On the Use of Spreading Activation Methods in Automatic Information Retreival

Gerard Salton
Chris Buckley*

88-907
April 1988

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*This study was supported in part by the National Science Foundation under grant IRI 87--02735.

On the Use of Spreading Activation Methods in Automatic Information Retrieval

Gerard Salton and Chris Buckley*

Abstract

Spreading activation methods have been recommended in information retrieval to expand the search vocabulary and to complement the retrieved document sets. The spreading activation strategy is reminiscent of earlier associative indexing and retrieval systems. Some spreading activation procedures are briefly described, and evaluation output is given, reflecting the effectiveness of one of the proposed procedures.

*Department of Computer Science, Cornell University, Ithaca, NY 14853.
This study was supported in part by the National Science Foundation under grant IRI 87-02735.

1. Associative Information Retrieval

It is well-known that simple matching procedures between the vocabularies contained in the query formulations and the stored documents do not always produce acceptable retrieval output. For this reason, methods have been introduced to expand the query formulations by adding to the initial queries new terms, or expressions, that are related to the originally available terms. Analogously, the retrieved document subsets can be augmented by using items that are similar, or related, to the originally available items. [1,2]

The difficulty in applying such *associative* retrieval methods is the identification of related terms and documents that will improve the retrieval operations. If few related terms, or documents, are identified, the retrieval output is unlikely to be substantially improved. On the other hand, when the query vocabulary or the retrieved document sets are drastically altered, the advantages gained from some useful added terms, or documents, are often lost because inappropriate terms will also have been supplied.

In principle, it is possible to use generally valid term or document associations for the expansion operations. Thus term associations might be derived by using available term *thesauruses*, specifying groups of related terms, or word *dictionaries* that contain definitions and other characteristics of the stored vocabulary. Analogously, associations between documents could be defined formally by observing similarities between documents derived from library classification schedules. In practice, generally valid vocabulary or document relations are difficult to find, so that it is necessary to rely on locally valid relations that are applicable only for particular document collections in particular situations. The well-known *relevance feedback* process, which augments the initial query vocabulary by adding terms contained in previously retrieved documents known to be relevant to the query, is an example of such a local term augmentation process. [3-5]

Various methods have been suggested for computing locally valid term and document associations. One such process, known as *associative linear retrieval*, consists in taking the original assignment of terms to documents specified in matrix C of Figure 1(a), and using it to compute a matrix T of pairwise term association factors, as well as a matrix D of pairwise document association factors. Assuming that C is of dimension $d \times t$, the $d \times d$ matrix D specifying the associations between pairs of documents is then obtained by using similarities in the term assignments to pairs of documents. For example, the similarity between documents D_i and D_j, stored as the ijth element of matrix D, is computed by comparing rows i and j of matrix C (see Fig. 1(b)). Similarly the $t \times t$ matrix T giving the term relations is obtained by noting similar postings patterns for pairs of terms (see Fig. 1(c)). In a linear retrieval system, the response vector r giving the retrieval values for all d documents of the collection can be computed by using the linear transformations shown in Fig. 1(d). [6-7]

While attempts have been made to produce workable term and document expansion systems, it is fair to say that completely satisfactory models have never been designed. For example, a query expansion method using term associations derived from a maximum spanning tree of term similarities led to the following conclusions: [8]

> "Our results on query expansion using the NPL data are disappointing. We have not been able to achieve any significant improvements over non-expansion... This leaves the whole question of the effectiveness of query expansion methods unresolved."

Other recent attempts to supply expanded document representations using citations and other bibliographic indicators attached to texts and documents have also led to the conclusion that effective term expansion methods valid for a variety of different collections are difficult to generate. [9]

Various reasons may be given for the lack of consistent improvements in the associative retrieval operations: First, the associations derived from particular pairs of documents, or pairs of terms, may be valid only locally in the particular environment from which the associations are derived. [10] Second, most practical methods for computing the linear document associations are based on the assumption that the terms are originally uncorrelated, that is, independent of each other; furthermore, the term association computations require an analogous assumption about the independence of the documents. [11] Finally, methods based on simplified theoretical models may not reflect the reality of existing relationships between documents and terms in operational situations. [12,13]

Recently, the term and document expansion methods have been revised using so-called spreading activation techniques based on mechanisms of human memory operations. [14-16] These developments are briefly covered in the remainder of this note.

2. Spreading Activation Techniques

The spreading activation techniques used in information retrieval are based on the existence of maps specifying the existence of particular relations between terms or concepts, or between documents, as the case may be. A typical arrangement of concepts of interest in cardiology is shown in Fig. 2. An alternative map including both terms as well as documents appears in Fig. 3. In both cases, the terms or documents are represented by network nodes, and the relationships between terms or documents are specified by labelled links between the nodes. In the examples of Figs. 2 and 3, the documents may be related through bibliographic citation (when one document cites another one), or by being close neighbors of each other (because the term sets assigned to the respective items are sufficiently similar). Analogously, two terms may be synonyms of each other, or one of the terms may be specified as a part, or a component of some other term.

The node activation process used in the spreading model starts by placing a specified *activation weight* at some starting term or document node. Typically the starting node might represent a term included in an initial query formulation, or a document retrieved in an earlier search operation. The initial activation weight then spreads through the network along the links originating at the starting node. The spreading action first affects those nodes located closest to the starting node, and spreads through the network one link at-a-time. Normally, the activation weight of a node is computed as a function of the weighted sum of the inputs to that node from directly connected nodes. If a_j is the original activation weight of node j, and w_{ij} is the link weight between nodes i and j, representing the influence of node j on node i, the new activation weight a_i' of node i may then be computed as

$$W_i = \sum_{\substack{\text{connected}\\ \text{nodes}}} w_{ij} \cdot a_j \tag{1}$$

$$a_i' = f(W_i)$$

where the summation extends over all nodes j directly connected to a given node.

Most spreading activation systems differ from the earlier associative indexing and retrieval system in several respects: [15,16]

a) Each node is assigned a special activation weight, which depends on the starting activation weight, and the link and node types traversed in the activation process.

b) Distance constraints may be imposed in the activation process by stopping the activity at some specified distance from the original node.

c) Nodes with a large branching ratio (fan-out) that are connected to many other nodes (therefore possibly representing ambiguous, or general notions or documents) may be bypassed in the activation process, or may otherwise receive special treatment.

d) The activation process follows specific rules. Thus for topic activations, the activation rules might be of the form

[(topic (x) with weight w(x)) *and* (link type L(x,y))]
implies (topic (y) with weight w(y))

Obviously, the node activation process used in the spreading system is much more refined than the earlier node association procedures where all nodes and all links receive equal treatment. This implies that, potentially at least, much better retrieval results can be obtained by the spreading activation process than by the ordinary node association system. The effectiveness of the process is, however, crucially dependent on the availability of a representative node association map, and on the use of activation rules that can distinguish the useful from the extraneous nodes. It remains to be seen whether representative term and document maps can be designed for the subject areas covered by ordinary document collections, and whether the refined spreading rules made possible by the system can actually be translated into workable node activation methods.

3. Evaluation of a Particular Spreading Activation Method

A complete spreading activation system that makes use of diverse link types, and of spreading rules with distance and fan-out constraints, has not so far been implemented with ordinary document collections. However, various simplified systems are discussed in the literature. One typical approach consists in starting with an initial query Q, using an activation weight w_{q_i} equal to the weight of each term i in the query as shown at the top of Fig. 4. Each query term node then activates the nodes of the corresponding terms in all documents in which that term occurs. The activation weight of the document term nodes is a combination of the weights w_{q_i} and w_{d_i} of terms i in queries and documents respectively. Finally, each document node is activated using some function of the term weights of all terms present in that document. Using such a computation of document weights, the items could be retrieved in decreasing order of the weights of all documents. A simplified representation of a document weighting operation using the spreading activation model is shown in Fig. 4. In that case, distance and fan-out restrictions are not used, and there are no explicit relations between terms or documents.

The following typical node weighting system has been proposed for a spreading activation model of document retrieval. [17] Assuming that the query node first receives an initial activation weight A, the activation weight a_t of a query term t is defined as

$$a_t = A \frac{w_{q_t}}{\sum_{i=1}^{m} w_{q_i}} \qquad (2)$$

where w_{q_t} is the initial weight of term t in the query, and the denominator represents a normalization factor equal to the sum of the weights of all query terms. A document node D reached from term t is next activated with a weight a_d as follows

$$a_d = a_t \frac{w_{dt}}{\sum_{j=1}^{N} w_{jt}} \tag{3}$$

where w_{dt} once again represents the original weight of term t in the document, and the summation in the denominator covers the tth term in all N documents of the collection. In a binary indexing system, where terms present in a given document (or query) receive a weight of 1, and term absent from the document (or query) are assigned a weight of 0, the denominator in (2) is equal to the number of query terms, whereas the denominator in (3) represents a document frequency factor equivalent to the postings frequency of term t, that is, the number of documents to which term t is assigned.

The total activation for a given document D received from all query terms assigned to that document can be summed to obtain the corresponding query-document similarity: [17]

$$similarity(D,Q) = A \sum_{i=1}^{m} \frac{w_{qi} \cdot w_{di}}{\left[\sum_{k=1}^{m} w_{qk}\right]\left[\sum_{j=1}^{n} w_{ji}\right]} \tag{4}$$

The retrieval effectiveness of a system based on the query-document similarity measure of expression (4) can be compared with that produced by other well-known document and query term weighting systems. Effective term weighting systems are known to be based on three different components: [18-20}

a) A *term frequency* factor (tf) that measures the frequency of occurrence of a term in a given document or query. Alteratively, an enhanced term frequency factor normalized to lie between 0.5 and 1, and computed as (0.5 + 0.5 tf/maxtf), can be used where max tf is the highest value of the term frequency for any term in a particular document or query.

b) An *inverse document frequency* (idf) factor, typically computed as $\log N/n_i$, which increases as the postings frequency n_i of term i decreases for a collection of N documents.

(c) And a length normalization factor which reduces all document and query vectors to the same length. For a given document vector $(w_{d_1}, w_{d_2}, \ldots, w_{d_m})$ the length normalization might be $1/\sqrt{\sum_{i=1}^{m} (w_{d_i})^2}$).

It is known that an effective term weighting system for document terms uses a composite $tf \times idf$ (term frequency multiplied by inverse document frequency) factor that is normalized for document length so that all documents are treated equally, regardless of the number and the weight of the assigned terms. For query terms, the length normalization becomes unnecessary because queries are normally short and more homogeneous in length than documents. However, the query terms may be given extra importance by insisting that the query term frequency factors lie in the range from 0.5 to 1. Useful term weighting factors for documents and queries may then be computed in the following way [20]:

a) For document terms

$$W_d = \frac{tf \cdot \log (N/n)}{\sqrt{\sum_{vector} [tf_i \cdot \log \frac{N}{n_i}]^2}} \tag{5}$$

b) For query terms

$$W_q = \left(0.5 + \frac{0.5\ tf}{\max tf} \right) \log \frac{N}{n} \tag{6}$$

Using the term weight of expressions (5) and (6), a document retrieval value may be obtained for each document by computing the standard vector product between query and document vectors as follows:

$$\text{similarity}\,(D,Q) = \sum_{i=1}^{m} w_{d_i} \cdot w_{q_i} \ . \tag{7}$$

When the vectors jointly contain a number of highly weighted terms, the similarity value of expression (7) will be large, and the corresponding documents may then be retrieved early in a given search operation.

By interpreting the query and document activation factors used in the spreading activation system as ordinary term weights, a comparison becomes possible between the spreading activation method of expression (4) and other previously used vector processing methods. In the spreading activation system, the inverse document frequency factor is absent for the query terms, and the document terms are not normalized for vector length. Furthermore, the inverse document frequency factors used for document terms differ substantially between the spreading activation formula (4) and the $(tf \times idf)$ system of expressions (5) and (6).

A number of query and document term weighting coefficients are included in Table 1. In each case the query document similarity is obtainable by multiplying the weights of equivalent query and document terms, and summing over all terms. The first four weighting systems of Table 1 are derived from the spreading activation method of expression (4); the other four systems represent more general vector processing strategies.

The first method of Table 1 is the exact spreading activation system of expression (4). Method (2) adds a normalization factor for document length. Method (3) represents the basic spreading activation with more conventional (L2) normalization factors in the denominator. Method (4) is equivalent to (2), but uses an L2 normalization. The four vector processing methods (5) to (8) represent a standard term frequency weight for both queries and documents (5), a standard term frequency multiplied by inverse document frequency $(tf \times idf)$ (6), a length normalized term frequency for documents and enhanced $(tf \times idf)$ weights for queries (7), and finally a normalized $tf \times idf$ weight for documents and enhanced $(tf \times idf)$ weight for queries (8). These last two weighting strategies are known to produce a high order of retrieval effectiveness. [20]

The weighting systems of Table 1 are evaluated using six document collections in different subject areas, ranging in size from 1033 documents and 30 queries in biomedicine (MED) to 12,684 documents and 84 queries in electrical engineering (INSPEC). The collection characteristics are listed in Table 2. The evaluation output is shown in Table 3 for the eight term weighting systems using the six sample document collections. The performance is measured by giving average search precision values evaluated at three fixed recall points of 0.25, 0.50, and 0.75. The average precision values are further averaged over the number of queries available for each collection. The base run is the spreading activation system of expression (4) described earlier in [17]. Percentage improvement or deterioration figures are shown in Table 3 reflecting the differences in performance between each particular method and the corresponding base run.

Table 3 shows that the simple spreading activation system does not produce a high performance level. Improvement in retrieval effectiveness are easily obtained by using better normalization factors, and in particular by taking into account the document length. Furthermore, with the exception of the term frequency weights of method (5), the vector processing model produces better output than the spreading activation system. The weighting system of method (8) improves the output by factors ranging from 23 percent for the NPL collection to over 40 percent for CISI and CRAN above the performance level of the spreding activation system.

The simple spreading activation system considered in this study may not be sufficiently powerful to produce acceptable retrieval output. More refined systems may become available in the future; such systems must be evaluated before actual implementations in retrieval environments appear warranted.

References

[1] L.B. Doyle, *Information Retrieval and Processing*, Melville Publishing Company, Los Angeles, CA, 1975.

[2] G. Salton, *Automatic Information Organization and Retrieval*, McGraw Hill Book Company, NY, 1968.

[3] J.J. Rocchio, Jr., Relevance Feedback in Information Retrieval, in *The SMART Retrieval System--Experiments in Automatic Document Processing*, ed. G. Salton, Englewood Cliffs, NJ, Prentice-Hall, Inc., 1971, Chapter 14, 313-323.

[4] E. Ide, New Experiments in Relevance Feedback, in *The SMART Retrieval System--Experiments in Automatic Document Processing*, ed. G. Salton, Englewood Cliffs, NJ, Prentice-Hall, Inc., 1971, 337-354.

[5] E. Ide and G. Salton, Interactive Search Strategies and Dynamic File Organization in Information Retrieval, in *The SMART Retrieval System--Experiments in Automatic Document Processing*, ed. G. Salton, Englewood Cliffs, NJ, Prentice-Hall, Inc., 1971, Chapter 18, 373-393.

[6] V.E. Giuliano and P.E. Jones, Linear Associative Information Retrieval, in P. Howerton (ed.) *Vistas in Information Handling*, Spartan Books, Inc., Washington, DC 1983.

[7] P.E. Jones, V.E. Giuliano and A.M. Curtice, *Papers on Automated Language Processing*, Vol. 2, Linear Models for Associative Retrieval, A.D. Little Inc., Report ESD TR 67-202, February 1967.

[8] S.E. Robertson, C.J. van Rijsbergen and M.F. Porter, Probabilistic Models of Indexing and Searching, in *Information Retrieval Research*, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, P.W. Williams, editors, Butterworths, London, 1981, 35-56.

[9] E.A. Fox, Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types, Doctoral Dissertation, Department of Computer Science, Cornell University, Ithaca, NY, 1983.

[10] M.E. Lesk, Word-word Associations in Document Retrieval Systems, *American Documentation*, 20:1, January 1989, 27-38.

[11] V.V. Raghavan and S.K.M. Wong, A Critical Analysis of the Vector Space Model for Information Retrieval, *Journal of the American Society for Information Science*, 37:5, September 1986, 279-287.

[12] C.J. van Rijsbergen, A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval, *Journal of Documentation*, 33, 1977, 106-119.

[13] C.T. Yu, C. Buckley, K. Lam and G. Salton, A Generalized Term Dependence Model in Information Retrieval, *Information Technology: Research and Development*, 2:4, October 1983, 129-154.

[14] A.M. Collins and E.F. Loftus, A Spreading Activation Theory of Semantic Processing, *Psychological Review*, 82, 1975, 407-428.

[15] J.R. Anderson, The Architecture of Cognition, Harvard University Press, 1983, Cambridge, MA, Chapter 3.

[16] P.R. Cohen and R. Kjeldsen, Information Retrieval by Constrained Spreading Activation in Semantic Networks, *Information Processing and Management*, 23:4, 1987, 255-268.

[17] W.P. Jones and G.W. Furnas, Pictures of Relevance--A Geometric Analysis of Similarity Measures, *Journal of the American Society for Information Science*, 38:6, 1987, 420-442.

[18] G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, 29:4, December 1973, 351-372.

[19] G. Salton, C.S. Yang and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, *Journal of the American Society for Information Science*, 26:1, January-February 1975, 33-44.

[20] G. Salton and C. Buckley, Term Weighting Approaches vs Automatic Text Retrieval, Technical Report 87-881, Department of Computer Science, Cornell University, Ithaca, NY, November 1987.

		Term Weight in Documents	Term Weight in Queries
(1)	Spreading Activation L_1 term normalization (document normalization extends over jth term in all N documents)	$\dfrac{tf_{d_i}}{\sum_{j=1}^{N} tf_{ji}}$	$\dfrac{tf_{q_i}}{\sum_{k=1}^{m} tf_{qk}}$
(2)	Spreading Activation L_1 term normalization and document length normalization	$\dfrac{tf_{d_i}}{\sqrt{\sum_{k=1}^{m} (tf_{d_i})^2} \cdot \sum_{j=1}^{N} tf_{j_i}}$	$\dfrac{tf_{q_i}}{\sum_{k=1}^{m} tf_{q_k}}$
(3)	Spreading Activation L_2 term normalization	$\dfrac{tf_{d_i}}{\sqrt{\sum_{j=1}^{N} (tf_{j_i})^2}}$	$\dfrac{tf_{q_i}}{\sqrt{\sum_{k=1}^{m} (tf_{q_k})^2}}$
(4)	Spreading Activation L_2 term normalization and document length normalization	$\dfrac{tf_{d_i}}{\sqrt{\sum_{k=1}^{m} (tf_{l_i})^2} \sqrt{\sum_{j=1}^{N} (tf_{j_i})^2}}$	$\dfrac{tf_{q_i}}{\sqrt{\sum_{k=1}^{m} (tf_{q_k})^2}}$
(5)	Vector Process standard tf weights for documents and queries	tf_{d_i}	tf_{q_i}
(6)	Vector Process standard ($tf \times idf$) weights for documents and queries	$\left[tf_{d_i} \cdot \log \dfrac{N}{n_i}\right]$	$\left[tf_{q_i} \cdot \log \dfrac{N}{n_i}\right]$
(7)	Vector Process tf documents length normalized; enhanced $tf \times idf$ queries	$\dfrac{tf_{d_i}}{\sqrt{\sum_{k=1}^{m} (tf_{d_i})^2}}$	$\left[0.5 + \dfrac{0.5\, tf_{q_i}}{\max tf_q}\right] \left[\log \dfrac{N}{n_i}\right]$
(8)	Vector Process (best) $tf \times idf$ documents length normalized; enhanced $tf \times idf$ queries	$\dfrac{tf_{d_i} \cdot \log \dfrac{N}{n_i}}{\sqrt{\sum_{k=1}^{m} (tf_{d_i} \cdot \log \dfrac{N}{n_i})^2}}$	$\left[0.5 + 0.5\, \dfrac{tf_{q_i}}{\max tf_q}\right] \left[\log \dfrac{N}{n_i}\right]$

Weighting Formulas for Document and Query Terms
(N documents in collection, m distinct terms)

Collection	Number of Documents	Average Number of Document Terms	Number of Queries	Average Number of Query Terms
CACM	3204	24.5	64	10.1
CISI	1460	46.6	112	28.3
CRAN	1398	53.1	225	9.2
INSPEC	12684	32.5	84	15.6
MED	1033	51.6	30	10.1
NPL	11429	20.0	100	7.2

Collection Specifications

Table 2

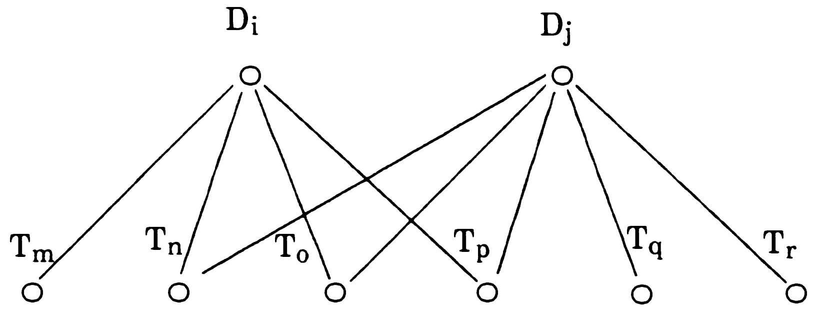
	Processing Method	CACM 3204 docs 64 queries	CISI 1460 docs 112 queries	CRAN 1397 docs 225 queries	INSPEC 12684 docs 84 queries	MED 1033 docs 30 queries	NPL 11429 docs 100 queries
(1)	Standard Spreading Activation (base runs)	.2830	.1844	.2724	.2110	.4539	.1577
(2)	Spreading Activation with Document Length Normalization	.2636 − 7%	.1991 + 8%	.3115 +14%	.2174 + 3%	.4732 +21%	.1486 − 6%
(3)	Spreading Activation (L_2 norm)	.3167 +12%	.1745 − 5%	.2744 + 1%	.2142 + 2%	.4279 − 6%	.1697 + 8%
(4)	Spreading Activation (L_2 norm) with Length Normalization	.3299 +17%	.2081 +13%	.3540 +30%	.2434 +15%	.4779 + 5%	.1848 +17%
(5)	Vector Process standard *tf* weights	.1647 −42%	.1322 −28%	.1834 −37%	.1060 −50%	.3899 −14%	.1197 −24%
(6)	Vector Process standard $tf \times idf$ weights	.3248 +15%	.2166 +40%	.2991 +10%	.2365 +12%	.5177 +14%	.1846 +17%
(7)	Vector Process *tf* length normed documents; enhanced $tf \times idf$ queries	.3252 +15%	.2189 +42%	.3950 +45%	.2626 +24%	.5542 +22%	.2170 +38%
(8)	Vector Process $tf \times idf$ length normed documents; enhanced $tf \times idf$ queries	.3630 +28%	.2189 +42%	.3841 +41%	.2626 +24%	.5628 +24%	.1933 +23%

Performance Evaluation for 6 Document Collection
(3-point precision averages)

Table 3

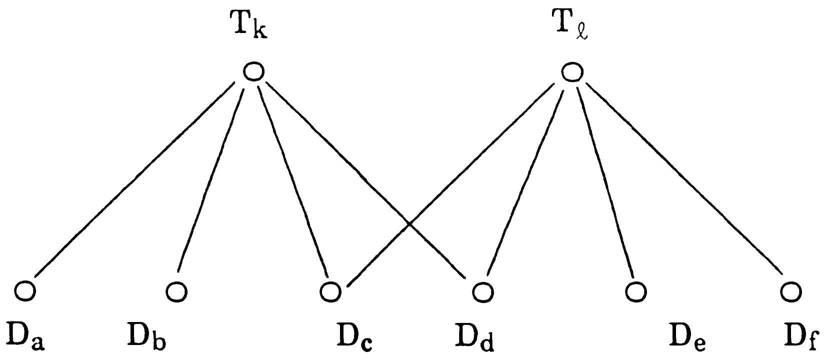
$$
\underline{C} = \begin{array}{c|cccc}
 & T_1 & T_2 & \cdots & T_t \\
\hline
D_1 & c_{11} & c_{12} & \cdots & c_{it} \\
D_2 & c_{21} & c_{22} & \cdots & c_{2t} \\
\cdot & \cdot & & & \\
\cdot & \cdot & & & \\
\cdot & \cdot & & & \\
D_d & c_{d1} & c_{d2} & \cdots & c_{dt}
\end{array}
$$

a) Original matrix C showing term assignment to documents



b) Computation of document association matrix D using overlapping vocabulary between document pairs.

Fig. 1. Linear Associative Retrieval

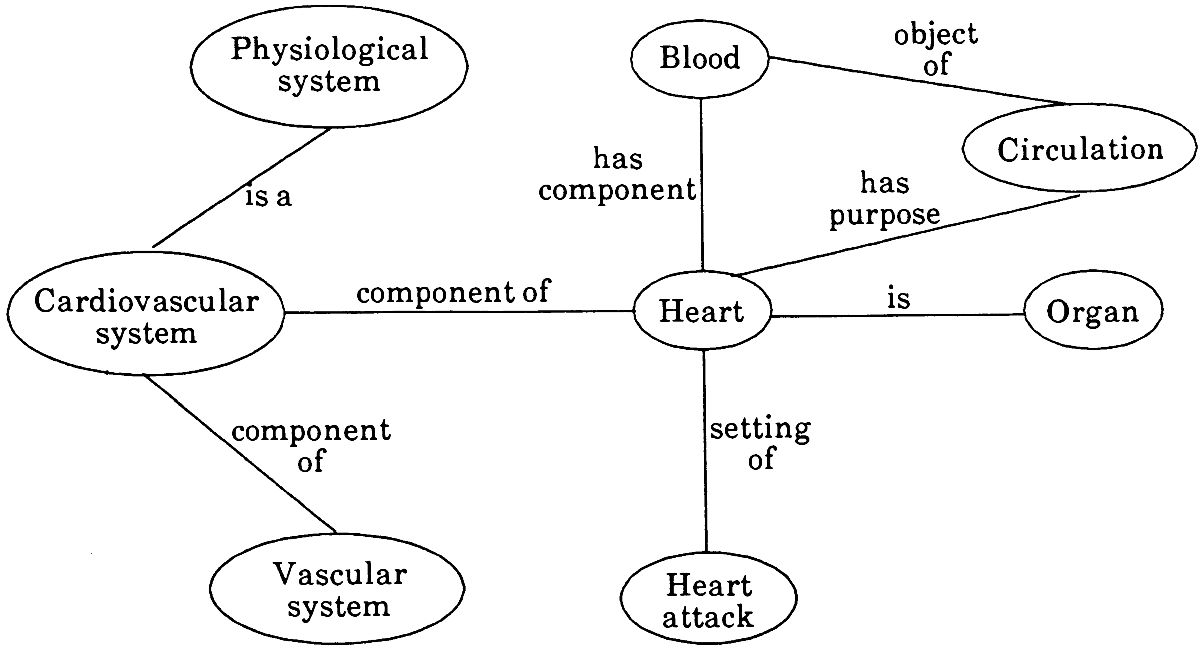


c) Computation of term association matrix T using overlapping postings to documents between pairs of terms

Initial retrieval operation	$\underset{(d)}{\mathbf{r}} = \underset{(d \times t)}{C} \cdot \underset{(t)}{\mathbf{q}}$
Retrieval with augmented terms	$\underset{(d)}{\mathbf{r}} = \underset{(d \times t)}{C} \cdot \underset{(t \times t)}{T} \cdot \underset{(t)}{\mathbf{q}}$
Retrieval with augmented documents	$\underset{(d)}{\mathbf{r}} = \underset{(d \times d)}{D} \cdot \underset{(d \times t)}{C} \cdot \underset{(t)}{\mathbf{q}}$
Retrieval with augmented terms and documents	$\underset{(d)}{\mathbf{r}} = \underset{(d \times d)}{D} \cdot \underset{(d \times t)}{C} \cdot \underset{(t \times t)}{T} \cdot \underset{(t)}{\mathbf{q}}$

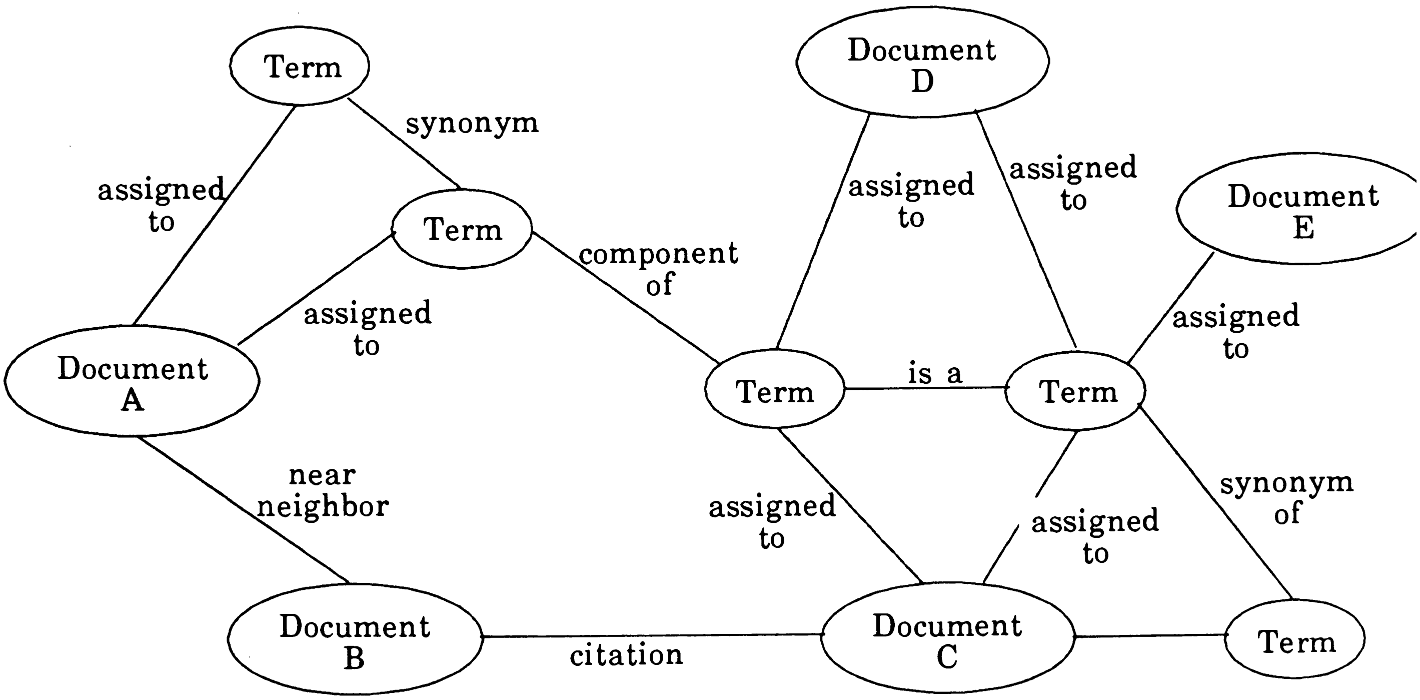
d) Computation of document retrieval values

Fig. 1 Linear Associative Retrieval (cont.)



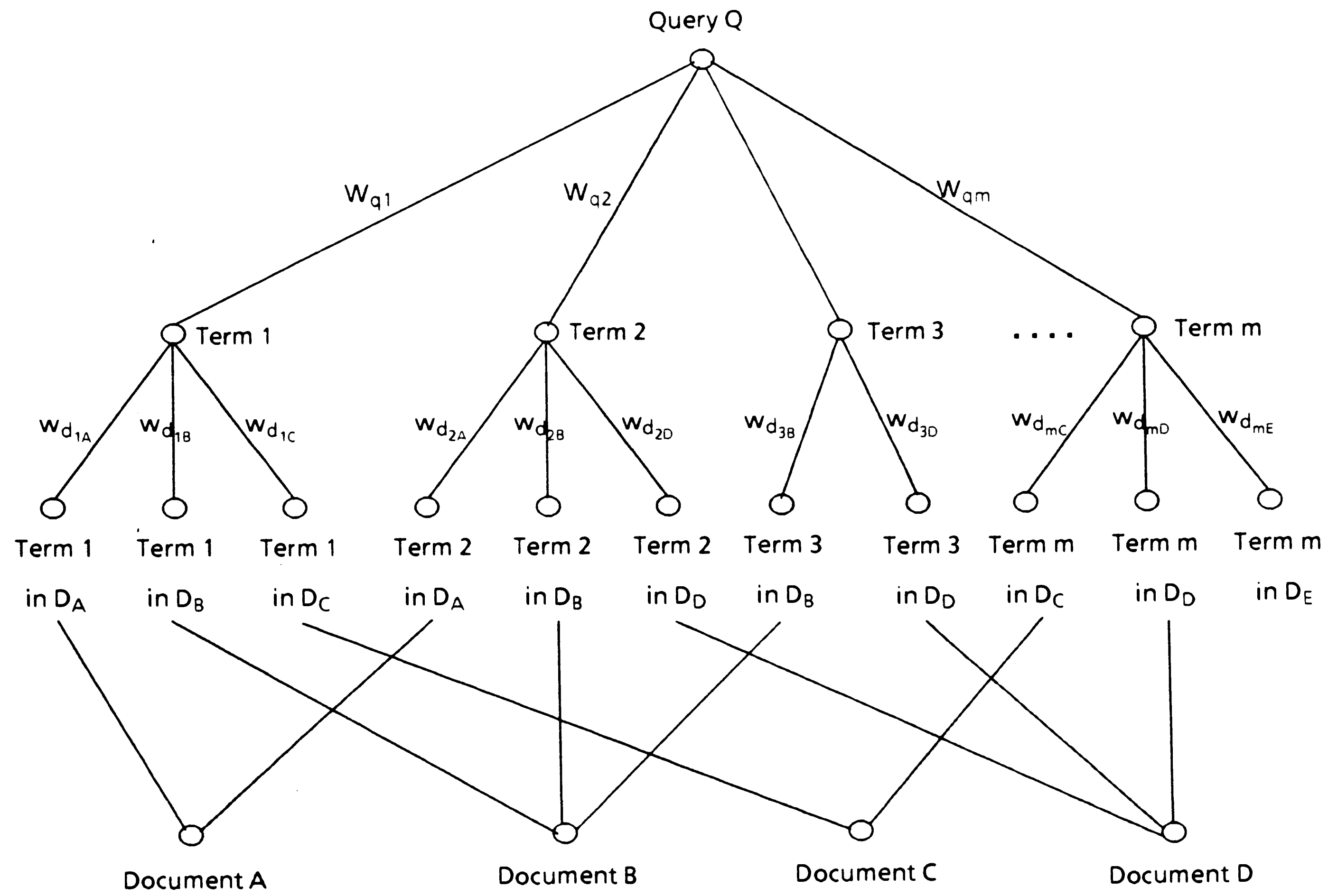
Typical Concept Arrangement (excerpted from [15])

Fig. 2



Typical Term-Document Map (adapted from [16])

Fig. 3



Simplified Spreading Activation from Query
to Corresponding Documents

Fig. 4