath (1)	$\frac{\sum (x_k \cdot x_j)}{2\sum (x_k)^2 - 2\sum (x_j)^2 - 2\sum (x_k \cdot x_j)}$	0 to 1	$\frac{a}{a+2b+2c}$	0 to 1
A5	Kulczyński (1)	$\frac{\sum (x_k \cdot x_j)}{\sum (x_k)^2 + \sum (x_j)^2 - 2\sum (x_k \cdot x_j)}$	0 to ∞	$\frac{a}{b+c}$	0 to ∞
A6	Simple matching	-	-	$\frac{a+d}{n}$	0 to 1
A7	Hamann	-	-	$\frac{a+d-b-c}{n}$	-1 to +1
A8	Sokal/Sneath (2)	-	-	$\frac{2a+2d}{a+d+n}$	0 to 1
A9	Rogers/Tanimoto	-	-	$\frac{a+d}{b+c+n}$	0 to 1
A10	Sokal/Sneath (3)	-	-	$\frac{a+d}{b+c}$	0 to ∞
A11	Baroni-Urbani/Buser	-	-	$\frac{\sqrt{ad+a}}{\sqrt{ad+a+b+c}}$	0 to 1
A12	Ochiai	$\frac{\sum (x_k \cdot x_j)}{\sqrt{\sum (x_k)^2} \cdot \sqrt{\sum (x_j)^2}}$	0 to 1	$\frac{a}{\sqrt{(a+b)(a+c)}}$	0 to 1
A13	Kulczyński (2)	$\frac{\sum (x_k \cdot x_j) / (\sum (x_k \cdot j) + \sum (x_j \cdot i))}{\sum (x_k \cdot j) \cdot \sum (x_j \cdot i)}$	0 to 1	$\frac{a(2a-b-c)}{(a+b)(a+c)}$	0 to 1

Table 2 Continued

ID	Common name	Non-binary		Binary	
		Formula	Range	Formula	Range
A14	Forbes	$\frac{n \sum (x_k \cdot x_j)}{\sum (x_k)^2 \cdot \sum (x_j)^2}$	0 to ∞	$\frac{n \cdot a}{(a+b)(a+c)}$	0 to ∞
A15	Fossum	$\frac{n \left(\sum (x_k \cdot x_j) - \frac{1}{2} \right)^2}{\sum (x_k)^2 \cdot \sum (x_j)^2}$	0 to ∞	$\frac{n \left(a - \frac{1}{2} \right)^2}{(a+b)(a+c)}$	0 to ∞
A16	Simpson	$\frac{\sum \min(x_k, x_j)}{\min(\sum x_k, \sum x_j)}$	0 to 1	$\frac{a}{\min(a-b, a-c)}$	0 to 1
C1	Pearson	$\frac{\sum (x_k - \bar{x}_k)(x_j - \bar{x}_j)}{\sqrt{\sum (x_k - \bar{x}_k)^2 \cdot \sum (x_j - \bar{x}_j)^2}}$	- 1 to + 1	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	- 1 to + 1
C2	Yule	-	-	$\frac{ad-bc}{ad+bc}$	- 1 to + 1
C3	McConnaughey	-	-	$\frac{a^2-bc}{(a+b)(a+c)}$	- 1 to + 1
C4	Stiles	-	-	$\log_e \frac{n \left(ad-bc - \frac{n}{2} \right)^2}{(a+b)(a+c)(b+d)(c+d)}$	$-\infty$ to $+\infty$
C5	Dennis	-	-	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	0 to ∞

only to what we call association coefficients. To confuse matters further, some of these (see, for example, Clifford and Stephenson (1975)) re-allocate the term 'association coefficients' to what we call correlation coefficients. Others use 'coefficients of association' for correlation coefficients, and 'coefficients of co-occurrence' for association coefficients (see, for example, Jackson, Somers and Harvey (1989)). Several authors make a distinction of a different kind, between the measurement of the *similarity* (or resemblance) of *objects*, and the measurement of the *association* (or affinity) of *attributes*. Calculations of these types are said to use the *Q* technique and the *R* technique, respectively (Sokal and Sneath, 1963). Although coefficients of all types have at some time been used for both *Q*-type and *R*-type studies, association coefficients are generally preferred for the former and correlation coefficients for the latter. At a more mundane level, van Rijsbergen (1979) calls the inner product the *simple matching coefficient*, which is actually the common name for another coefficient described below. The formula for the mean Euclidean distance given above is the one

quoted by Willett (1987): the normalizing factor in the formula given by Jardine and Sibson (1971) and Sneath and Sokal (1973) is $\frac{1}{\sqrt{n}}$. Our formula for the *mean Canberra metric* is identified by Clifford and Stephenson (1975) as that for the Canberra metric itself. Both Cormack (1971) and Willett (1987) express the Canberra metric using the formula that is normally used to express the Bray/Curtis coefficient.

Association coefficients

As we have noted, association coefficients are based upon the inner product. Most, but not all, vary in value from 0 (indicating no similarity) to 1 (indicating complete similarity). Two simple examples are the coefficient of *Jaccard* (1901) and the coefficient of *Dice* (1945). In the field of information retrieval, the former is sometimes known as the coefficient of *Tanimoto* (1958), and we shall continue this usage. Other early uses of the Tanimoto coefficient in this field are described by Parker-Rhodes and Needham (1960), Doyle (1962) and Hooper (1965). Similarly, the Dice coefficient is sometimes known either as the

coefficient of Czekanowski (1913) or of Sørensen (1948).

The Tanimoto coefficient is given by

$$S_{A1} = \frac{\sum (x_{jk} \cdot x_{jl})}{\sum (x_{jk})^2 + \sum (x_{jl})^2 - \sum (x_{jk} \cdot x_{jl})}$$

and the Dice coefficient by

$$S_{A2} = \frac{2 \sum (x_{jk} \cdot x_{jl})}{\sum (x_{jk})^2 + \sum (x_{jl})^2}$$

where the summation in all cases is over $j = 1$ to $j = n$.

We may define contingency-table and set-theoretic forms of each of these coefficients for use with binary data. The equivalent for binary data of the inner product is a , or $|X_k \cap X_l|$, and the basic formula of an association coefficient for binary data is formed by dividing a or $|X_k \cap X_l|$, the *actual* number of attributes whose value in each of the two vectors is 1 (i.e., *positive matches*), by a factor representing the number of attributes whose value in each of the two vectors could *possibly* be 1. Different coefficients vary in the exact composition of this factor. The Tanimoto coefficient may be expressed by

$$S_{A1} = \frac{a}{a + b + c}$$

$$S_{A1} = \frac{|X_k \cap X_l|}{|X_k| + |X_l| - |X_k \cap X_l|}$$

or

$$S_{A1} = \frac{|X_k \cap X_l|}{|X_k \cup X_l|}$$

and the Dice coefficient by

$$S_{A2} = \frac{2a}{2a + b + c}$$

or

$$S_{A2} = \frac{2 |X_k \cap X_l|}{|X_k| + |X_l|}$$

We can now see more clearly that, in the Tanimoto coefficient, attributes whose value is 1 in each of the two vectors and attributes whose value is 1 only in one vector or the other are equally weighted; whereas, in the Dice coefficient, positive matches carry twice the weight of mismatches. Many other association coefficients have been suggested that use different weighting schemes in the composition of the normalising factor.

Related to the Tanimoto and Dice coefficients are the coefficients of Russell and Rao (1940), Sokal and Sneath (1963) and Kulczyński (1927). The formulae for each of

these, and of other coefficients described below, are presented in Table 2. We give these particular coefficients of Sokal/Sneath and Kulczyński the suffix '(1)' to distinguish them from other coefficients described by the same authors (and discussed below). The coefficient of Russell/Rao is simply the inner product normalized by the number of attributes n . The coefficient of Kulczyński (1) is not normalized satisfactorily, and its values vary from 0 to an indefinitely large number according to the value of n .

A separate class of coefficients is composed of those that consider, in addition to the number of positive matches (attributes whose value is 1 in each of the two vectors) and the number of mismatches (attributes whose value is 1 in only one of the two vectors), the number of *negative matches* (attributes whose value is 0 in each of the two vectors). Such coefficients are suitable only for use with binary data. The most well-known member of this class is the *simple matching coefficient* (Sokal and Michener, 1958; Zubin, 1938), which is formed by adding the number of negative matches to both the numerator and the denominator of the Tanimoto coefficient:

$$S_{A6} = \frac{a + d}{a + b + c + d}$$

or

$$S_{A6} = \frac{|X_k \cap X_l| + |X_k \cup X_l|}{|X_k| + |X_l| + |X_k \cap X_l|}$$

The denominator of the simple matching coefficient is equivalent to n , the total number of elements in each vector. Other members of this class are: the coefficient of Hamann (1961), which subtracts the number of mismatches from the numerator of the simple matching coefficient and whose values vary from -1 to $+1$; the coefficient of Sokal and Sneath (2) (1963), which adds twice the number of negative matches to the numerator and denominator of the Dice coefficient; the coefficient of Rogers and Tanimoto (1960), which adds the number of negative matches to the numerator and denominator of the coefficient of Sokal/Sneath (1); the coefficient of Sokal and Sneath (3) (1963), which adds the number of negative matches to the numerator only of the coefficient of Kulczyński (1), and whose upper value similarly varies according to the value of n ; and the coefficient of Baroni-Urbani and Buser (1976), which adds the square root of the product of the number of positive matches and the number of negative matches to the numerator and denominator of the Tanimoto coefficient.

Another class is that of the *angular coefficients*. Just as distance coefficients may be viewed as functions of the geometric *distances* between vectors plotted in n -dimensional space, *angular coefficients* may be considered as functions of the *angles* between these vectors. The best-known example is the coefficient of Ochiai (1957; Driver

and Kroeber, 1932), often known in information retrieval circles as the *cosine* coefficient (Salton, 1963), given by

$$S_{A11} = \frac{\sum (x_{jk} \cdot x_{jl})}{\sqrt{\sum (x_{jk})^2 \cdot \sum (x_{jl})^2}}$$

where the summation in each case is again over $j = 1$ to $j = n$. The angle between vectors which are identical is 0° , and the cosine of this angle is 1; vectors that are completely dissimilar will be at right angles to each other, and the cosine of 90° is 0; hence the cosine coefficient, like most other association coefficients, varies from 0 to 1. The form of the cosine coefficient for use with binary data is

$$S_{A11} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

or

$$S_{A11} = \frac{|X_k \cap X_l|}{\sqrt{|X_k| \cdot |X_l|}}$$

Other, more exotic, coefficients may be identified whose denominator is similar in form to that of the cosine coefficient: the coefficient of Kulczyński (2) (1927); the coefficient of Forbes (1907) (sometimes known in information retrieval as that of Kochen and Wong (1962)); and the coefficient of Fossum (1966). The upper values of the latter two again vary according to the value of n .

The cosine coefficient is the most popular of similarity measures amongst proponents of the vector processing model. The validity of its basis in Euclidean geometry has, however, been questioned in a similar way to that in which the appropriateness to applications in text retrieval of coefficients based on geometric distance has been questioned. As Wong and Raghavan point out (1984; Raghavan and Wong, 1986) (see the earlier section on distance coefficients), early expositions of the vector processing model did not refer to the formal notions of vector space and independence, treating 'vectors' simply as one-dimensional arrays (i.e., *tuples* of a length equal to the number of attribute-values of an object). Salton, Wong and Yang (1975) were the first to make explicit reference to the formal correspondence between index terms and the dimensions of a multi-dimensional space, and their interpretation has been accepted in much subsequent work. Wong and Raghavan conclude, however, that this work has not depended on the representation of documents in vector space, and that the use in text retrieval of formal concepts such as that of multi-dimensional space should be regarded as 'casual flirtings'.

Finally, we may consider the coefficient of Simpson (1943), known in information retrieval as the *overlap* or *asymmetric* coefficient (Lesk, 1964), given by

$$S_{A15} = \frac{\sum \min(x_{jk}, x_{jl})}{\min(\sum x_{jk}, \sum x_{jl})}$$

where the summations are over $j = 1$ to $j = n$. Values vary from 0 to 1. The form of the overlap coefficient for use with binary data is

$$S_{A15} = \frac{a}{\min(a+b, a+c)}$$

or

$$S_{A15} = \frac{|X_k \cap X_l|}{\min(|X_k|, |X_l|)}$$

The values of the *complement* of a function may be calculated by subtracting the values of that function from its maximum value. The complements of many association coefficients (i.e., those whose values vary from 0 to 1) may be considered as distance coefficients. For example, a value for a distance coefficient based on the Tanimoto coefficient may be calculated by subtracting the value of S_{A1} from 1. The complement of the simple matching coefficient may be equated with the mean squared Euclidean distance, and is therefore metric; when the values being compared are in binary form, it may also be equated with the mean Manhattan metric. The complements of the Tanimoto and the Dice coefficients, however, are non-metric (Sneath and Sokal, 1973).⁴ The complement of the binary form of the Dice coefficient may be equated with the Bray/Curtis coefficient in its binary form.

Correlation coefficients

A special class of angular coefficients is formed by the *correlation coefficients*, which require the calculation of *moments* - i.e., the differences between each attribute-value and the mean of all attribute-values in a vector. The most common of these is the *Pearson product-moment correlation coefficient* (often denoted by r), given by

$$S_{C1} = \frac{\sum (x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)}{\sqrt{\sum (x_{jk} - \bar{x}_k)^2 \sum (x_{jl} - \bar{x}_l)^2}}$$

or by

$$S_{C1} = \frac{\sum (x_{jk} \cdot x_{jl}) - (n \cdot \bar{x}_k \cdot \bar{x}_l)}{\sqrt{\sum (x_{jk}^2 - n \cdot \bar{x}_k^2) \sum (x_{jl}^2 - n \cdot \bar{x}_l^2)}}$$

where the summation in each case is over $j = 1$ to $j = n$. Values of the Pearson coefficient vary within the range -1 to $+1$. The contingency-table form of the Pearson coefficient for use with binary data (Yule, 1912), often known as the phi coefficient and denoted by ϕ , is

$$S_{C1} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

The phi coefficient is related closely to the chi-squared statistic mentioned in Note 2: $\phi = \sqrt{\frac{\chi^2}{n}}$. The 'association factor' of Stiles (1961) is similarly related, being equivalent to the logarithmic value of the modification of χ^2 suggested by Yates (1934) for applications where the values of a, b, c, d in a contingency table are each less than 10. Other related coefficients, like Stiles' suitable only for use with binary data, are: the coefficient of Yule (1990), also known in information retrieval as the coefficient of Maron and Kuhns (1960); the coefficient of McConnaughey (1964); and the coefficient of Dennis (1965). The formulae of these coefficients are presented in Table 2.

The typical form taken by the numerator of correlation coefficients for use with binary data (i.e., $ad - bc$, or the *determinant* of the 2×2 contingency table) may be justified as a comparison between the *observed* number of co-occurrences (given by a) and the *expected* number of co-occurrences on the assumption of the independent, random distribution of attributes (given by $\frac{(a+b)(a+c)}{n}$). To calculate the excess of observed co-occurrences over and above the expected number of co-occurrences, we use the formula $a - \frac{(a+b)(a+c)}{n}$, which is equal to $\frac{ad - bc}{n}$. The precise forms of the various correlation coefficients are the result of different normalizations of this formula. Any correlation coefficient whose values vary within the range -1 to $+1$ may be converted into one whose values vary within the range 0 to 1 using the formula $\frac{S_r + 1}{2}$.

THE CHARACTERISTICS OF SIMILARITY COEFFICIENTS

More than a quarter of a century ago, Jones and Curtice (1967) prefaced their discussion of existing formulae for the measurement of the degree of association between index terms with the following grumble: 'What is annoying is that no clear-cut criterion for choice among the alternatives has emerged. As a result, few candidate measures have been permanently dismissed from consideration, and a rather large set of formulae remains available.' A decade later, the same problem remained: Sager and Lockemann (1976) began their evaluation of 14 ranking algorithms with the observation that '(although) it is known that not all techniques yield equal results in a particular situation, a systematic collection, classification and comparison seems to be missing.' In a much larger study of the effectiveness of 504 different ranking algorithms, McGill, Koll and Noreault (1979) stated that 'conclusive tests of similarity measures for performance differences have not been conducted.' By 1992, Wang, Wong and Yao (1992) were able to draw on the algebraic and empirical analyses undertaken by the writers quoted above, and the geometric analysis of seven similarity

coefficients offered by Jones and Furnas (1987), but were still led to conclude that the precise conditions under which a particular coefficient should be used had not been identified in any previous work.

Presented in these terms, the history of research into the use of similarity coefficients in text retrieval appears to betray a lack of progress. However, there are a number of points of significance that can be made regarding the selection of appropriate measures for applications in text retrieval.

Equivalence and monotonicity

There are several coefficients that are referred to in the literature by various names, or that are variously represented by different forms of what may be revealed on close inspection to be the same formulae. Examples of these are provided in the discussion in the section on the composition of similarity coefficients. For binary data, for example, the mean squared Euclidean distance and the Canberra metric are equivalent to the Manhattan metric, and their complements are each equivalent to the simple matching coefficient. In turn, if its values are converted so that they fall in the range 0 to 1 using the formula $\frac{S_{sr} + 1}{2}$, the Hamann coefficient is equivalent to all of these. Similarly, for binary data, the complement of the Bray/Curtis coefficient is equivalent to the Dice coefficient; and, when its values are converted to fall in the range 0 to 1 , the McConnaughey coefficient is equivalent to the coefficient of Kulczyński (2).

We may also identify, as well as those coefficients that are identical to each other, many coefficients that are *monotonic* with each other. Two coefficients may be said to be jointly monotonic if it can be shown that the ranking of all measurements of similarity between pairs of objects in a specific set is the same using one coefficient as it is using the other. Considering the case where the objects in such a set were the documents in a database, for example, the output of ranking algorithms that differ only in the similarity coefficients used would be the same if those coefficients were monotonic with each other. This monotonicity may be demonstrated by a simple algebraic comparison of the two formulae, or by means of empirical simulation. The latter method involves the random generation of the data elements of a set of vectors, and the calculation of a value for each coefficient for each pair of vectors. The values for each coefficient may then themselves be considered as the elements of a vector representing that coefficient, which may be compared with the elements of another such vector using the formula for the Pearson correlation coefficient. The existence of monotonicity may be inferred from the level of correlation: the closer the value of r to 1.00 , the greater the evidence of joint monotonicity.

Using these methods, McGill, Koll and Noreault (1979) managed to whittle an initial list of 67 similarity

coefficients down to 24 that each produce a unique ranking. For instance, they suggest that, for binary data, the coefficients of Dice (and its equivalent, the Bray/Curtis coefficient), Sokal/Sneath (1) and Kulczyński (1) are all monotonic with the Tanimoto coefficient (and hence with each other). McGill *et al.* also note the joint monotonicity, for binary data, of the complements of the mean Manhattan metric (and its equivalents), the mean Euclidean distance, and the divergence coefficient, together with the simple matching coefficient (and its equivalents) and the Rogers/Tanimoto coefficient. To this list, we may add the coefficient of Sokal/Sneath (2).

McGill *et al.* also assert that the coefficients of Forbes and Dennis are both monotonic with the cosine coefficient; and that the coefficient of Stiles is monotonic with the Pearson coefficient. Some other pairs of coefficients displayed levels of correlation above $r = 0.7$: for example, the cosine and Pearson coefficients, and the coefficients of Dice and Kulczyński (2). In contrast, some coefficients exhibited conspicuously low correlations with most others: for example, the coefficient of Fossum. Some of these findings do not, however, match the results of the empirical trial described by Hubálek (1982). From an exhaustive list of 43 association and correlation coefficients, Hubálek selected 20 that passed each of a number of 'admissibility conditions', including the specification that coefficients should take their maximum value at $b = c = 0$ and their minimum at $a = d = 0$, and that they should be able to discriminate between positive and negative associations. Again using the Pearson correlation coefficient, Hubálek compared each pair of the sets of values produced by these 20 different coefficients, and grouped the measures into clusters using two different clustering methods. At the level of $r = 0.96$, he confirmed the monotonicity of the Dice, Tanimoto, Sokal/Sneath (1) and Kulczyński (1) measures; but he also found all these to be monotonic at the same level with the cosine, Kulczyński (2) and Pearson coefficients. Initial results from our own simulations (Ellis, Furner-Hines and Willett, 1993b) appear to suggest that the Fossum and overlap coefficients, and, once its values have been converted so that they fall in the range 0 to 1, the coefficient of Dennis, may all be added to the list of those that are monotonic with the Dice coefficient.

The consideration of negative matches

The differences between those coefficients that are *not* monotonic with each other are such that significantly different rankings or clusterings may result from the use of different coefficients to measure the similarities amongst the same set of objects. Each coefficient gives different weightings to different properties of the data, and the behaviour of an individual coefficient depends largely on the specific nature of the data under observation. The typical nature of data studied in the field of text retrieval militates against the use of certain types of coefficient, and it is instructive that very few coefficients have actually been used

in operational mechanisms for ranking documents in response to queries. The cosine coefficient, with its clear theoretical affinity to the vector space model, has become the preferred measure of many (notably those who build on the work of Salton or van Rijsbergen); the Dice and Tanimoto coefficients also have their advocates. The simple matching coefficient, on the other hand, has seldom, if ever, been used in the text retrieval environment – even though, as its name suggests, it is one of the simplest (and oldest) of all similarity coefficients. The reason is that the objects under consideration in the context of information retrieval are typically characterized by sets of attributes that form small proportions of the total number of attributes that might possibly be assigned. The number of negative matches in these cases is therefore likely to be much greater than the number of positive matches; and the use of any similarity coefficient that takes negative matches into account, such as the simple matching coefficient, will result in measurements of similarity that are not especially discriminatory. Even if $a = 0$, the simple matching coefficient takes high values as $d \rightarrow n$. As Clifford and Stephenson (1976) point out, 'in some circumstances it would seem ridiculous to regard two (objects) as similar largely on the basis of them both lacking something.' On the other hand, 'in other circumstances it would seem improper to neglect conjoint absences.' For applications in information retrieval, any of the measures that are equivalent to, or monotonic with, the Dice coefficient would be a satisfactory choice; whereas the use of any of those that are equivalent to, or monotonic with, the simple matching coefficient would be regarded as perverse.

Geometric analyses

Galvanised, perhaps, by the challenge laid down by Wong and Raghavan (1984; Raghavan and Wong, 1986) (see the section on methods of expressing the formulae of similarity coefficients), several studies have attempted to explain the differences in the output of different similarity coefficients by analysing more precisely their geometric 'meaning'. Whereas McGill and his co-workers put most emphasis on discovering relationships amongst formulae that are used with binary data, these later writers have used geometrical techniques to investigate the behaviour of coefficients when used with non-binary data.

Wong *et al.* (Bollmann and Wong, 1987; Wang, Wong and Yao, 1992; Wong and Yao, 1990; Wong, Yao and Bollmann, 1988; Wong *et al.*, 1991) develop a generalized definition of the *linearity* of a similarity coefficient. A measure may be said to be linear if there exists a normalizing function by which each vector may be multiplied such that the value of the inner product of the normalized vectors is equal to the value of the measure under consideration. By this definition, the cosine, Dice and Pearson coefficients may all be said to be linear. The cosine coefficient, for instance, is formed by multiplying

the inner product of a pair of vectors by the inverse of the product of their Euclidean lengths. It is thus equal to the inner product of the corresponding pair of vectors that may be derived by normalizing the original pair so that their Euclidean lengths each equal 1. In this context, we may note the similarity in form of the Pearson coefficient to the cosine coefficient. The numerator of the Pearson coefficient takes the form of the inner product of a pair of vectors, *viz.* those whose elements are given by subtracting from each of the original vectors' elements the mean of those elements. Without further normalization, this function is known as the *covariance* of a pair of vectors. Just as the cosine coefficient may be formed by dividing the inner product of a pair of vectors by the product of their Euclidean lengths, the Pearson coefficient is formed by dividing the inner product of another pair of vectors, equal to the original vectors normalized by moment, by the product of *those* vectors' Euclidean lengths. Using techniques based on the identification of the contour sets introduced by Jones and Furnas (1987) (see below), Wang, Wong and Yao (1992) identify the necessary and sufficient conditions under which any linear similarity coefficient might produce an acceptable representation of the structure formed by a set of documents and the relation of 'user preference' that we discussed in the section on document-query similarity. Wang *et al.* assert that their work establishes the formal basis for adopting linear similarity measures in information retrieval, and their geometrical justification for the use of such measures goes some way towards rehabilitating certain concepts inherent in the vector space model that have previously been questioned.

Jones and Furnas (1987) assert that the behaviour of a coefficient depends on its *semantic* sensitivity to the information of various kinds that is contained in a set of document vectors. The two most important of these kinds of information are (i) the *topic* or subject area of the content of each document, indicated by the set of relationships between each individual term weight and others in the same document vector, and (ii) the *intensity*⁵ of the content of each document, indicated by the set of relationships between each individual term weight and other weights for the same term in different document vectors. The *direction* of a vector in vector space may be said to represent the topic of a document or query, while its Euclidean *length* may be said to represent the object's intensity. Jones and Furnas attempt to explain the differences amongst coefficients in their sensitivity to topic and intensity using a technique based on the construction of *iso-similarity contours* – i.e., lines in vector space that join the end-points of all those vectors whose degree of similarity to a particular query vector is equal. This technique allows them to identify five important properties of similarity coefficients:

1. If, for instance, given two document vectors of equal

length, the one with the smaller angle from a particular query vector is rated by a coefficient as the more similar, then we may say that the coefficient has the property of *angular monotonicity*. All measures that are linear (by Wong *et al.*'s definition) have this property, albeit to different degrees. In the case of the cosine coefficient, for example, angle monotonicity does not even depend on the two document vectors being of equal length, because vectors with the same direction but different lengths are transformed by the cosine formula to the same vector of unit length.

2. If, given two document vectors of the same angle from the query vector, the longer one is rated by a coefficient as the more similar, then we may say that the coefficient has the property of *radial monotonicity*. The inner product is radial-monotone, but the cosine and Pearson coefficients are not. The latter two measures are thus completely dominated by a sensitivity to topic, and intensity is ignored.
3. If an increase in the value of any element in a document vector to a value greater than that of the corresponding element in a query vector has the effect only of increasing the degree of that vector's similarity with the query vector, or has no effect at all, then we may say that the coefficient has the property of *component-wise monotonicity*. If a measure is not component-monotone, just as the cosine and Pearson coefficients are not, then it might be argued that it can penalize documents for their 'richness', or the high weighting in their vectors of terms that are zero-weighted in the query vector.
4. If it is possible that an increase in the value of any *single* element in a document vector can have the effect of increasing the degree of that vector's similarity with a query vector to an arbitrarily high level, even if the document has very little to do with the query's topic (and hence its vector's angle of separation from the query vector is very large), then we may say that the coefficient is subject to *unbounded single-component influence*. The inner product is subject to this influence, whereas the cosine and Pearson coefficients are not.
5. Finally, if there is no upper limit on the range of possible values that may be produced by a similarity coefficient, we may say that the measure is *unbounded*. As we have already noted, the inner product is indeed unbounded, in contrast to its normalized counterparts. Measures that are not unbounded allow the identification of documents that are maximally similar, or 'ideal', with respect to a given query.

The meaning of correlation coefficients

Correlation coefficients were originally developed in order to measure the degree of correlation between the *attributes* of a sample of independent objects taken from a population (i.e., an R-type study – see the section on distance coefficients). Their values could then be subjected to tests

that would determine the statistical significance of such a correlation. Discussion of the use of correlation coefficients in text retrieval often points to this amenability to statistical analysis as a benefit (Maron and Kuhns, 1960; Salton, 1968). However, the validity of correlation coefficients for the measurement of similarity between vectors representing *objects* (Q-type studies) can be questioned.

- When such coefficients are used in R-type studies, it is assumed that the objects are distributed randomly and independently of each other; it is usually not the case in Q-type studies that attributes are similarly distributed randomly and independently.
- The significance of the mean of a set of values for different attributes is unclear, especially if the attributes of an object are each measured on an individual rather than a standardized scale. Jardine and Sibson (1971) less hesitantly describe the calculation of this mean as 'absurd'.
- Once the values of a correlation coefficient are converted into values of a distance coefficient using the formula $\frac{1-S_c}{2}$, it can be shown that complementary coefficients of this kind are generally non-metric. It is possible for them to disobey not only the fourth metric axiom, because the attribute-values of two non-identical objects might exhibit perfect correlation (as would be the case if one of the vectors could be formed simply by multiplying the other by a scaling factor), but also the axiom of triangle inequality.

Hubálek (1982) notes the relationship of many correlation coefficients to the chi-squared statistic, and draws a more general conclusion: 'Statistical tests of significance of an association must not be confused with the measures of the association, and both should be regarded as separate characteristics of the association. A significance test thus indicates nothing but whether two (objects or attributes) are associated or not at a pre-selected level, and provides no measurement of the degree of the association. Therefore any association analysis that makes use of, for example, chi-square values as a measure of the association could be the subject of controversy.' For these reasons, the use of association coefficients in Q-type studies is usually preferred.

Nevertheless, it should be noted that the product-moment correlation coefficient has been used with rather less uncertainty as to its validity in several contexts related to those with which we have been concerned:

- in comparisons of the results produced by different methods of clustering the same data (Sokal and Rohlf, 1962);
- in comparisons of the set of data contained in a similarity matrix and the results of any method of clustering those data - i.e., in the measurement of the

distortion imposed by the clustering method (Griffiths, Robinson and Willett, 1984; Rohlf, 1974);

- in comparisons of the sets of data contained in two similarity matrices produced by different similarity coefficients (as undertaken in the studies of Hubálek (1982) and Ellis, Furner-Hines and Willett (1993b)).

In each of these contexts, it is the ratios between the values in each set that are the object of attention rather than the absolute values themselves, and most authors have found the product-moment correlation coefficient (or the *cophenetic coefficient*, as it is called by Sokal and Rohlf (1962)) to be appropriate. Jardine and Sibson (1971) point out, however, that even in these contexts 'justification for its use rests on its practical utility rather than on any resemblance to statistical correlation measures'. In an earlier article (Jardine and Sibson, 1968), the same authors introduce a family of alternative measures of distortion, Δ_{μ}^* , each of which normalizes the variation in the corresponding Minkowski metric that occurs due to the variable range of the attribute-values making up the data. In other words, like the mean Canberra metric and the coefficient of divergence (and indeed the product-moment correlation coefficient), these measures are scale-invariant. Δ_{μ}^* is given by

$$\Delta_{\mu}^* = \left(\sum \left| \frac{x_{jk}}{(\sum x_{jk})^{\frac{1}{\mu}}} - \frac{x_{jl}}{(\sum x_{jl})^{\frac{1}{\mu}}} \right|^{\mu} \right)^{\frac{1}{\mu}}$$

CONCLUSIONS

Even in the field of numerical taxonomy, where the use of similarity coefficients has been even more widespread than in information retrieval, Jackson, Somers and Harvey (1989) were moved to conclude that 'the choice of a similarity coefficient is largely subjective and often based on tradition or on *a posteriori* criteria such as the "interpretability" of the results', and went on to quote Gordon (1987): 'Human ingenuity is quite capable of providing a *post hoc* justification of dubious classifications.' Such conclusions are essentially no different from that reached by Yule (1912) (quoted in Hubálek (1982)), who compared the range of available similarity coefficients with the choice of arithmetic mean, geometric mean or median: 'All forms of average are measures of analogous properties, but do not give the same values; the various coefficients that have been suggested for measuring association and correlation differ in precisely the same way.'

To identify the 'best' of his list of 20 'admissible' coefficients, Hubálek (1982) suggests a number of optional conditions. He prefers those measures that vary within the

ranges 0 to 1 or -1 to +1 rather than 0 to ∞ , as those of the latter type are generally over-sensitive to small changes in a . He prefers measures that are linear, such as the Dice and cosine coefficients, to those that are not, such as the Tanimoto coefficient. He also prefers measures that are expressed by simple formulae, and whose computation is therefore most efficient. He suggests that studies would do well to select one coefficient from each of the sub-clusters identified in his empirical trial: in other words, the Dice (or Tanimoto), Kulczyński (2), cosine, Pearson and Baroni-Urbani/Buser coefficients. Our general conclusion is the same as Hubálek's, who quotes Goodman and Kruskal (1954): 'Each scientific area that has use for different measures of association should, after appropriate argument and trial, settle down on those measures most useful for its needs.' For most applications in information retrieval, the historical attachment to the simple, linear, association coefficients provided by the Dice and cosine formulae is in no need of revision.

ACKNOWLEDGEMENTS

We thank the British Library Research and Development Department for funding this work under grant number RDD/G/142, and Dr Geoff Downs for a careful reading of this manuscript.

NOTES

1. Cleverdon *et al.*'s measure (ascribed to Vickery) reduces to the complement of the coefficient of Sokal/Sneath (1), Gebhardt's measure to the cosine coefficient, Heine's measure to the complement of the Tanimoto coefficient, and van Rijsbergen's E measure (when its component scaling factor is made equal to 0.5) to the complement of the Dice coefficient (i.e., the Bray/Curtis coefficient). The forms of these coefficients are defined in the section on the composition of similarity coefficients.
2. Contingency tables are widely used in data analysis, specifically in order to determine whether the degree of association between two attributes is statistically significant. This involves the use of chi-squared (χ^2), or related statistics, to test the (non-binary) data presented in a contingency table against the null hypothesis that the attributes are independent (Kinnucan *et al.*, 1987). Here, however, we shall consider the use of the 2×2 table only as a basis for calculating the degrees of similarity between objects, rather than for determining in addition the statistical significance of these calculations.
3. If we define Y to be the set of all possible values y_r that may be represented by each element x_j , where r varies from 1 to p , we may say that a contingency table consists of a matrix of values, where the sum of the values in each row k , is equal to the number of elements x_{jk} in \bar{X}_k that are equal to y_r , and where the sum of the values in each column l , is equal to the number of elements x_{il} in \bar{X}_k that are equal to y_r . Each element of the matrix is therefore a value representing the number of attributes A_j whose value in \bar{X}_k is equal to y_{rk} and whose value for \bar{X}_k is equal to y_{rl} . The sum of all these values is equal to the number of attributes of each object.
4. Jackson *et al.* (1989) confirm the non-metric nature of the complements of the Dice and cosine coefficients, but assert that the complement of the Tanimoto coefficient is metric.
5. Jones and Furnas suggest that 'intensity' might be interpreted variously as 'quantity', 'quality', or 'accessibility'.

REFERENCES

- Al-Hawamdeh, S. and Willett, P. (1989) Paragraph-based nearest neighbour searching in full-text documents. *Electronic Publishing: Origination, Dissemination and Design*, 2, 179-192
- Al-Hawamdeh, S., Smith, G. and Willett, P. (1991) Paragraph-based access to full-text documents using a hypertext system. *Program*, 25, 119-131
- Anderberg, M. R. (1973) *Cluster analysis for applications*. New York, NY: Academic Press
- Andersen, M. H., Nielsen, J. and Rasmussen, H. (1989) A similarity-based hypertext browser for reading the UNIX network news. *Hypermedia*, 1, 255-265
- Angell, R. C., Freund, G. E. and Willett, P. (1983) Automatic spelling correction using a trigram similarity measure. *Information Processing and Management*, 19, 255-261
- Baroni-Urbani, C. and Buser, M. W. (1976) Similarity of binary data. *Systematic Zoology*, 25, 251-259
- Becker, J. and Hayes, R. M. (1963) *Information storage and retrieval: Tools, elements, theories*. New York, NY: Wiley
- Belkin, N. J. and Croft, W. B. (1987) Retrieval techniques. *Annual Review of Information Science and Technology*, 22, 109-145
- Bernstein, M. (1990) An apprentice that discovers hypertext links. In *Hypertext: Concepts, systems and applications*, eds. A. Rizk, N. Streitz and J. André, pp. 212-223. Cambridge: Cambridge University Press
- Bollmann, P. and Wong, S. K. M. (1987) Adaptive linear information retrieval models. In *Research and development in information retrieval: Proceedings of the Tenth Annual International ACM SIGIR Conference*, eds. C. Yu and C. van Rijsbergen, pp. 157-163. New York, NY: Association for Computing Machinery
- Burgin, R. (1992) Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, 28, 619-627
- Clark, P. J. (1952) An extension of the coefficient of divergence for use with multiple characters. *Copeia*, 2, 61-64
- Cleverdon, C. W., Mills, J. and Keen, M. (1966) *Factors determining the performance of indexing systems. Volume I: Design*. Cranfield: College of Aeronautics

- Clifford, H. T. and Stephenson, W. (1976) *An introduction to numerical classification*. New York, NY: Academic Press
- Cooper, W. S. (1969) Is interindexer consistency a hobgoblin? *American Documentation*, **20**, 268–278
- Cormack, R. M. (1971) A review of classification. *Journal of the Royal Statistical Society, Series A (General)*, **134**, 321–353
- Crouch, D. B., Crouch, C. J. and Andreas, G. (1989) The use of cluster hierarchies in hypertext information retrieval. In *Hypertext '89 proceedings*, pp. 225–237. New York, NY: Association for Computing Machinery
- Cuadra, C. A. and Katter, R. V. (1967) *Experimental studies of relevance judgments*. Santa Monica, CA: Systems Development Corporation
- Czekanowski, J. (1913) Zarys metod statystycznych w zastosowaniu do antropologii. *Travaux de la Société des Sciences de Varsovie. III. Classes des Sciences Mathématiques et Naturelles*, **5**
- Dennis, S. F. (1965) The construction of a thesaurus automatically from a sample of text. In *Statistical association techniques for mechanized documentation: Symposium proceedings*, eds. M. E. Stevens, V. E. Guiliano and L. B. Heilprin, pp. 61–148. Miscellaneous publication 269, National Bureau of Standards. Washington, DC: US Department of Commerce
- Dice, L. R. (1945) Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302
- Doyle, L. B. (1962) Indexing and abstracting by association. *American Documentation*, **13**, 378–390
- Driver, H. E. and Kroeber, A. L. (1932) Quantitative expression of cultural relationships. *The University of California Publications in American Archaeology and Ethnology*, **31**, 211–256
- Eisenberg, M. B. (1988) Measuring relevance judgments. *Information Processing and Management*, **24**, 373–389
- Ellis, D., Furner-Hines, J. and Willett, P. (1993a) Measuring the consistency of assignment of hypertext links in full-text documents. Paper presented at *The 15th British Computer Society Information Retrieval Colloquium* (University of Strathclyde, Glasgow, 29–30 March 1993)
- Ellis, D., Furner-Hines, J. and Willett, P. (1993b) On the consistency of assignment of sets of hypertext links. In preparation
- Fossum, E. G. (1966) *Optimization and standardization of information retrieval language and systems*. Springfield, VA: Clearinghouse for Federal Scientific and Technical Information
- Freund, G. E. and Willett, P. (1982) Online identification of word variants and arbitrary truncation searching using a string similarity measure. *Information Technology: Research and Development*, **1**, 177–187
- Frisse, M. E. (1988) Searching for information in a hypertext medical handbook. *Communications of the ACM*, **31**, 880–886
- Garfield, E. (1979) *Citation indexing – its theory and application in science, technology, and humanities*. New York, NY: Wiley
- Gebhardt, F. (1975) A single probabilistic model for the relevance assessment of documents. *Information Processing and Management*, **11**, 59–65
- Gomez, L. M., Lochbaum, C. C. and Landauer, T. K. (1990) All the right words: Finding what you want as a function of richness of indexing vocabulary. *Journal of the American Society for Information Science*, **41**, 547–559
- Goodman, L. A. and Kruskal, W. H. (1959) Measures of association for cross classifications. *Journal of the American Statistical Association*, **49**, 732–764
- Gordon, A. D. (1981) *Classification*. London: Chapman and Hall
- Gordon, A. D. (1987) A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A (General)*, **150**, 119–137
- Griffiths, A., Robinson, L. A. and Willett, P. (1984) Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation*, **40**, 175–205
- Hamann, U. (1961) Merkmalbestand und Verwandtschaftsbeziehungen den Farinosae: Ein Betrag zum System der Monokotyledonen. *Willdenowia*, **2**, 639–768
- Heincke, F. (1898) Naturgeschichte des Herings. 1. Die Likalformen and die Wanderungen des Herings in den europäischen Meeren. *Deutscher Seefischereiverein, Abhandlung*, **2**, 1–223
- Heine, M. H. (1973) Distance between sets as an objective measure of retrieval effectiveness. *Information Storage and Retrieval*, **9**, 181–198
- Hooper, R. S. (1965) *Indexer consistency tests – origin, measurements, results and utilization*. Bethesda, MD: IBM
- Hubálek, Z. (1982) Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews of the Cambridge Philosophical Society*, **57**, 669–689
- Iivonen, M. (1990) Interindexer consistency and the indexing environment. *International Forum on Information and Documentation*, **15**, (2), 16–21
- Jaccard, P. (1901) Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 241–272
- Jackson, D. A., Somers, K. M. and Harvey, H. H. (1989) Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist*, **133**, 436–453
- Jardine, N. and Sibson, R. (1968) The construction of hierarchic and non-hierarchic classifications. *The Computer Journal*, **11**, 177–184
- Jardine, N. and Sibson, R. (1971) *Mathematical taxonomy*. London: Wiley
- Jardine, N. and van Rijsbergen, C. J. (1971) The use of hierarchic clustering in information retrieval.

- Information Storage and Retrieval*, 7, 217–240
- Johnson, M. A. (1989) A review and examination of the mathematical spaces underlying molecular similarity analysis. *Journal of Mathematical Chemistry*, 3, 117–145
- Jones, P. E. and Curtice, R. M. (1967) A framework for comparing term association measures. *American Documentation*, 18, 153–161
- Jones, W. P. and Furnas, G. W. (1987) Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38, 420–442
- King, D. W. and Bryant, E. C. (1971) *The relevance of information services and products*. Washington, DC: Information Resources Press
- Kinnucan, M. T., Nelson, M. J. and Allen, B. L. (1987) Statistical methods in information science research. *Annual Review of Information Science and Technology*, 22, 147–178
- Kochen, M. and Wong, E. (1962) Concerning the possibility of a co-operative information exchange. *IBM Journal of Research and Development*, 6, 270–271
- Kulczyński, S. (1927) Zespoły roślin w Pieninach. *Bulletin Internationale de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série B (Sciences Naturelles), Supplement II*, 57–203
- Lance, G. N. and Williams, W. T. (1966) Computer programs for hierarchical polythetic classification ('similarity analyses'). *The Computer Journal*, 9, 60–64
- Lance, G. N. and Williams, W. T. (1967) Mixed-data classificatory programs. I. Agglomerative systems. *Australian Computer Journal*, 1, 15–20
- Leonard, L. E. (1975) *Interindexer consistency and retrieval effectiveness: Measurement of relationships*. PhD dissertation, Graduate School of Library Science, Urbana-Champaign, IL: University of Illinois
- Leonard, L. E. (1977) *Interindexer consistency studies, 1954–1975: A review of the literature and summary of study results*. Occasional paper 131, Graduate School of Library Science, Urbana-Champaign, IL: University of Illinois
- Lesk, M. E. (1964) *Procedures for statistical processing and request alteration*. Information storage and retrieval report ISR-7. Washington, DC: National Science Foundation
- Lesk, M. E. (1969) Word–word association in document retrieval systems. *American Documentation*, 20, 27–38
- Lesk, M. E. and Salton, G. (1968) *Relevance assessments and retrieval system evaluation*. Information storage and retrieval report ISR-14. Washington, DC: National Science Foundation
- Leydesdorff, L. (1987) Various methods for the mapping of science. *Scientometrics*, 11, 295–324
- Li, Z., Davis, H. and Hall, W. (1993) Hypermedia links and information retrieval. In *14th Information Retrieval Colloquium: Proceedings of the BCS 14th Information Retrieval Colloquium* (University of Lancaster, 13–14 April 1992), eds. T. McEnery and C. Paice. London: Springer-Verlag. In press
- McConnaughey, B. H. (1964) The determination and analysis of plankton communities. *Penelitian laut di Indonesia (Marine research in Indonesia)*, Special number, 1–40
- McGill, M., Koll, M. and Noreault, T. (1979) *An evaluation of factors affecting document ranking for information retrieval systems*. NTIS report PB80–119506. Springfield, VA: US Department of Commerce
- Markey, K. (1984) Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research*, 6, 155–177
- Maron, M. E. and Kuhns, J. L. (1960) On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery*, 7, 216–244
- Motyka, J., Dobrzanski, B. and Zawazki, S. (1950) Wstępne badania nad lagami południo-wschodniej Lubelszczyzny (Preliminary studies on meadows in the southeast of the province Lublin). *Universitas Mariae Curie-Skłodowska, Annales, Sectio E (Nauki Rolnicze)*, 5, 367–447
- Noreault, T., McGill, M. and Koll, M. B. (1981) A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In *Information retrieval research*, eds. R. N. Oddy et al., pp. 57–76. London: Butterworths
- Ochiai, A. (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletins of the Japanese Society for Scientific Fisheries*, 22, 526–530
- Parker-Rhodes, A. F. and Needham, R. M. (1960) *The theory of clumps*. Report 136. Cambridge: Cambridge Language Research Unit
- Raghavan, V. V. and Wong, S. K. M. (1986) A critical analysis of vector space model in information retrieval. *Journal of the American Society for Information Science*, 37, 279–287
- Rees, A. M. and Schultz, D. G. (1967) *A field experimental approach to the study of relevance assessments in relation to document searching, final report*. Cleveland, OH: Center for Documentation and Communication Research, School of Library Science, Case Western University
- Robertson, A. M. and Willett, P. (1992) *Identification of word variants in historical text databases*. London: British Library Research and Development Department
- Robertson, S. E. and Sparck Jones, K. (1976) Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129–146
- Robertson, S. E., Maron, M. E. and Cooper, W. S. (1982) Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1, 1–21

- Rogers, D. J. and Tanimoto, T. T. (1960) A computer program for classifying plants. *Science*, **132**, 1115–1118
- Rogers, H. J., and Willett, P. (1991) Searching for historical word forms in text databases using spelling correction methods: reverse error and phonetic coding methods. *Journal of Documentation*, **47**, 333–353
- Rohlf, F. J. (1974) Methods of comparing classifications. *Annual Review of Ecology and Systematics*, **5**, 101–113
- Rolling, L. (1981) Indexing consistency, quality and efficiency. *Information Processing and Management*, **17**, 69–76
- Russell, P. F. and Rao, T. R. (1940) On habitat and association of species of anopheline larvae in south-eastern Madras. *Journal of the Malaria Institute of India*, **3**, 153–178
- Sager, W. K. H. and Lockemann, P. C. (1976) Classification of ranking algorithms. *International Forum on Information and Documentation*, **1**, 12–25
- Salton, G. (1963) Associative document retrieval techniques using bibliographic information. *Journal of the Association for Computing Machinery*, **10**, 440–457
- Salton, G. (1968) *Automatic information organization and retrieval*. New York, NY: McGraw-Hill
- Salton, G. (ed.) (1971) *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall
- Salton, G. (1991) Developments in automatic text retrieval. *Science*, **253**, 974–980
- Salton, G. and Buckley, C. (1988) Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**, 513–523
- Salton, G. and Buckley, C. (1991) Global text matching for information retrieval. *Science*, **253**, 1012–1015
- Salton, G., Buckley, C. and Allan, J. (1992) Automatic structuring of text files. *Electronic Publishing: Origination, Dissemination and Design*, **5**, 1–17
- Salton, G. and McGill, M. J. (1983) *Introduction to modern information retrieval*. New York, NY: McGraw-Hill
- Salton, G., Wong, A. and Yang, C. S. (1975) A vector space model for automatic indexing. *Communications of the ACM*, **18**, 613–620
- Small, H. (1973) Cocitation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, **24**, 265–269
- Small, H. (1982) Citation context analysis. In *Progress in communication sciences*, **3**, eds. B. Dervin and M. Voight, pp. 287–310. Norwood, NJ: Ablex
- Small, H., and Sweeney, E. (1985) Clustering the Science Citation Index using co-citations: 1. A comparison of methods. *Scientometrics*, **7**, 391–409
- Smeaton, A. F. (1992) Information retrieval and hypertext: Competing technologies or complementary access methods. *Journal of Information Systems*, **2**, 221–233
- Sneath, P. H. A. and Sokal, R. R. (1973) *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco, CA: W H Freeman
- Sokal, R. R. (1961) Distance as a measure of taxonomic similarity. *Systematic Zoology*, **10**, 70–79
- Sokal, R. R. and Michener, C. D. (1958) A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin*, **38**, 1409–1438
- Sokal, R. R. and Rohlf, F. J. (1962) The comparison of dendrograms by objective means. *Taxon*, **11**, 33–40
- Sokal, R. R. and Sneath, P. H. A. (1963) *Principles of numerical taxonomy*. San Francisco, CA: W H Freeman
- Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter*, **5**, 1–34
- Sparck Jones, K. (1971) *Automatic keyword classification for information retrieval*. London: Butterworths
- Sparck Jones, K. (1973) Index term weighting. *Information Storage and Retrieval*, **9**, 619–633
- Stiles, H. E. (1961) The association factor in information retrieval. *Journal of the Association for Computing Machinery*, **8**, 271–279
- Tanimoto, T. T. (1958) An elementary mathematical theory of classification and prediction. *IBM Internal Report*, November 17
- Tenopir, C. (1988) Search strategies for full text databases. In *ASIS '88: Proceedings of the 51st Annual Meeting of the American Society for Information Science*, Volume 25, ed. C. L. Borgman and E. Y. H. Pai, pp. 80–86. Medford, NJ: Learned Information
- van Rijsbergen, C. J. (1977) A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, **33**, 106–119
- van Rijsbergen, C. J. (1979) *Information retrieval* 2nd ed. London: Butterworths
- Wang, Z. W., Wong, S. K. M. and Yao, Y. Y. (1992) An analysis of vector space models based on computational geometry. In *SIGIR '92: Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ed. N. Belkin, P. Ingwersen and A. M. Pejtersen, pp. 152–160. New York, NY: Association for Computing Machinery
- White, H. D. (ed.) (1990) Perspectives on author cocitation analysis. *Journal of the American Society for Information Science*, **41**, 429–468
- White, H. D., and Griffith, B. C. (1981) Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, **32**, 163–172
- White, H. D. and McCain, K. W. (1989) Bibliometrics. *Annual Review of Information Science and Technology*, **24**, 119–186
- Willett, P. (1987) *Similarity and clustering in chemical information systems*. Letchworth: Research Studies Press
- Willett, P. (1988) Recent trends in hierarchic document

- clustering: A critical review. *Information Processing and Management*, **24**, 577–597
- Wong, S. K. M. and Raghavan, V. V. (1984) Vector space model of information retrieval – a reevaluation. In *Research and development in information retrieval: Proceedings of the third joint BCS and ACM symposium*, ed. C. J. van Rijsbergen, pp. 167–185. Cambridge: Cambridge University Press
- Wong, S. K. M. and Yao, Y. Y. (1990) Query formulation in linear retrieval. *Journal of the American Society for Information Science*, **41**, 334–341
- Wong, S. K. M., Yao, Y. Y. and Bollmann, P. (1988) Linear structure in information retrieval. In *ACM SIGIR '88: Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, ed. Y. Chiaramella, pp. 219–232. New York, NY: Association for Computing Machinery
- Wong, S. K. M. et al. (1991) Evaluation of an adaptive linear model. *Journal of the American Society for Information Science*, **42**, 723–730
- Yates, F. (1934) Contingency tables involving small numbers and χ^2 test. *Journal of the Royal Statistical Society, Supplement I*, 217–235
- Yu, C. T., Lam, K. and Salton, G. (1982) Term weighting in information retrieval using the term precision model. *Journal of the Association for Computing Machinery*, **29**, 152–170
- Yule, G. U. (1900) On the association of attributes in statistics. *Philosophical Transactions of the Royal Society, A*, **194**, 257–319
- Yule, G. U. (1912) On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, **75**, 579–642
- Zubin, T. (1938) A technique for measuring like-mindedness. *Journal of Abnormal and Social Psychology*, **33**, 508–516