A Theory of Relevance for Automatic Document Classification

H. S. Heaps

Department of Computer Science, Sir George Williams University, Montreal, Canada

A method of automatic document classification is optimized to give best agreement with manually assigned classification of a set of test documents. The effectiveness of the automatic classification depends on the correlation between the relevance and the mutual keyword content of pairs of documents. An expression is obtained to measure the indexing worth of additional keywords. The resolution of the automatic classification is dependent on the correlation between mutual keyword content and the uniqueness of the association of categories with documents.

1. INTRODUCTION

The use of probabilistic indexing for classification of documents into given categories was introduced by Maron and Kuhns (1960), and applied by Maron (1961) through the computation of "attribute numbers." The use of factor analysis was described by Borko and Bernick (1963, 1964).

The method of attribute analysis is based on prediction of the probability that a document indexed by certain keywords will belong to a given category. It is usually applied in a form based on the assumption of statistical independence of the keywords associated with documents within each category. The assumption allows the frequencies of multiple keyword occurrences to be predicted from a knowledge of the probabilities of occurrence of single keywords in documents of each category. However, the assumption is not correct and it is difficult to assess its effect on the efficiency of the resulting document classification.

Document classification based on latent class analysis has been proposed by Baker (1962) and Winters (1965). The method depends on determination of eigenvalues of certain matrices whose elements are probabilities. If the eigenvalues are found to be real and to have values between 0 and 1 they may be used as probabilities that determine the latent class structure. Unfortunately, the equations that describe latent class structure are based on the assumption of statistical independence of keywords within each
category. Since such independence is seldom true the required eigenvalues are rarely found to lie between 0 and 1 and hence cannot be assumed to predict the required probabilities.

When documents are assigned to categories by use of nonautomatic methods the assignment is based on estimation of document content whether by examination of the title, the abstract, or the entire text. Automatic assignment of categories should therefore depend primarily on prediction of subject relevance rather than on assumed probability relations between keyword occurrences within separate categories.

In the present paper the problem of automatic document classification is first formulated as an optimization problem in which classification errors are to be minimized. The solution is then expressed in terms of statistical properties of the document database and the relation between categories. The mathematical techniques are similar to those used in a previous paper to determine the optimum form of question for use in information retrieval from a document database (Heaps, 1971). A similar approach has also been used to formulate a method of computer-aided medical diagnosis (Heaps, 1972).

The approach of the present paper is very general. It is supposed that automatic classification of a document may be effected through computation of a set of numbers that estimate the relevance of the document to each of a given set of categories. Each number is a function of the set of keywords used to index the document. Thus the relevance of the \( n \)-th document to the \( k \)-th category is expressed through an equation of the form

\[
y_k(n) = f_k(\{w(n)\}),
\]

where \( \{w(n)\} \) denotes the set of keywords used to index the \( n \)-th document.

A set of test documents \( \{D\} \) is assumed to have been assigned relevance numbers \( r_k(m) \) by nonautomatic means and is used to measure the effectiveness of the automatic classification scheme. Thus, for maximum effectiveness, the classification functions \( f_k \) should be chosen so that for the test documents the values of \( y_k \) are in best agreement with the \( r_k \).

For a general document, that is not necessarily one of the test documents, the \( y_k(n) \)'s are dependent on the keywords \( w(n) \), the set of test documents, and the relevance values \( r_j(m) \) assigned to the test documents. The dependency may be expressed in the form

\[
y_k(n) = f_k(\{w(n)\}, D, \{r_j(m)\}).
\]

The efficiency of the automatic classification of the test documents is...
measured by the extent to which the computed values of $y_k(n)$ are in agreement with the assigned values of $r_k(n)$. It is clearly dependent on the form assumed for the classification function $f_k$. In the present paper the functions $f_k$ are expressed in a general form in terms of unknown parameters that are chosen to minimize the mean square difference between the $y_k(n)$ and $r_k(n)$ for the best documents.

The method used in the present paper allows determination of the manner in which the effectiveness of automatic document classification is dependent on the choice of keywords used to index the documents, and also on the choice of categories as indicated by the assigned relevance values.

2. Procedure for Automatic Classification

Consider a set of documents $z(n)$ each of which is indexed by reference to a set of keywords. Each indexed document may be represented by a vector $w(n)$ whose $i$-th component $w_i(n)$ specifies the extent to which the $n$-th document is relevant to the $i$-th keyword. In the special instance that each $w_i(n)$ is chosen to be 1 or 0 then its value indicates the presence or absence of the keyword as an indexing term for the $n$-th document.

Suppose each document $z(n)$ is given a manual rating $r_k(n)$ of its relevance to the $k$-th category. Then with each document there is associated a rating vector $r(n)$ of dimension equal to the number of document categories.

Each document $z(n)$ is also given an automatic rating $y(n)$ whose components are linear functions of the $w_i(n)$. Thus

$$y_k(n) = \sum_i a_{ki}w_i(n)$$

(3)

where the $a_{ki}$ are the parameters of the linear process.

Let the parameters $a_{ki}$ be chosen so that, with respect to the entire set of test documents, the automatic ratings form the best mean-square approximations to the $r_k(n)$. The $a_{ki}$ must therefore be chosen to minimize

$$E = \frac{1}{N} \sum_n \frac{1}{K} \sum_k [y_k(n) - r_k(n)]^2$$

$$= \frac{1}{KN} \sum_n \sum_k \left[ \sum_i a_{ki}w_i(n) - r_k(n) \right]^2$$

$$= \frac{1}{K} \sum_k \left[ \sum_i \sum_j a_{ki}a_{kj}w_i(n) - 2 \sum_i a_{ki}w_i(n) + \frac{1}{N} \sum_n r_k(n)^2 \right],$$

(4)
where \( N \) is the number of documents, \( K \) is the number of categories, and

\[
 u_{ij} = u_{ji} = \frac{1}{N} \sum_n w_i(n) w_j(n),
\]

\[
 v_{ki} = \frac{1}{N} \sum_n r_k(n) w_i(n).
\]

The minimum value of \( E \) results when the \( a_{ij} \) are chosen to satisfy the equations \( \frac{\partial E}{\partial a_{ij}} = 0 \) for all combinations of \( i \) and \( j \). These equations have the form

\[
 (2/K) \sum_j a_{kj} u_{ji} - (2/K) v_{ki} = 0,
\]

and hence the \( a_{kj} \) must be chosen to satisfy the equations

\[
 \sum_j a_{kj} u_{ji} = v_{ki},
\]

which may be expressed in matrix form as

\[
 AU = V,
\]

and hence

\[
 A = VU^{-1}.
\]

Substitution of Eq. (8) into Eq. (4) shows that the minimum value of \( E \) is

\[
 E_{\text{min}} = \frac{1}{KN} \sum_k \sum_n r_k(n)^2 - \frac{1}{K} \sum_i \sum_k a_{ki} v_{ki}.
\]

The element \( u_{ij} \) is a measure of the extent to which the \( i \)-th and \( j \)-th keywords tend to be associated with the same documents. If each \( w_k(n) \) is always chosen to be 1 or 0 then \( Nu_{ij} \) is equal to the number of documents indexed by both the \( i \)-th and \( j \)-th keywords. Similarly \( Nv_{ki} \) then represents the number of documents that contain the \( i \)-th keyword and are in the \( k \)-th category.

In many instances the matrix elements \( u_{ij} \) for \( i \neq j \) are small in comparison to the diagonal elements. The matrix \( U^{-1} \) may then be expanded to the form

\[
 U^{-1} = (U_d + U_n)^{-1}
 = U_d^{-1} - U_d^{-1} U_n U_d^{-1} + U_d^{-1} U_n U_d^{-1} U_n U_d^{-1} - \cdots,
\]
where \( U_d \) is a diagonal matrix and \( U_n \) has zero diagonal elements. Equation (10) may then be written as

\[
a_{ij} = \frac{v_{ij} - \frac{1}{u_{jj}} \sum_{k \neq j} u_{ik} u_{kj}}{u_{jj} - \frac{1}{u_{jj}} \sum_{k \neq j} u_{ik} u_{kj}} + \frac{1}{u_{jj}} \sum_{k \neq j} \sum_{r \neq j, k} \frac{u_{kr} u_{rj}}{u_{rr}},
\]  

subject to neglect of higher order terms.

In the special instance that the \( w_{i}(n) \) have values only of 1 or 0 the terms in (13) may be interpreted as probabilities, so that

\[
a_{ij} = p(C_i | W_j) - \sum_{k \neq j} p(C_i | W_k) p(W_k | W_j)
\]

\[+ \sum_{k \neq j, k} \sum_{r \neq j, k} p(C_i | W_k) p(W_k | W_r) p(W_r | W_j),
\]

where

\[p(C_i | W_j) = \text{probability that a document which contains the} \]

\[j\text{-th keyword } W_j \text{ belongs to the } i\text{-th category } C_i; \]

and

\[p(W_k | W_j) = \text{probability that a document which contains the} \]

\[j\text{-th keyword also contains the } k\text{-th keyword.} \]

3. Mutual Relevance and Keyword Content of Documents

When the parameters \( a_{kj} \) are chosen to satisfy (8) the value of \( E \) is a minimum and given by (11). Substitution of the first two terms of \( a_{ki} \) from (13) leads to

\[E_{min} = \frac{1}{KN} \sum_{k} \sum_{n} r_k(n)^2 - \frac{1}{K} \sum_{i,j} \frac{v_{ij}^2}{u_{jj}} + \frac{1}{K} \sum_{i,j,k} \frac{u_{ij} v_{ik} v_{kj}}{u_{ii} u_{jj}}.\]

If the \( u_{ij} \) and \( v_{ij} \) are represented as the sums (5) and (6) then \( E_m \) may be expressed as follows, subject to neglect of higher order terms

\[E_{min} = (1/N) \sum_{n} r(n, n) - (1/N^2) \sum_{m} \sum_{n} r(m, n)[w(m, n) - c(m, n)],\]
and where

\[ r(m, n) = \frac{1}{K} \sum_{i} r_i(m) r_i(n), \]  

\[ w(m, n) = \sum_{j} \frac{w_j(m) w_j(n)}{(1/N) \sum_{s} w_j(s)^2}, \]  

\[ c(m, n) = \frac{1}{N} \sum_{j,k,s \neq k} \frac{w_j(m) w_j(s) w_k(s) w_k(n)}{u_{js} u_{kk}}. \]  

The function \( r(m, n) \) provides a measure of the relevance of the \( m \)-th to the \( n \)-th document as indicated by their mutual associations through relevance to the same categories. In contrast, the function \( w(m, n) \) measures the similarity between the \( m \)-th and \( n \)-th documents as indicated by their association with the same keywords.

The quantity \( w(m, n) \) measures the extent of the keyword vocabulary associated with both the \( m \)-th and \( n \)-th documents. The value of \( w(n, n) \) is a measure of the amount of keyword vocabulary associated with the \( n \)-th document. The denominator in (18) introduces a normalizing effect so that each factor \( w_j(m) \) in the numerator affects the summation according to its value with respect to the root-mean-square relevance of all documents to the \( j \)-th keyword.

The function \( w(m, n) \) will be termed the “mutual keyword content” of the \( m \)-th and \( n \)-th documents. Similarly \( w(n, n) \) will be termed the “keyword content” of the \( n \)-th document.

The “mutual relevance” between the \( m \)-th and \( n \)-th documents will be defined to be the value of \( r(m, n) \). The value of \( r(n, n) \) is the mean-square value of the relevance of the \( n \)-th document to each category and may be termed the “relevance” of the document. A document that is classed as highly relevant to many categories will have a high value of “relevance” and may be regarded as important from many points of view.

The function \( c(m, n) \) may be termed the “connectivity” between the \( m \)-th and \( n \)-th documents. It measures the extent to which keywords relevant to the \( m \)-th document, and different keywords relevant to the \( n \)-th document, are also relevant to the same other documents. The value of \( c(m, n) \) measures the extent to which the \( m \)-th document contains pairs of keywords common to other documents.

With optimum choice of the \( a_{kj} \) the automatic rating \( y(n) \) of the \( n \)-th document may also be expressed in terms of the functions \( r(m, n) \) and \( w(m, n) \).
Thus
\[ y_s(n) = \left(\frac{1}{N}\right) \sum_m \left[ w(m, n) - c(m, n) \right] r_s(m) \]
\[ = \left(\frac{1}{N}\right) \sum_m s(m, n) r_s(m), \quad (20) \]

where \( s(m, n) \) denotes \( w(m, n) - c(m, n) \) and may be called the "significant keyword content" common to the \( m \)-th and \( n \)-th documents.

The function \( s(m, n) \) is a measure of the extent to which the \( m \)-th and \( n \)-th documents contain common keywords but do not contain keywords mutually common to other documents.

In terms of \( s(m, n) \) the expression for \( E_{\text{min}} \) may be written in the form
\[ E_{\text{min}} = (\text{Average Document Relevance}) - \left(\frac{1}{N^2}\right) \sum_m \sum_n s(m, n) r(m, n), \quad (21) \]

which illustrates how \( E_{\text{min}} \) may be reduced by choosing the keywords so that the significant keyword content between documents is highly correlated with their mutual relevance.

Equation (21) may be used to compare the relative suitability of title words versus abstract words for use as indexing words for document classification. The indexing words should be chosen from the set for which
\[ \left(\frac{1}{N^2}\right) \sum_m \sum_n s(m, n) r(m, n) \quad (22) \]
assumes its maximum value.

It may be remarked that if further terms are included in (13) their effect is to add further terms to \( w(n, m) - c(m, n) \) in (16) and (20). Such additional terms measure the extent to which there are higher order connectivity links between the \( m \)-th and \( n \)-th documents. If these terms are included in the definition of \( s(m, n) \) then Eqs. (20) and (21) become quite general and not dependent on neglect of higher order terms in (13).

4. INDEXING WORTH OF A KEYWORD

Suppose that a subset of the keywords has been used to determine either the optimum \( a_{ki} \) for use in (3) or the optimum \( s(m, n) \) for use in (20). Addition
of a further indexing word \( W_q \), and subsequent determination of the optimum \( a_{ki} \), has the effect of decreasing \( E_{\text{min}} \) in (16) by

\[
\Delta q = \frac{1}{N^2} \sum_{m,n} r(m, n) \left[ \frac{w_q(m) w_q(n)}{u_{qq}} - \frac{2}{N} \sum_s \sum_{k \neq q} \frac{w_k(m) w_k(s) w_k(n)}{u_{kk}} \right]
\]

\[
= \frac{1}{N^2 u_{qq}} \sum_m \sum_n r(m, n) \left[ w_q(m) w_q(n) - \frac{(2/N)}{s} \sum_s w(s, n) w_q(m) w_q(s) \right]
\]

\[
= \frac{1}{N^2 u_{qq}} \sum_m \sum_n w_q(m) w_q(n) \left[ r(m, n) - \frac{(2/N)}{s} \sum_s r(m, s) w(n, s) \right]. \quad (23)
\]

In (23) the function \( w(n, s) \) denotes the value computed in the absence of the \( q \)-th keyword. Addition of this keyword as an index term allows \( E_{\text{min}} \) to be reduced by \( \Delta q \), which is therefore a measure of the “worth” of the \( q \)-th keyword as an additional indexing term. Worth, defined in this manner, is a function not only by the \( q \)-th keyword, but also of the keywords already in use.

The worth \( \Delta q \) is the difference of two terms. The first term is the summation of \( w_q(m) w_q(n) r(m, n) \) and measures the extent to which the \( q \)-th keyword occurs in mutually relevant documents. The second term is the summation of \( w_q(m) w_q(n) r(m, s) w(n, s) \), and measures the extent to which one of the documents \( z(n) \) has mutual keyword content with documents \( z(s) \) that are relevant to the other document \( z(m) \). Thus, to have high worth as an additional indexing term a keyword should occur in a large number of relevant documents that do not already have in common a large number of indexing words.

The form of (23) suggests the manner in which further indexing words should be chosen. The pairs of documents for which

\[
r(m, n) - \frac{(2/N)}{s} \sum_s r(m, s) w(n, s) \quad (24)
\]

assumes its largest values should be examined for the presence of keywords in common since these are the keywords whose use as index terms will effect the greatest reduction in the value of \( E_{\text{min}} \).

Similarly, for a keyword \( W_q \) already in use as an index term for automatic determination of categories, if the expression (23) is found to be of negligible value then the keyword is redundant as an index term. In computation of (23) for an existing index word \( W_q \) the value of \( w(n, s) \) should be reduced by \( w_q(n) w_q(s)/u_{qq} \) to reduce to its value in the absence of the \( q \)-th index word.
5. Resolution of the Automatic Classification

The manual ratings \( r_i(n) \) and \( r_j(n) \) describe the relevance of the \( n \)-th document to the \( i \)-th and \( j \)-th categories respectively. A large value of \( |r_i(n) - r_j(n)| \) results for each document that is highly relevant to only one of the \( i \)-th or \( j \)-th categories. Thus the following expression provides a measure of the number of documents that are manually rated as relevant to either the \( i \)-th or \( j \)-th category but not to both

\[
\sum_n [r_i(n) - r_j(n)]^2. \tag{25}
\]

The similar expression in terms of the automatic ratings may be computed as follows in which the formula (20) is used with neglect of the \( c(m, n) \) term

\[
\sum_n [y_i(n) - y_j(n)]^2 = \frac{1}{N^2} \sum_{m, s} [r_i(m) - r_j(m)][r_i(s) - r_j(s)] \sum_n w(m, n) w(s, n)
\]

\[
= \frac{1}{N} \sum_{m, s} [w(m, s) + c(m, s)][r_i(m) - r_j(m)][r_i(s) - r_j(s)]
\]

\[
= \frac{1}{N} \sum_{m, s} w(m, s)[r_i(m) - r_j(m)][r_i(s) - r_j(s)], \tag{26}
\]

subject to neglect of the term \( c(m, s) \).

The resolution of the automatic classification may be defined as the ability of the system to assign sufficiently distinct category relevances to a document whose corresponding manual ratings are sufficiently distinct. The average resolution may therefore be measured by the ratio of (26) to (25). Hence, if \( r_{ij}(n) \) is used to denote \( r_i(n) - r_j(n) \), the resolution may be expressed in the form

\[
R = \frac{\sum_m \sum_s w(m, s) r_{ij}(m) r_{ij}(s)}{N \sum_n r_{ij}(n)^2}. \tag{27}
\]

6. More General Classification Function

The linear classification function of type (3) may be generalized to the form

\[
y_k(n) = \sum_i a_{ki} w_i(n) + \sum_{i,j \neq j} a_{kj} w_i(n) w_j(n), \tag{28}
\]
which allows the automatic rating to depend on the occurrence of word pairs in a manner that is not a linear function of the occurrence of the two separate words.

However, the function \( y_k(n) \) of (28) is still a linear function of the \( a_{kij} \).

Thus the analysis of Sections 2–3 may be repeated with the understanding that each summation of the index \( i \) in \( a_{kij} \) is to be extended to include summation over the index pair \( i, j \) in \( a_{kij} \).

Equation (16) remains valid provided \( w(m, n) \) is redefined as

\[
w(m, n) = \sum_j \frac{w_j(m) w_j(n)}{(1/N) \sum_s w_j(s)^2} + \sum_{i \neq j} \frac{w_i(m) w_i(n) w_j(m) w_j(n)}{(1/N) \sum_s w_i(s)^2 w_j(s)^2}, \tag{29}
\]

with a similar extension of the definition of \( c(m, n) \).

The effect of adding a single nonlinear term \( w_j(n) w_j(n) \) to (3) is to reduce \( E_{\text{min}} \) by \( \Delta q \) of (23) in which \( w_i(m) w_o(n) \) is replaced by \( w_i(m) w_j(n) w_j(m) w_j(n) \).

The indexing worth \( \Delta q \) of the nonlinear term is dependent on the number of occurrences of the pair of words in relevant documents that do not already have in common a large number of indexing words. However, all documents that contain the pair of words also contain the index words counted separately. Thus, in general the worth of adding a nonlinear term to (3) is likely to be less than the worth of a single new indexing term.

It is therefore believed that the form of (23) suggests that, in general, it may prove satisfactory to base automatic document classification on use of linear functions of form (3) applied to a large set of keywords. In many instances this is likely to be more efficient than use of a smaller number of keywords with nonlinear classification functions of type (28).

Received: April 9, 1971; Revised: September 19, 1972

References


