Pergamon

Information Processing & Management, Vol. 33, No. 1, pp. 1–14, 1997

0306-4573/97 $17.00+0.00

S0306-4573(96)00043-X

PERFORMANCE STANDARDS AND EVALUATIONS IN IR TEST COLLECTIONS: CLUSTER-BASED RETRIEVAL MODELS

W. M. SHAW Jr,[1] ROBERT BURGIN[2] and PATRICK HOWELL[1]

[1] School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, U.S.A. and [2] School of Library and Information Sciences, North Carolina Central University, Durham, NC 27707, U.S.A.

(Received August 1995; accepted April 1996)

Abstract—Low performance standards for the group of queries in 13 retrieval test collections have been computed. Derived from the random graph hypothesis, these standards represent the highest levels of retrieval effectiveness that can be obtained from meaningless clustering structures. Operational levels of cluster-based performance reported in selected sources during the past 20 years have been compared to the standards. Comparisons show that typical levels of operational cluster-based retrieval can be explained on the basis of chance. Indeed, most operational results in retrieval test collections are lower than those predicted by random graph theory. A tentative explanation for the poor performance of cluster-based retrieval reveals weaknesses in both fundamental assumptions and operational implementations. The cluster hypothesis offers no guarantee that relevant documents are naturally grouped together, clustering algorithms may not reveal the inherent structure in a set of documents, and retrieval strategies do not reliably retrieve the most effective cluster or clusters of documents. That most cluster-based retrieval implementations implicitly rely on topical relatedness to be equivalent to a relevance relationship contributes to the poor performance. Clustering strategies capable of adapting to relevance information may succeed where static clustering techniques have failed. Copyright © 1997 Elsevier Science Ltd

INTRODUCTION

Retrieval test collections consist of a set of documents, a set of queries, and a subset of the document collection considered to be relevant to each query. Relevance judgments are generally provided by subject experts. Retrieval experimentation focuses on the capacity of combinations of document representation, query representation, and retrieval strategy to discriminate relevant documents associated with each query from all other documents in such a database. Retrieval performance values derived from diverse test collections with fixed queries and relevance evaluations can, at least in principle, be presented on a measurement scale bounded by conditions defining 'complete failure', no relevant document retrieved, and 'complete success', all and only relevant documents retrieved. Reporting optimal, operational, and 'expected' (by chance) levels of performance on such a scale could influence a research agenda for the field. For any retrieval model, high levels of optimal performance, such as those detected for probabilistic models (Shaw, 1995), accompanied by low operational results, call attention to weaknesses in the implementation, but they also suggest that progress can be made by understanding why the system fails to meet expectations. Low levels of optimal performance, such as those detected for cluster-based retrieval (Burgin, 1995), challenge fundamental assumptions of the model. Distinguishing between assumptions based on mathematical or conceptual convenience and assumptions grounded in empirical reality, for example, can be expected to lead to effective retrieval models. Whether dictated by poor designs or inappropriate assumptions, low levels of operational performance are difficult to interpret. The distinction between acceptable, but low, and unacceptably low levels of performance, for example, is

unclear. Such distinctions could be facilitated by establishing low performance standards for queries in retrieval test collections. Such standards would guard against the fallacy that small but 'statistically significant' improvements in retrieval performance are meaningful, when neither value exceeds expectations based on chance.

Reported here are average low peformance standards for the group of queries in 13 traditional retrieval test collections. Standards are expressed in terms of a composite measure of retrieval effectiveness varying from 0, indicating complete failure, to 1, representing complete success. Derived from random graph theory, low performance standards are based on the highest level of retrieval effectiveness attributable to meaningless clustering structures and are directly comparable to operation cluster-based results reported in the literature. Comparisons challenge both fundamental assumptions and operational implementations of cluster-based retrieval.

PERFORMANCE MEASURES

In the context of retrieval experimentation, recall and precision are suitable measures of retrieval performance, representing, respectively, the fraction of relevant documents that is retrieved and the fraction of retrieved documents that is relevant. Both measures are required to demonstrate the extent to which relevant documents have been distinguished from non-relevant documents. Computation of recall and precision for unranked retrieval outcomes is generally straightforward, given the set of retrieved documents and the set of relevant documents. Complications can arise, however, if choices are presented for the set of documents to be retrieved. For example, efforts to establish the optimal performance of cluster-based retrieval confront the problem of selecting the 'best' cluster for each query from the set of available clusters (Burgin, 1995; Shaw, 1994). The absence of expressions of information need in retrieval test collections provides no guidance for selecting a cluster favoring recall or precision. Objectives of retrieval experimentation are met by retrieving the cluster with the highest combination of recall and precision values. Selecting such a cluster can be facilitated by a composite measure of retrieval effectiveness.

Dice's coefficient (Dice, 1945) can be used to measure the similarity of the set of documents retrieved in response to a query and the set of documents relevant to that query, and it provides a measure of retrieval effectiveness (F) given by

$$F=\frac{2r}{n+R_{\mathrm{d}}}, \tag{1}$$

where r is the number of documents retrieved and relevant, n is the number of documents retrieved, and R_{d} is the total number of relevant documents. The complement of Dice's coefficient, referred to as the normalized symmetric difference and denoted by E, is customarily employed in cluster-based retrieval as a measure of retrieval effectiveness (Van Rijsbergen, 1979, p. 167). Neither mathematical properties nor practical considerations compel the adoption of E (Shaw, 1986); the complement of Dice's coefficient is not a distance measure, satisfying metric conditions, and the resulting function is inversely related to retrieval performance, which is awkward.

Dividing the numerator and denominator of eqn (1) by r leads immediately to a composite measure of retrieval effectiveness expressed in terms of R and P:

$$F=\frac{2}{\frac{1}{R}+\frac{1}{P}}. \tag{2}$$

The magnitude of F varies from 0 to 1 and is directly related to retrieval performance. When no relevant document is retrieved, $R=P=0$ and $F=0$; when all relevant documents and only relevant documents are retrieved, $R=P=1$ and $F=1$. The resulting effectiveness measure (F) is the harmonic mean of R and P (Kenney, 1947, pp. 50–59). Properties of the harmonic mean

Table 1. Illustration of the relationship among harmonic, geometric, and arithmetic means for hypothetical values of recall and precision

Recall (R)	Precision (P)	Harmonic mean (F)	Geometric mean	Arithmetic mean
0.10	0.90	0.18	0.30	0.50
0.20	0.80	0.32	0.40	0.50
0.30	0.70	0.42	0.46	0.50
0.40	0.60	0.48	0.49	0.50
0.50	0.50	0.50	0.50	0.50

favor its use as a composite measure of retrieval effectiveness in relation to other common mean values. For the same set of data, the harmonic, geometric, and arithmetic means are ordered in magnitude; the harmonic mean is less than or equal to the geometric mean, and the geometric mean is less than or equal to the arithmetic mean. The means are equal when all values in the dataset have the same magnitude (Bullen *et al.*, 1988, chap. II). The arithmetic mean weights the magnitudes of all values equally, while the geometric, or especially the harmonic, mean weights low values more heavily than high values. In the present context, low values of F occur when both R and P are low, or when either R or P is low and the other is high; high values of F occur only when both R and P are high, as illustrated in Table 1. Selecting the unranked retrieval outcome that maximizes F is consistent with the objectives of retrieval experimentation.

LOW PERFORMANCE STANDARDS

The success of cluster-based retrieval depends upon several assumptions. Firstly, pairs of documents, associated by the similarity of their representations, must impose clustering structure on the documents constituting a database. The presence or absence of clustering structure in a set of data elements can be assessed by hypotheses derived from random graph theory (Dubes & Jain, 1979, 1980), and evidence for clustering structure in retrieval test collections has been investigated for subject, citation, and combinations of subject and citation representations (Shaw, 1990a, b, 1991a, 1993). Secondly, clustering structure resulting from pairwise associations must manifest a tendency for documents relevant to the same query to be grouped together and documents relevant to distinct queries to be separated in some document space. Providing the basis upon which cluster-based retrieval is founded, the cluster hypothesis asserts that 'closely associated documents tend to be relevant to the same requests' (Van Rijsbergen, 1979, p. 45). The validity of the cluster hypothesis has been tested in a variety of information retrieved (IR) test collections (Van Rijsbergen & Sparck Jones, 1973; Voorhees, 1985; El-Hamdouchi & Willett, 1987). Thirdly, implemented by a clustering algorithm and applied to an array of pairwise associations among documents, the clustering criterion must be capable of revealing the natural groupings or clusters of relevant documents posited by the cluster hypothesis (Dubes & Jain, 1980). Finally, the cluster search strategy must retrieve one or more suitable cluster of documents in response to a query. The operational and optimal effectiveness of cluster-based retrieval techniques have been tested for a variety of clustering criteria and search strategies (Burgin, 1995; Willett, 1988).

The first assumption is fundamental to subsequent considerations. If there is no clustering structure in a document collection, documents relevant to a query are no more likely to be associated than any other pair of documents, clusters are an artifact of the clustering technique rather than a representation of inherent structure in the data, and retrieval strategies are confronted with the hopeless task of selecting from a group of clusters distinguished, perhaps, only by their size. Predictions derived from random graph theory offer evidence for the presence or absence of clustering structure in a document collection and provide an opportunity to establish low performance standards in the context of cluster-based retrieval.

A set of documents, referred to as 'points', and a set of document pairs, referred to as 'lines', constitute a document graph. A document graph is disconnected if the documents are distributed

among a set of distinct, connected subgraphs, referred to as 'components'. For example, a document graph consisting of N_d documents and one line ($q=1$) is disconnected and includes one component with two documents and N_d-2 components with one document. A document graph with one component is connected (Harary, 1969, chap. 2; Shaw, 1990a; Van Rijsbergen, 1979, pp. 48–50).

Under the random graph hypothesis, it is assumed that the lines of a graph are randomly selected from the set of all possible lines (Dubes & Jain, 1979, 1980). The assumption defines a set of points and lines for which there is no clustering structure and invalidates the cluster hypothesis. By constructing a large number of random graphs with specified numbers of points and lines, expected properties of the random structure, including the effectiveness of cluster-based retrieval, can be computed (Ling, 1973, 1975; Ling & Killough, 1976; Shaw, 1990a, c, 1991b, 1993, 1994). In the present application, the number of points in the random graph is dictated by the number of documents in the test collection being investigated (N_d), and the number of lines (q) is varied from one to a large number, insuring a complete pattern of results. Composed of N_d documents and q randomly selected pairwise associations, each random graph is constructed by the single-link clustering criterion. For each random graph, every available component with two or more documents, hereinafter referred to as a 'cluster', is tested as a response to each query; the cluster with the highest effectiveness (F) value is retrieved for each query; and average values of recall, precision, and effectiveness are computed for the complete set of queries. The process is completed for a sufficiently large number of random graphs to insure stable outcomes, and the mean values of average recall, average precision, and average effectiveness (referred to as 'expected, average values') are computed as a function of q for each value of N_d.

Expected, average values of recall, precision, and effectiveness represent the highest, average cluster-based performance that can be produced by a meaningless structure at one level of an arbitrary single-link hierarchy with N_d documents and q pairwise associations. Nevertheless, the expected, average performance values represent conservative low performance thresholds in this context; higher levels of random performance than those defined here can be derived from complete hierarchies and alternative clustering criteria (Burgin, 1995).

Low performance standards derived from the random graph hypothesis have two disadvantages. Extensive computational effort renders these measures impractical for large databases. Tied to a specific retrieval model, the resulting performance standards might also lack the generality required for adoption. Directly comparable to cluster-based retrieval outcomes, however, these standards offer surprising insights, as will be demonstrated.

RESULTS

Collection statistics

Descriptions of 13 test collections investigated here can be found in several sources (Davies, 1983; Fox, 1983; Shaw *et al.*, 1991; Sparck Jones & Van Rijsbergen, 1976). Collection statistics, pertinent to the present investigation, are presented in Tables 2 and 3. Summary statistics for the cystic fibrosis (CF) test collection (Shaw *et al.*, 1991) reported in Table 2 require explanation.

The CF test collection includes 100 queries with exhaustive relevance evaluations from physicians and scientists involved in CF care and research. Three subject experts examined the full-text of each document; judged the document to be 'highly relevant', 'marginally relevant', or 'not relevant' to each query; and assigned relevance scores of 2, 1, or 0, respectively. In the CF test collection, highly relevant documents provide an answer or partial answer to the query, marginally relevant documents are topically related but do not directly address the question, and non-relevant documents are unrelated to the question. A relevance-weight threshold (RT) is used to control the specificity of the relevance requirement, a document being considered relevant to a query if the sum of its relevance scores is greater than or equal to RT. Low levels of RT signify

Table 2. CF test collection statistics

Test collection	Number of documents (N_d)	Relevance threshold (RT)	Number of queries	Average number of relevant documents ($\bar{r}_q$)
CF	1239	1	100	31.9
CF	1239	2	100	18.1
CF	1239	3	99	14.9
CF	1239	4	99	14.1
CF	1239	5	99	10.7
CF	1239	6	94	6.4

a comprehensive search, in which marginally relevant documents have meaning, and high values of RT signify a specific search, in which only highly relevant documents have merit. For example, at $RT=1$, a document is considered relevant to a query if at least one expert considers it to be topically related; at $RT=6$, a document is considered relevant only if all three experts agree that it provides an answer or partial answer to a query.

For the 12 other test collections investigated here, a document is considered to be relevant or not relevant to a query, and $RT=1$ is reported in Table 3 for consistency with Table 2. Because of the strong influence of the average number of relevant documents per query ($\bar{r}_q$) on the magnitude of low performance standards, test collections in Table 3 are listed in inverse order of the magnitude of $\bar{r}_q$ for consistency with the presentation of other results.

Performance standards

Expected, average, cluster-based effectiveness has been computed as a function of RT and the number of randomly generated lines (q) in the CF test collection, and the results are shown in Fig. 1. Results manifest two general patterns of interest in this investigation. First, for each of the six values of RT, the expected, average effectiveness of random structures is low when there are few lines, reaches a maximum for an intermediate number of lines, and is constant and low for any number of lines beyond some critical number. The pattern is expected. A small number of randomly selected pairwise associations (q) can only produce a few small clusters, including few relevant documents. For example, random graphs with only one line ($q=1$) can produce only one cluster with two documents as a response to every query, leading to a minimal level of effectiveness for most queries ($F=0$). For a sufficiently large number of lines, random graphs are connected, all documents appear in one cluster, and the entire database is retrieved as a response to each query, leading to a constant, low level of average retrieval performance for any

Table 3. IR test collection statistics

Test collection	Number of documents (N_d)	Relevance threshold (RT)	Number of queries	Average number of relevant documents ($\bar{r}_q$)
UKCIS	27,361	1	182	58.9
CISI	1460	1	76	41.0
INSPEC	12,684	1	77	33.0
MED	1033	1	30	23.2
EVANS	2542	1	39	23.1
HARDING	2472	1	65	22.6
NPL	11,429	1	93	22.4
CACM	3204	1	52	15.3
KEEN	800	1	63	14.9
LISA	6004	1	35	10.8
CRAN	1400	1	225	8.2
TIME	425	1	83	3.9

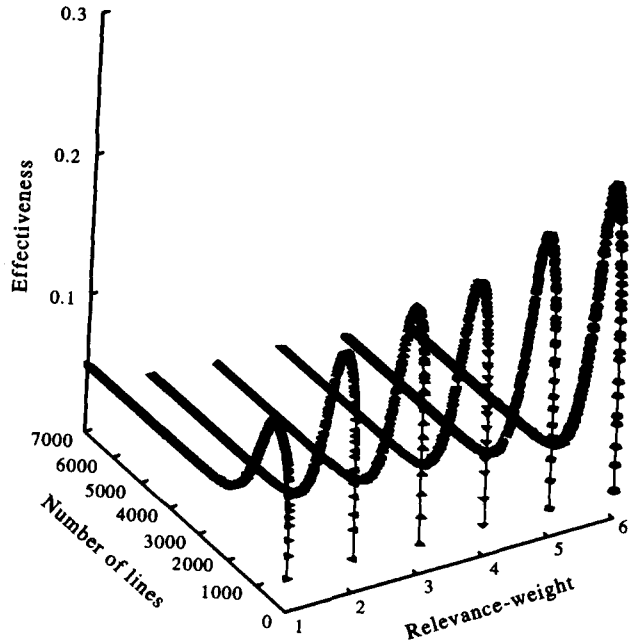

Fig. 1. Expected, average effectiveness, derived from the random graph hypothesis, as a function of the number of lines (q) in random graphs and relevance-weight thresholds (RT) in the CF test collection.

number of lines exceeding the critical value. The expected, average effectiveness is maximized when the number of clusters in random graphs is maximized (Shaw, 1990a). The few large clusters and many small clusters provide an opportunity for finding a cluster with at least one relevant document for most queries. The second pattern of interest is the systematic increase in the maximum, expected, average effectiveness—hereinafter referred to as 'expected effectiveness', for simplicity—as RT is increased from 1 to 6. A tentative explanation for this pattern can be inferred from interactions of cluster size and number of relevant documents per query. As RT is increased, and the average number of relevant documents per query is decreased, retrieving one small cluster with perhaps one relevant document increases expected recall for relatively constant levels of expected precision, causing the associated expected effectiveness to increase. For example, consider a randomly generated cluster with two documents and suppose one is relevant to a query. As the number of relevant documents associated with the query decreases, expected precision remains constant ($P=0.50$), while expected recall and effectiveness increase. An examination of the interacting factors is informative.

Table 4 gives the expected number of documents retrieved ($\bar{n}_s$), the expected number of relevant documents retrieved ($\bar{r}_s$), the expected recall ($\bar{R}_s$), the expected precision ($\bar{P}_s$), and the

Table 4. Expected random graph retrieval outcomes in the CF test collection

Test collection	Expected number retrieved* ($\bar{n}_s$)	Expected number retrieved and relevant* ($\bar{r}_s$)	Expected recall ($\bar{R}_s$)	Expected precision ($\bar{P}_s$)	Expected effectiveness ($\bar{F}_s$)
CF ($RT=1$)	10.7	3.7	0.1154	0.3454	0.1263
CF ($RT=2$)	5.8	2.4	0.1324	0.4113	0.1636
CF ($RT=3$)	6.1	2.4	0.1624	0.3965	0.1841
CF ($RT=4$)	5.5	2.3	0.1628	0.4148	0.1911
CF ($RT=5$)	5.9	2.3	0.2099	0.3783	0.2108
CF ($RT=6$)	5.6	1.9	0.2921	0.3321	0.2360

*The expected number of documents retrieved and relevant ($\bar{r}_s$) is estimated by the product $\bar{r}_q \times \bar{R}_s$ and the expected number of documents retrieved ($\bar{n}_s$) is estimated by the quotient $\bar{r}_s / \bar{P}_s$.

associated expected effectiveness ($\bar{F}_x$) as a function of RT in the CF test collection.* As anticipated, the expected number of documents retrieved and the expected number of relevant documents retrieved are small and relatively constant for most values of RT; an exception occurs at $RT=1$, where $\bar{n}_x$ is almost twice the number for all other relevance thresholds. It can also be seen that expected effectiveness increases, as RT is increased, because expected recall increases steadily while expected precision remains relatively constant. Because the highest level of effectiveness attributable to meaningless structures is sought, the expected effectiveness, defined here, constitutes the low performance standard for the cluster-based retrieval model and the CF test collection.

The expected, average cluster-based effectiveness has been computed as a function of q for 12 other test collections, and the results are shown in Figs 2(a) and 2(b). To examine the influence of the average number of relevant documents per query ($\bar{r}_q$) when database size varies, test collections are ordered inversely by the magnitude of $\bar{r}_q$ and numbered accordingly; test collection 1 (UKCIS) in Fig. 2(a) has the highest value ($\bar{r}_q=58.9$), and test collection 12 (TIME) in Fig. 2(b) has the lowest value ($\bar{r}_q=3.9$), as shown in Table 3. Results are essentially consistent with those presented for the CF test collection. The (maximum) expected effectiveness ($\bar{F}_x$) can be detected for each test collection and generally increases as the average number of relevant documents per query decreases.

The influence of database size is also evident in Figs 2(a) and 2(b). As the number of documents in a test collection increases, the distribution of expected effectiveness values becomes more uniform. Consider the UKCIS(1) and TIME(12) test collections in Figs 2(a) and 2(b), respectively. UKCIS(1) has the greatest number of documents ($N_d=27{,}361$), and expected effectiveness reaches a maximum value at approximately 9000 lines and remains essentially constant until about 12,000 lines, after which the value declines; it is the only test collection for which the maximum value is not defined by a single point on the curve. INSPEC(3) has the second greatest number of documents ($N_d=12{,}684$), and the maximum value, which is difficult to detect visually in Fig. 2(a), occurs at $q=4645$. The TIME(12) test collection has the fewest number of documents ($N_d=425$), and the distribution is sharply spiked, with the (maximum) expected effectiveness occurring at $q=134$. Visual inspection of the figures suggests that test collection 9 (KEEN) might have the second fewest number of documents, which is correct. Table 5 gives the expected number of documents retrieved ($\bar{n}_x$), the expected number of relevant documents retrieved $\bar{r}_x$, the expected recall ($\bar{R}_x$), expected effectiveness ($\bar{F}_x$) for the 12 test collections, which are listed in descending order of the average number of relevant documents per query ($\bar{r}_q$). Although the character of the relevance evaluations in these distinct test collections can be expected to vary, statistical characteristics influence retrieval performance, as in the CF test collection. The expected effectiveness associated with the test collections generally increases as the average number of relevant documents ($\bar{r}_q$) decreases. Excluding the UKCIS and CISI test collections, with exceptionally high average numbers of relevant documents per query ($\bar{r}_q=58.9$ and $\bar{r}_q=41.0$, respectively), and the CRAN and TIME test collections, with exceptionally low values ($\bar{r}_q=8.2$ and $\bar{r}_q=3.9$, respectively), the increase in expected effectiveness can be attributed to increases in expected recall, while expected precision remains relatively constant, as in the CF test collection (Table 4).

Test collection, average number of relevant documents per query ($\bar{r}_q$), and expected effectiveness values ($\bar{F}_x$), derived from the random graph hypothesis, are summarized in Table 6. Reported to the customary degree of accuracy for IR research, these values represent low performance standards for 13 test collections, including six for the CF database. These standards are directly comparable to operational cluster-based retrieval effectiveness. These standards can also be compared to standards derived from the hypergeometric distribution (Shaw *et al.*, 1997). Although based on distinct assumptions, similar values for the same test collection suggest that alternative interpretations of chance have identified a general standard for retrieval performance.

* At $RT=2$ and $RT=5$, two combinations of expected recall and precision produce the same expected effectiveness. In these and all subsequent such cases, the combination of expected recall and precision minimizing the difference between expected recall and precision is reported.

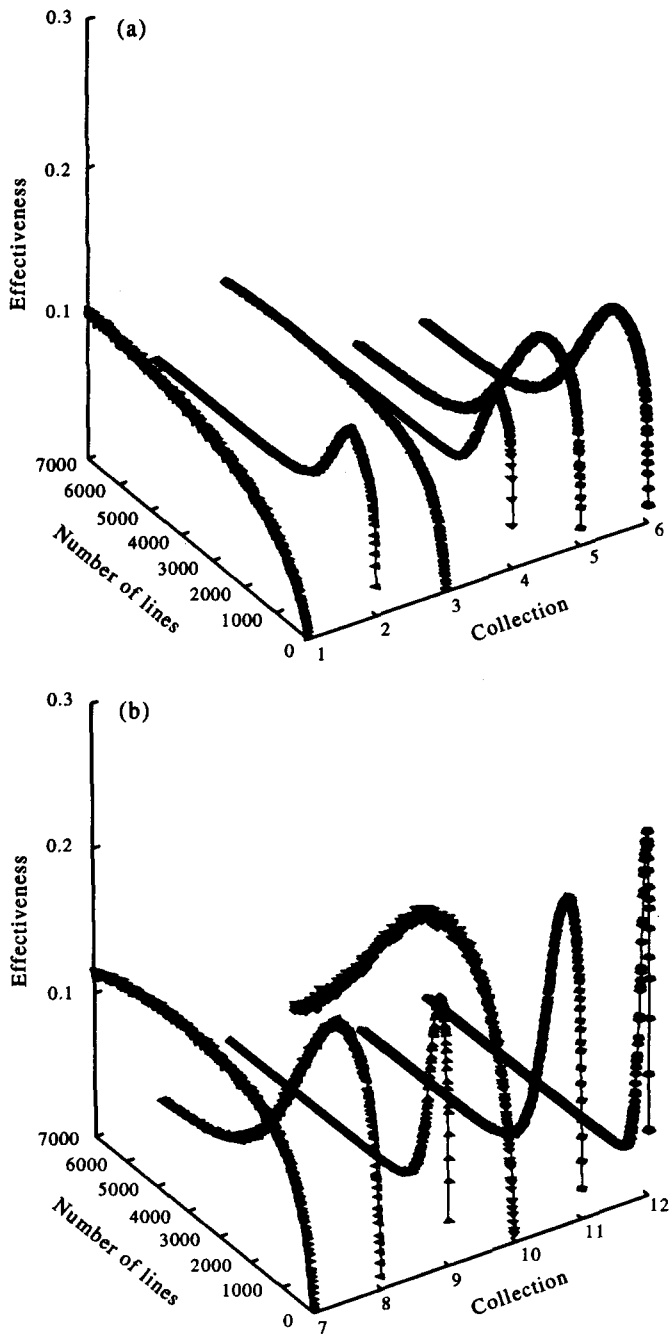

Fig. 2. (a) Expected, average effectiveness, derived from the random graph hypothesis, as a function of the number of lines (q) in random graphs and IR test collections: UKCIS(1), CISI(2), INSPEC(3), MED(4), EVANS(5), and HARDING(6). (b) Expected, average effectiveness, derived from the random graph hypothesis, as a function of the number of lines (q) in random graphs and IR test collections: NPL(7), CACM(8), KEEN(9), LISA(10), CRAN(11) and TIME(12).

Performance evaluations

Low performance standards derived from the random graph hypothesis are compared to operational levels of cluster-based rerieval effectiveness reported in the literature. Two hundred and ninety one operational results from nine papers are compared to outcomes based on chance. (Croft, 1980; El-Hamdouchi & Willett, 1989; Griffiths *et al.*, 1984; Griffiths *et al.*, 1986; MacLeod & Robertson, 1991; van Rijsbergen & Croft, 1975; van Rijsbergen, 1974; Voorhees,

Table 5. Expected random graph retrieval outcomes in IR test collections

Test collection	Expected number retrieved* ($\bar{n}_s$)	Expected number retrieved and relevant* ($\bar{r}_s$)	Expected recall ($\bar{R}_s$)	Expected precision ($\bar{P}_s$)	Expected effectiveness ($\bar{F}_s$)
UKCIS	13.1	5.6	0.0953	0.4273	0.1081
CISI	19.2	5.8	0.1415	0.3023	0.1139
INSPEC	5.0	2.3	0.0682	0.4535	0.0979
MED	4.4	1.8	0.0766	0.4046	0.1143
EVANS	4.3	1.9	0.0820	0.4369	0.1244
HARDING	4.7	2.1	0.0908	0.4370	0.1301
NPL	5.5	2.4	0.1068	0.4335	0.1296
CACM	5.4	2.3	0.1469	0.4151	0.1634
KEEN	6.0	2.4	0.1611	0.4011	0.1791
LISA	5.3	2.2	0.1994	0.4036	0.1926
CRAN	3.3	1.4	0.1682	0.4197	0.2135
TIME	4.1	1.3	0.3222	0.3052	0.2522

*The expected number of documents retrieved and relevant ($\bar{r}_s$) is estimated by the product $\bar{r}_q \times \bar{R}$, and the expected number of documents retrieved ($\bar{n}_s$) is estimated by the quotient $\bar{r}_s / \bar{P}$.

1985; Wilbur & Coffee, 1994). Reports of optimal cluster-based performance are not included in this investigation. Operational results are associated with 10 of the 13 test collections for which low performance standards have been established: CACM, CISI, CRAN, EVANS, HARDING, INSPEC, KEEN, LISA, MED, and UKCIS. A summary of operational results from each source can be obtained from the Appendix.*

Associated by topical relatedness, documents have been clustered by several hierarchical clustering techniques: complete link, group average, single link, and Ward's method. Documents have also been clustered by nearest neighbor criteria and a neural network algorithm. Retrieval strategies include top-down and bottom-up searches in hierarchies and selections from a set of nearest neighbor or neural network clusters. Clustering criteria producing many small clusters (such as complete link, group average, Ward's method, and nearest neighbor) and retrieval techniques favoring the selection of small clusters (such as bottom up searching in hierarchical

Table 6. Low performance standards for IR test collections

Test collection	Average number of relevant documents ($\bar{r}_q$)	Performance standard ($\bar{F}_s$)
UKCIS	58.9	0.11
CISI	41.0	0.11
INSPEC	33.0	0.10
CF (*RT*=1)	31.9	0.13
MED	23.2	0.11
EVANS	23.1	0.12
HARDING	22.6	0.13
NPL	22.4	0.13
CF (*RT*=2)	18.1	0.16
CACM	15.3	0.16
CF (*RT*=3)	14.9	0.18
KEEN	14.9	0.18
CF (*RT*=4)	14.1	0.19
LISA	10.8	0.19
CF (*RT*=5)	10.7	0.21
CRAN	8.2	0.21
CF (*RT*=6)	6.4	0.24
TIME	3.9	0.25

*For each source, test collection, and cluster-based retrieval strategy, the number of results and the minimum, maximum, and mean operational effectiveness are accessible in a web site (URL: http://ils.unc.edu/faculty_papers.html).

structures) are more successful than clustering criteria producing a radically skewed distribution of cluster sizes (such as single link) and retrieval techniques allowing large clusters to be selected (such as top-down searches in hierarchies) (Willett, 1988).

Cluster-based retrieval performance for each query in a database is customarily expressed in terms of E (Van Rijsbergen, 1979) and is easily converted to the effectiveness measure employed here: $F=1-E$. Mathematical properties of F and E are dictated by the harmonic mean of recall (R) and precision (P). Imposed on the natural weighting of the harmonic mean toward the value of R or P of lesser magnitude, reported values of E frequently include two weighting schemes: R is weighted twice or half as important as P. Following Van Rijsbergen (1979), eqn (2) can be expressed in terms of the weighting factor (β):

$$F=\frac{(\beta^2+1)}{\beta^2\frac{1}{R}+\frac{1}{P}}, \tag{3}$$

where $\beta=2$ assigns twice as much importance to recall than precision, $\beta=1$ imposes no additional importance to either recall or precision, and $\beta=\frac{1}{2}$ assigns half as much importance to recall than precision. If R and P have similar magnitudes, the effect of β on F is small, and, if R and P have dissimilar magnitudes and the value of R or P of lesser magnitude is weighted more heavily than the other, the effect of β is also small because the function is naturally weighted toward the smaller value. If R and P have dissimilar magnitudes and the value of R or P of greater magnitude is weighted more heavily than the other, β can have a noticeable effect on F. Interaction of the inherent weighting of the harmonic mean and β might explain the typically small differences between effectiveness values for $\beta=2$ and $\beta=\frac{1}{2}$, as noted by Keen (1992, p. 495).

The low performance standard for each test collection derived from the random graph hypothesis, the effectiveness expected by chance ($\bar{F}_x$), assumes no artificial weighting of R or P, i.e. $\beta=1$. Consequently, the average effectiveness of an operational cluster-based retrieval strategy reported in the literature and denoted here by $\bar{E}_o$ must be based on $\beta=1$ and must be converted to the harmonic average to be comparable to $\bar{F}_x$. In most cases, $\bar{E}_o$ ($\beta=1$) is reported and the conversion is straightforward: $\bar{F}_o$ ($\beta=1$)$=1-\bar{E}_o$ ($\beta=1$). In a comprehensive investigation of cluster-based retrieval effectiveness, however, El-Hamdouchi and Willett (1989) summarize results with $\bar{E}_o$ ($\beta=2$) and $\bar{E}_o$ ($\beta=\frac{1}{2}$). Converted to harmonic means, these results lead to two equations of the form of eqn (3) for which the values of $\bar{F}$ and β are known and the corresponding values of R and P are not known. The two independent equations allow computations of the unknown values of R and P. Substituting computed values of R and P into eqn (3) with $\beta=1$ yields an estimate of the needed effectiveness value for each clustering criterion, search strategy, and database investigated by El-Hamdouchi and Willett.

Differences between (average) operational and (average) expected effectiveness values ($\Delta\bar{F}$) are determined by subtracting the expected value for a database from the corresponding operational value: $\Delta\bar{F}=\bar{F}_o-\bar{F}_x$. Negative values of $\Delta\bar{F}$ identify operational results that are worse than those expected on the basis of chance, and positive values of $\Delta\bar{F}$ identify operational results that exceed expectations based on chance. The distribution of $\Delta\bar{F}$ values for operational, cluster-based retrieval results is shown in Fig. 3. Differences ($\Delta\bar{F}$) range from -0.19 to 0.45, and the mean difference is 0.00, which is influenced by the exceptional, unbalanced results for which $\Delta\bar{F}\geq 0.20$. Unfortunately, these unusually high results appear to represent the influence of a test collection, not the influence of an effective cluster-based retrieval strategy.

All retrieval outcomes for which $\Delta\bar{F}\geq 0.20$ are associated with the MED test collection. As noted by Kwok (1990), no document in the MED test collection is relevant to more than one query. If, as Kwok speculates, the MED test collection is composed of retrieval outcomes from distinct queries submitted to a large medical database, subject descriptions of documents associated with one query are not likely to be related to the representations of documents associated with other queries; consequently, a document relevant to one query is not relevant to another because the subject of that document is unrelated to the subjects of documents associated with any other query. Salton (1969) describes the construction of a small medical test

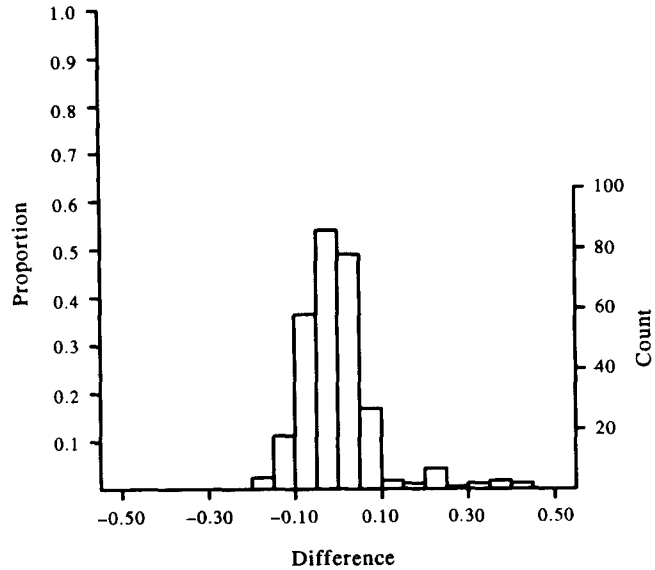

Fig. 3. Proportion (per standard unit) and count (number) of occurrences of differences between operational and expected effectiveness ($\Delta\bar{F}$) for cluster-based retrieval results. Proportion per standard unit is normalized by the sample standard deviation and allows histograms based on different scales to be compared (Freedman *et al.*, 1980, pp. 284–288).

collection following such a procedure. Queries representing distinct subject fields covered by MEDLARS and a subset of documents retrieved from MEDLARS by these queries constituted a test collection, and the resulting 18 queries and 273 documents were used to compare the retrieval performance of manual indexing and Boolean searching associated with MEDLARS with the performance of automatic indexing and a vector space search engine employed in SMART (Salton, 1969). Enhanced by additional queries and by citations to relevant documents (Salton, 1972), the small test collection appears to be a precursor to the MED test collection with 30 queries and 1033 documents; construction of the MED test collection has not been documented. Additional evidence confirms that documents relevant to a query in the MED test collection can be distinguished from other documents in the database because they appear in a small cluster of topically related documents associated with the query (Voorhees, 1985). With an average of 34.4 documents per query and an average of 23.3 relevant documents per query, selecting any or all documents topically related to a query in the MED test collection can be expected to produce a high level of retrieval performance for that query.

Excluding results from the MED test collection produces a concentrated, unimodal distribution, as shown in Fig. 4. The minimum, maximum, and mean differences between operational and expected cluster-based retrieval results ($\Delta\bar{F}$) are -0.19, 0.13, and -0.02, respectively; 91% of the differences are between $\Delta\bar{F}=-0.10$ and $\Delta\bar{F}=0.10$. The typical cluster-based retrieval outcome reported in the literature can be explained by selecting the most effective cluster for each query from a meaningless clustering outcome.

DISCUSSION

With more than 90% of the average, operational effectiveness values ($\bar{F}_o$) varying from about 0.03 to 0.29, on a scale ranging from 0 (complete failure) to 1 (complete success), it is apparent that cluster-based retrieval performance is poor. Nevertheless, it is surprising to find that such results can be explained on the basis of chance. Several explanations can be considered:

(1) the cluster hypothesis is not valid;
(2) the cluster hypothesis is valid, but clustering criteria employed to date have failed to

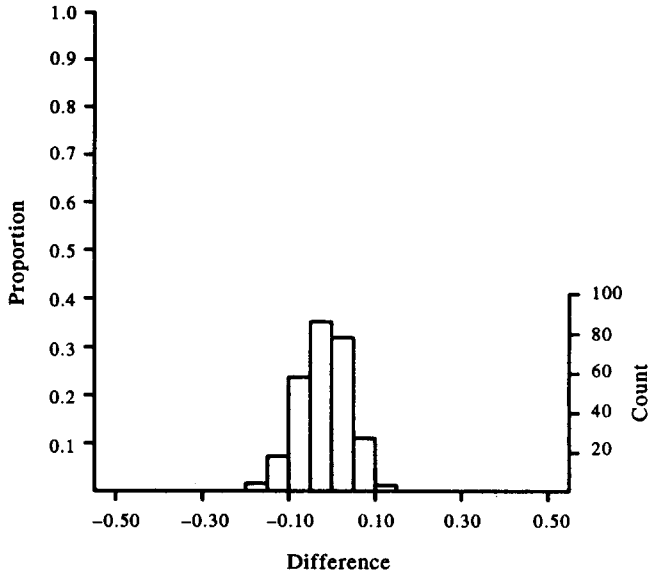

Fig. 4. Proportion (per standard unit) and count (number) of occurrences of differences between operational and expected effectiveness ($\Delta\bar{F}$) for cluster-based retrieval results, excluding results from the MED test collection. Proportion per standard unit is normalized by the sample standard deviation and allows histograms based on different scales to be compared (Freedman *et al.*, 1980, pp. 284–288).

reveal the inherent tendency for documents relevant to the same query to be grouped together;

(3) clustering criteria reveal the validity of the cluster hypothesis, but operational search strategies fail to retrieve the appropriate clusters.

It is informative to compare $\Delta\bar{F}$ values for perfect, optimal, and operational levels of performance in an internally consistent manner, and the CRAN test collection, which has been shown to have a high level of clustering tendency relative to other test collections (El-Hamdouchi & Willett, 1987), allows such a comparison.

Suppose for each query a clustering outcome includes a cluster with all relevant documents and only relevant documents. Such an outcome would provide a definitive endorsement of the cluster hypothesis and the clustering criterion, which are critical to the success of cluster-based retrieval. Suppose, further, an optimally effective search strategy unfailingly retrieves the appropriate cluster for each query. Perfect effectiveness for each query is clearly 1 ($F=1.00$), and the average effectiveness of perfect outcomes for the group of queries associated with a database is also 1 ($\bar{F}=1.00$). Among the 10 test collections investigated in this portion of the study, the difference between perfect and expected effectiveness ranges from 0.79 for the CRAN test collection to 0.96 for UKCIS. These results are off the scale of Fig. 4. Optimizing document representations for each query and employing an optimally effective search strategy, Burgin (1995, p. 567) has shown that the group average, the weighted average, or Ward's clustering criterion produces optimal effectiveness (0.62) for the CRAN test collection. The difference between optimal and expected effectiveness ($\Delta\bar{F}$) is 0.41 for CRAN. Employing four operational search strategies, the most effective of which assumes a known relevant document to initiate the search, El-Hamdouchi and Willett (1989) have shown that group average consistently produces the most effective retrieval outcomes for several test collections, including CRAN; the operational effectiveness for the CRAN test collection and the group average clustering criterion ranges from 0.23 to 0.34. Thus, if implications of the cluster hypothesis are entirely valid and flawlessly manifest by a clustering criterion, and the search strategy is optimally effective, $\Delta\bar{F}=0.79$ for the CRAN test collection; if implications of the cluster hypothesis are manifest by the most effective hierarchical clustering criterion, and the search strategy is optimally effective, $\Delta\bar{F}=0.41$; and if the most effective clustering criterion and operational search strategy are employed, $\Delta\bar{F}=0.13$. Clearly, implications of the cluster hypothesis are not fully realized in the CRAN test collection, despite its high clustering

tendency. The degradation of effectiveness from perfect to optimal levels might not invalidate the cluster hypothesis (Shaw & Willett, 1993), but it does establish a reduced target for operational implementations. Unfortunately, operational search strategies are less effective than might be hoped, reducing average effectiveness to a level precariously close to that expected by chance for CRAN and other test collections. The success of both clustering criteria and search strategies is related to an overarching assumption; topical relatedness is sufficient to discriminate relevant documents from non-relevant documents in the absence of term relevance weighting (Robertson & Sparck Jones, 1976; Shaw, 1995) or adaptive clustering techniques (Gordon, 1991). Exceptions to this assumption can distort clustering outcomes based on topical relatedness and confound searching outcomes based exclusively on subject representations of documents and information need (Barry, 1994).

CONCLUSION

Derived from random graph theory, low performance standards for cluster-based retrieval performance have been computed for 13 retrieval test collections. Differences between operational levels of cluster-based retrieval effectiveness and low performance standards are negligible; typical cluster-based retrieval results reported in the literature can be derived from meaningless clustering outcomes. Analyses reveal weaknesses in fundamental assumptions and operational implementations. The structure imposed on a set of documents by topical relatedness may not reliably associate documents relevant to the same query. Exceptions distort the posited, natural groupings of documents relevant to the same query. Clustering algorithms, which are inherently exploratory, may not reveal the natural structure and cannot reveal a structure that is not present. Search strategies exploiting the topical relatedness of queries and clusters do not always select the most effective cluster of documents. Weak realizations of expectations at all stages of cluster-based retrieval processes yield operational performance levels that are attributable to chance. Implicit in much of cluster-based retrieval research is the expectation that topical relatedness manifests relevance relationships. That this is not true in the field is widely accepted. The failure of cluster-based retrieval may be related to the weakness of this assumption in the laboratory as well. If cluster-based retrieval is to play a role in IR, it is likely to be demonstrated by adaptive clustering techniques and not by fixed clustering outcomes.

Acknowledgements—The authors are indebted to Chris Buckley, Donna Harmann, A. M. Robertson, Gerard Salton, Ellen Voorhees, Peter Willett, and Natalie Willman for access to or information about various IR test collections.

REFERENCES*

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*, 149–159.

Bullen, P. S., Mitrinović, D. S., & Vasić, P. M. (1988). *Means and their inequalities.* Dordrecht, The Netherlands: D. Reidel.

Burgin, R. (1995). The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Society for Information Science, 46*, 562–572.

*Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems, 5*, 189–195. [Crof80]

Davies, A. (1983). A document test collection for use in information retrieval. Unpublished master's thesis, University of Sheffield, U.K.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297–302.

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition, 11*, 235–254.

Dubes, R., & Jain, A. K. (1980). Clustering methodologies in exploratory data analysis. In M. C. Yovits (Ed.), *Advances in computers* (Vol. 19, pp. 113–228). New York: Academic Press.

El-Hamdouchi, A., & Willett, P. (1987). Techniques for the measurement of clustering tendency in document retrieval

*References marked with an asterisk indicate sources of operational results in the meta-analysis. The six or seven character code in square brackets accompanying these citations identifies sources in the Appendix (URL:http://ils.unc.edu/faculty_papers.html).

systems. *Journal of Information Science, 13*, 361–365.
*El-Hamdouchi, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal, 32*, 220–227. [ElHW89]
Fox, E. A. (1983). Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. Tech. Rep. 83-561, Department of Computer Science, Cornell University, Ithaca, NY.
Freedman, D., Pisani, R., & Purves, R. (1980). *Statistics*. New York: Norton.
Gordon, M. D. (1991). User-based document clustering by redescribing subject descriptions with a genetic algorithm. *Journal of the American Society for Information Science, 42*, 311–322.
*Griffiths, A., Luckhurst, H. C., & Willett, P. (1986). Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science, 37*, 3–11. [GriL86]
*Griffiths, A., Robinson, L. A., & Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation, 40*, 175–205. [GriR84]
Harary, F. (1969). *Graph theory*. Reading, Mass: Addison-Wesley.
Keen, E. M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing & Management, 28*, 491–502.
Kenney, J. F. (1947). *Mathematics of statistics*. Toronto: D. Van Nostrand.
Kwok, K. L. (1990). Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems, 8*, 363–386.
Ling, R. F. (1973). The expected number of components in random linear graphs. *The Annals of Probability, 1*, 876–881.
Ling, R. F. (1975). An exact probability distribution on the connectivity of random graphs. *Journal of Mathematical Psychology, 12*, 90–98.
Ling, R. F., & Killough, G. G. (1976). Probability tables for cluster analysis based on a theory of random graphs. *Journal of the American Statistical Association, 71*, 293–300.
*MacLeod, K. J., & Robertson, W. (1991). A neural algorithm for document clustering. *Information Processing & Management, 27*, 337–346. [MacR91]
Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 129–146.
Salton, G. (1969). A comparison between manual and automatic indexing methods. *American Documentation, 20*, 61–71.
Salton, G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science, 23*, 75–84.
Shaw, R. J., & Willett, P. (1993). On the non-random nature of nearest-neighbour document clusters. *Information Processing & Management, 29*, 449–452.
Shaw, W. M., Jr (1986). On the foundation of evaluation. *Journal of the American Society for Information Science, 37*, 346–348.
Shaw, W. M., Jr (1990a). An investigation of document structures. *Information Processing & Management, 26*, 339–348.
Shaw, W. M., Jr (1990b). Subject indexing and citation indexing—part I: Clustering structure in the cystic fibrosis document collection. *Information Processing & Management, 26*, 693–703.
Shaw, W. M., Jr (1990c). Subject indexing and citation indexing—part II: An evaluation and comparison. *Information Processing & Management, 26*, 705–718.
Shaw, W. M., Jr (1991a). Subject and citation indexing—part I: The clustering structure of composite representations in the cystic fibrosis document collection. *Journal of the American Society for Information Science, 42*, 669–675.
Shaw, W. M., Jr (1991b). Subject and citation indexing—part II: The optimal, cluster-based retreival performance of composite representations. *Journal of the American Society for Information Science, 42*, 676–684.
Shaw, W. M., Jr (1993). Controlled and uncontrolled subject descriptions in the CF database: A comparison of optimal cluster-based retrieval results. *Information Processing & Management, 29*, 751–763.
Shaw, W. M., Jr (1994). Retrieval expectations, cluster-based effectiveness, and performance standards in the CF database. *Information Processing & Management, 30*, 711–723.
Shaw, W. M., Jr (1995). Term-relevance computations and perfect retrieval performance. *Information Processing & Management, 31*, 491–498.
Shaw, W. M., Jr, Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research, 13*, 347–366.
Shaw, W. M., Jr, Burgin, R., & Howell, P. (1997). Performance standards and evaluations in IR test collections: Vector space and other retrieval models. *Information Processing & Management, 33*, 15–36.
Sparck Jones, K., & Van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of Documentation, 32*, 59–75.
*Van Rijsbergen, C. J. (1974). Further experiments with hierarchic clustering in document retrieval. *Information Processing & Management, 10*, 1–14. [VanR74]
Van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.
*Van Rijsbergen, C. J., & Croft, W. B. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 collection. *Information Processing & Management, 11*, 171–182. [VanC75]
Van Rijsbergen, C. J., & Sparck Jones, K. (1973). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation, 29*, 251–257.
*Voorhees, E. M. (1985). The cluster hypothesis revisited. *Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 188–196). New York: Association for Computing Machinery. [Voor85]
*Wilbur, W. J., & Coffee, L. (1994). The effectiveness of document neighboring in search enhancement. *Information Processing & Management, 30*, 253–266. [WilC94]
Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management, 24*, 577–597.