This work was supported in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract No. DA 28 043 AMC 00073(E); and in part by the National Science Foundation: Grant NSF GK-690.

Reproduction in whole or in part is permitted for any purpose of the United States Government.

Distribution of this report is unlimited. Qualified requesters may obtain copies of this report from DDC.
Abstract

This paper considers the problem of information retrieval from the point of view of graph theory. In this formulation documents are represented as nodes and relationships among the documents are represented by edges. Two types of graphs are introduced, namely the similarity graph which is based on subject-content correlation and the citation graph, which is derived from direct citation linkages among documents. Several distance measures are considered and evaluated with regard to retrieval operations.

I. Introduction

Within the scope of this paper we shall consider an information retrieval system to consist of two major components, namely, a document collection and a retrieval procedure, that is, a systematic way of selecting a subset of documents of the collection according to a given criterion.

The documents in the collection are coupled to one another in many different respects, such as subject content, form, authorship, citations, etc. Two of these facets, namely subject content and citations, have been exploited for application in retrieval.

In a great many modern information retrieval systems the characteristics in subject content are expressed in terms of subject descriptors. Attached to each document is a set of subject descriptors which characterizes the subject content of the document. A measure of the similarity between a pair of documents can then be obtained by comparing their assigned descriptors. Characterizations of documents through the use of subject descriptors is known as coordinate indexing.
In retrieval operation a query is presented to the system which describes a profile of the type of documents to be retrieved from the collection. In most systems employing coordinate indexing today the query is given in terms of a set of descriptors or some logical function thereof. For instance, we may ask for all documents that deal with the "decoding" of "Bose-Chandhuri-Hocquenghem Codes" that are published in the "Transactions of IEEE on Information Theory" since "1964," where those terms under quotation signs are descriptors.

Another type of retrieval systems are based on citation indexing. In this type of systems citation information among documents is stored in the system. The query is given in terms of specifying accession documents in the network. For instance, one might wish to retrieve all documents citing a document d or one might wish to retrieve all documents that are cited by document d. Retrieval operations based on multi-generation citations are theoretically feasible but so far have not received much attention.

In comparing the two popular schemes, citation indexing is easy to instrument but is limited in scope in that it derives information only from existing direct linkages in the document collection. This restriction is reflected in the usual incompleteness of retrieval results when one is interested in searches based on subject content.

On the other hand, coordinate indexing works well only if the indexed document collection is relatively homogeneous and the query well-defined. For requests from research scientists the query is always aimed at the intersection or the union of several narrow and ill-defined disciplines. As a result, the outcome is usually contaminated with large amounts of irrelevant material.
Aimed at retrieval procedures that will produce sharper and more complete responses, we propose the study of potential systems that combine the resources of both the coordinate-indexing approach and the citation methods. To minimize the inconsistency between indexing and retrieval we choose to represent all queries in terms of documents. To state it formally, the problem treated in this paper is one of finding an information retrieval system that combines the advantages of both the coordinate indexing and citation indexing. A typical retrieval operation would be the retrieval of a set of documents that is "close" in some reasonable measure to a given document profile. To facilitate instrumentation emphasis is placed on easily-implemented systems.

II. The Correlation Graph

The main consideration in this section will be document couplings that are subject-content based. Although a number of studies have been made in this area involving fairly complicated couplings and their interactions, the type of couplings to be investigated here will be relatively simple in nature as our chief objective dwells on the question of optimum combination of subject-content based indexing and non-subject-content based indexing.

Let us consider a coordinate indexing scheme in which each document is assigned a number of descriptors. For a typical system the total number of descriptors will be of the order of 10,000 while each document may be assigned ten to fifteen descriptors on the average. A typical curve for descriptor frequency is given in Figure 1. The behavior of the curve sketched in Figure 1 can be explained as follows. It is observed that typically there are two kinds of descriptors. Descriptors of the first
kind may be termed general descriptors and have a high probability of being used for many documents. Descriptors of the second kind are specialized in nature and have a low probability of being used but provide the system with a tremendous amount of selectivity whenever they are present.

The dichotomy of the descriptor population points up the difficulty in indexing resolution. In the interest of efficiency it is necessary to keep the number of descriptors, especially descriptors of the general type, small. The thesaurus of any practical system is therefore usually the result of compromises. While the initial resolution may be adequate for the initial collection and subject to most queries, the system may not perform satisfactorily when the document collection grows or when the system cannot be defined clearly with the system's limited vocabulary.

Let us consider the document descriptor matrix $A$ which has $m$ rows and $n$ columns. With each row $A$ is associated a document and each column a descriptor. The entry $a_{ij}$ takes the value one if the $j$th descriptor is assigned to the $i$th document and zero otherwise. We define the $m \times m$ correlation matrix as

$$C = AA^T = \|c_{ij}\|.$$
The correlation graph is defined by the following process. We assign each document a node and assign the value $c_{ij}$ as the weight of the link between nodes $i$ and $j$. Thus the weight $c_{ij}$ of the link in the correlation graph serves as a measure of "closeness" between documents $i$ and $j$.

It is noted that the number of rows of $C$ is equal to the number of documents, $m$, in the document collection. This is usually a large number. To compute $AA^T$ in the conventional way of matrix computation would not be an attractive approach. Since the number of descriptors assigned for each individual document is small the density of entries $a_{ij}$ in $A$ is very low. The computation of $C = AA^T$ can then be done efficiently by list processing techniques. A detailed discussed of the technique will be given in conjunction with the analysis of the citation graph in the next section.

III. The Citation Graph

Another class of structural organizations of a given collection of documents can be obtained by exploiting the bibliographic couplings. Several types of bibliographic couplings may be envisaged, such as those based on the number of shared references, citation, weighted citation, etc. Obviously, the simplest type of coupling is provided by direct citation, which may be considered as a first order association of documents. In this scheme, with each document we associate a set of documents, i.e. the documents it cites. Citation is interpreted as a directed relation between citing and cited: if we represent documents with nodes, citation can be adequately represented by directed edges from the citing document to the cited documents. We perform this representation for each document in the collection and the citation graph is constructed.
Formally, given a document collection $B = \{d_1, d_2, \ldots, d_n\}$ consisting of documents $d_1, d_2, \ldots, d_n$, the directed citation graph $G$ pertaining to $B$ is entirely described by an $n \times n$ matrix $E = \{e_{ij}\}$, where $e_{ij} > 0$ if and only if document $d_i$ cites document $d_j$.

As noted, citation indicates an association between documents and could be conveniently exploited in retrieval operations. Specifically, the citation structure may be particularly useful when the query is formulated by specifying a non-empty set of documents $Q$ and the retrieval goal is the extraction of a set $R$ of documents ($R \supseteq Q$) which are subject-related to the documents of $Q$. In the simplest instance, $Q = d_i$, i.e. it contains a single document $d_i$. $d_i$ is denoted as the access point.

The determination of the retrieved set could be conveniently performed in a mechanical fashion through the evaluation of some single-valued distance function defined between each pair of nodes of the graph.

Before analyzing the prerequisites of a distance function, we reconsider the directed citation graph $G$. If we take citation as a sign of subject-relation, we see that for the purpose of defining subject-areas the direction of citation loses its importance. This leads us to replacing the directed graph $G$ with the undirected graph $U$, simply denoted as the citation graph. $U$ is described by the $n \times n$ matrix

$$T = \|t_{ij}\| = C + C^T$$

where now $t_{ij} = t_{ji} > 0$ means that $d_i$ and $d_j$ are linked through direct citation. The weight of the linkage, $t_{ij}^*$, may be binary-valued ($0, 1$) if we are simply interested in the presence or absence of citation. In more
refined schemes it could be real-valued non-negative, its magnitude measuring the strength of coupling in a normalized interval (0,1).

We now make an attempt to formulate some properties which seem to be desirable for a distance function \( f_{ij} \) defined for every pair of nodes \( d_i, d_j \) of the graph \( U \): obviously \( f_{ij} \) must provide an intuitively satisfactory measure of connectivity.

First, suppose that a procedure has been given for the computation of \( f_{ij} \). It seems reasonable to require that, if the coupling strength \( t_{hk} \) between two generic documents \( d_h \) and \( d_k \) is increased (i.e. \( t_{hk} \) is a continuous parameter), the distance between any two distinct documents \( d_i, d_j \) cannot increase. Formally, in the hypothesis that coupling strengths are continuous parameters

\[
\frac{\partial f_{ij}}{\partial t_{hk}}
\]

must be continuous and for \( t_{hk} > 0, f_{ij} \geq 0 \) we must have

\[
\frac{\partial f_{ij}}{\partial t_{hk}} \leq 0
\]  \hspace{1cm} (1)

i.e. \( f_{ij} \) is a monotonically non-increasing function of the \( t_{hk} \)'s.

Secondly, assume that two documents \( d_i \) and \( d_j \) are linked exclusively through a third document \( d_k \), i.e. that every path \( P_{ij} \) between \( d_i \) and \( d_j \) contains \( d_k \). In this case, it seems natural to require that the distance function \( f_{ij} \) be additive, or

\[
f_{ij} = f_{ik} + f_{kj}
\]  \hspace{1cm} (2)

We must point out, at this stage, that more than to a semantic similarity between documents, we are aiming to some easily and mechanically
computable correlation based on the citation association.

Returning now to our main line, we notice that the well-known function "resistance" defined over the graph U would meet our previous requirements (1), (2). The graph U is considered as a resistive network, in which each edge $b_{hk}$ is assigned a resistance $1/t_{hk}$. Since the resistance $R_{ij}$ between any two nodes $d_i, d_j$ of U is well-defined we could let

$$f_{ij} = R_{ij}.$$ 

In addition to verifying (1) and (2), $R_{ij}$ is also a metric function.

Another well-known function which could be adopted as a measure of distance is the "reliability" between pairs of nodes. We recall that reliability $r_{ij}$ between $d_i$ and $d_j$ is the probability of establishing a transmission path between $d_i$ and $d_j$ if $t_{hk}$ is the probability of correct functioning for the edge $b_{hk}$. It is easy to recognize that both requirements (1) and (2) are verified by $r_{ij}$.

A number of topological techniques are known for the evaluation of either the resistance function or the reliability function respectively. These techniques are satisfactory for most applications. In computer based information retrieval systems however, the procedure must be applied many times for each retrieval operation and simplicity in methods employed is of utmost importance.

For this reason, we turn our attention to another function which can be defined for each pair of nodes of U. We recall that a circuit is a set of m undirected edges $b_1, b_2, \ldots, b_m$ such that: i) each $b_j$ can be oriented; ii) the terminal of $b_j$ coincides with the original of $b_{j+1}$; iii) the terminal of $b_m$ coincides with the origin of $b_1$. Obviously a circuit
$G_{ij}$ containing $d_i$ and $d_j$ is composed of two paths which are edge-disjoint (but not necessarily node-disjoint). We can now give the following

**Definition:** Let $G_{ij}^{(1)}, G_{ij}^{(2)}, \ldots, G_{ij}^{(n)}$ be the totality of distinct circuits containing two distinct nodes $d_i$ and $d_j$. We define as the length of the circuit $G_{ij}^{(s)}$ ($s = 1, 2, \ldots, n$)

$$\mathcal{L}[G_{ij}^{(s)}],$$

the sum of $1/t_{hk}$ over each edge belonging to $G_{ij}^{(k)}$. Then we let

$$f_{ij} = \min_{s} \mathcal{L}[G_{ij}^{(s)}]. \quad (3)$$

We note that $f_{ij}$ satisfies requirements (1) and (2). In fact, if $t_{hk}$ is the weight of edge $b_{hk}$ and $G_{ij}$ is a minimum length circuit, then

$$f_{ij} = \sum_{b_{hk} \in G_{ij}} \frac{1}{t_{hk}}.$$

It follows that

$$\frac{\partial f_{ij}}{\partial t_{hk}} = \begin{cases} 0 & \text{if } b_{hk} \notin G_{ij} \\ -\frac{1}{t_{hk}^2} & < 0 \text{ if } b_{hk} \in G_{ij} \end{cases}$$

By letting $f_{ii} = 0$ for each $i$, verification of property (2) follows from the stronger statement that $f_{ij}$, as given by (3), is a metric function. The proof of this assertion is considerably simplified by the following lemma.

**Lemma:** If there is a circuit $G_1$ containing $d_1$ and $d_2$ and a circuit $G_2$ containing $d_2$ and $d_3$, then there exists a circuit containing $d_1$ and $d_3$. 
Proof: Let $G_1$ consist of the two edge-disjoint paths $P_1$, $P_2$ and similarly $G_2$ consist of $P_3$, $P_4$. Since $P_1 \cap G_2$ is non-empty, (at least they contain node $d_2$) starting from $d_1$ and proceeding on $P_1$, let $d_1^*$ be the first node of $P_1$ which also belongs to $G_2$. Similarly, let $d_2^*$ be the analogous node on $P_2$. We have now the following two situations:

1) $d_1^*$, $d_2^*$ belong to the same path of $G_2$, say $P_3$. Then traversing $P_3$ from $d_3$ to $d_2$, assume, with no loss of generality, that we first reach $d_1^*$ (if $d_1^* = d_2^*$, it is immaterial which $d_j^*$ ($j = 1,2$) is chosen as the first node reached). Path $P_3$ is therefore partitioned into paths $d_3^*P_3d_1^*$, $d_1^*P_3d_2^*$, $d_2^*P_3d_2^*$, with $d_1^*P_3d_2^*$ possibly empty. We then form the following paths $P_1^*$, $P_2^*$:

$$
P_1^* : \quad d_1 \quad P_1 \quad d_1^* \quad P_3 \quad d_3
$$

$$
P_2^* : \quad d_3 \quad P_4 \quad d_2 \quad P_3 \quad d_2^* \quad P_2 \quad d_1
$$

We claim that $G^* = P_1^* \cup P_2^*$ is a circuit. In fact the path $d_1P_1d_1^*$ is edge-disjoint from $d_2P_2d_1$ by hypothesis and from $d_3^*P_3d_2^*$ by construction (since $d_4P_1d_1^*$ contains no edge of $G_2$). Similarly $d_1^*P_3d_3$ is edge-disjoint from $d_3^*P_3d_2^*$ by hypothesis and from $d_2^*P_2d_1$ by construction (since the latter contains no edge of $G_2$).

2) $d_1^*$, $d_2^*$ belong to different paths of $G_2$. Assume $d_1^* \in P_3$ and $d_2^* \in P_4$. Then we form the two paths

$$
P_1^* : \quad d_1 \quad P_1 \quad d_1^* \quad P_3 \quad d_3
$$

$$
P_2^* : \quad d_3 \quad P_4 \quad d_2 \quad P_2 \quad d_1
$$

and argue as in case 1.

Q.E.D.
We see therefore that $f_{ij}$, as given by (3), is real-valued, satisfies the reflexive property by definition and the symmetric property because of the undirectedness of $U$. The triangle inequality follows from Lemma 1, since, with the same symbols, $G^*$ consists of a subset (proper or improper) of the edges of $G_1 \cup G_2$. Hence

$$\mu[G^*] \leq \mu[G_1] + \mu[G_2]$$

and the inequality holds also when $G_1$ and $G_2$ are of minimal length. We have therefore proved

**Theorem**: The function $f_{ij}$ (3) is a metric function.

In addition to some other reason which we shall mention later, an interesting feature of function (3) is the relative ease with which it can be mechanically computed.

A **string** $S$ is a sequence over the set of symbols (integers) $1, 2, \ldots, n$. Over the set of strings we define the operation of a **string product**: The string product of $S_1$ and $S_2$ is their concatenation $S_1 \cdot S_2$. Clearly, the string product is associative but not commutative. With the symbol 0 we denote the zero string, i.e., the string of no symbols. By definition, for every $S$, $0 \cdot S = S \cdot 0 = 0$. Further a string product $S$ is 0 in the following circumstances (nullification rules):

**Rule i)** $S$ is of the form ...hk...hk... or ...hk...kh... (i.e. a given pair of consecutive symbols is repeated either in the same order or in reversed order).

**Rule ii)** $S$ is of the form h...h (i.e. the first and the last symbols of $S$ coincide).
Given these definitions, we construct the matrix $M$, obtained from $A$ by replacing each $t_{hk} > 0$ with the integer $k$, which is now regarded as a symbol in the sense specified above.

Assume now, for simplicity, that we aim to compute the distance with respect to $d_1$. We multiply the first row $u^{(1)}$ of $M$ by $M$ and replace the ordinary operation of multiplication with the just defined string product. We obtain the vector

$$u^{(2)} = u^{(1)}M.$$

We iterate this operation $s-1$ times and obtain

$$u^{(s)} = u^{(s-1)}M.$$

Let us analyze $u^{(s)}$ for $s \geq 3$. Its first component, which is then conventionally set to 0 (rule ii), gives a collection of circuits containing $d_1$ and composed of $s$ edges: in fact rules 1, 2) of nullification of the string product ensure us that no edge is traversed more than once. By this iterative procedure we can obtain all circuits containing $d_1$ with up to $s$ edges.

The computation of the distance becomes trivial in the particular case in which all edges are equally weighted, e.g. $t_{hk} = 1$ for any existing edge. In this case the distance is simply the number of edges of the shortest circuit containing the access node and the node under consideration. We can therefore give the following computer-oriented algorithm for the search of all documents up to distance $s$ from a specified document where $s$ is used as a control parameter. The algorithm takes advantage of the fact that the $T$ matrix is in effect very sparse: while its order could be around several tens of thousands, the number of non-zero entries per row (the degree of the node) is, on the average, close to 10.
Algorithm. Each document \( d_i \in B \) is specified through its accession number, for simplicity, \( i \). With each \( i \) we associate a list \( L_i \), i.e. a collection of integers which are the accession numbers of the documents directly linked through citation with \( i \); the integers belonging to \( L_i \) are assumed to be naturally ordered.

Let \( i \) be the document specified by the query, i.e. the access point. With \( L \) we designate the current list: each term of \( L \) is, in general, a sum of all the string products having equal last symbol; the terms are ordered by increasing last symbol.

1. Set \( r = 2 \). Let \( L = L_1 \).
2. Let \( \alpha_1, \alpha_2, ..., \alpha_n \) be the last symbols of the terms of \( L \). Set \( j = 1 \).
3. Call from the archive list \( L_{\alpha_j} \) and form the string product of the term ending with \( \alpha_j \) by each term of \( L_{\alpha_j} \). If \( j < n_r \), replace \( j \) with \( j+1 \) and repeat step 3; if \( j = n_r \) go to 4).
4. Sort all string products obtained in iterations of step 3 by increasing last symbol; form new terms by adding all string products with equal last symbol. For \( r > 2 \), the term ending with \( i \) provides all circuits of length \( r \).
5. Apply nullification rules i) and ii) on the list obtained in step 4. The resulting list is the new \( L \). If \( r = s \), the algorithm terminates. If \( r < s \), replace \( r \) with \( r+1 \) and return to step 2.

The described algorithm provides all circuits containing the access node and having up to \( s \) edges: the actual computation of the distance requires no further comment.
We must not overlook the possible objection that however simple the previous algorithm may appear, the length of the current list \( L \) may reach extremely high values for sufficiently high \( s \). This geometric explosion with ratio equal to the average degree of the nodes would certainly take place if document-links were assigned at random. In our case, however, it appears that the structure of the citation network, through the strong interconnection of documents in a given subject area, acts in favor of a much milder increase: simple manual trials appear to confirm this intuition, but only more extensive experiments can have a probatory value.

Another promising feature of the circuit concept is related to the remark that possibly irrelevant documents, relatively close through citation to the access document, are excluded from the retrieved set \( R \): the intuition, in fact, would suggest that if there is only one path from the access node to the node representative of a given document, the latter is most likely not subject-related to the query.

IV. Schemes for Combined Retrieval

In the two previous sections we have analyzed the correlation graph and the citation graph as two structural organizations which can be conveniently exploited for document retrieval. As mentioned in the introduction, it seems very attractive to combine the power of the two structures in order to mitigate their respective shortcomings, i.e. the disturbance or "noise" caused, for example, by homographs in coordinate indexing or by careless citation.
If the query is specified by a single document (and there seems to be no conceptual difficulty in passing from single to composite queries), by following the criteria presented in Sections II and III, we can compute two distances of each document $d_j$ from the query $d_i$: i.e. $f_{ij}^{(1)}$, as obtained from the correlation graph, and $f_{ij}^{(2)}$, as obtained from the citation graph. The combined distance $F_{ij}$ must very reasonably be an increasing function of $f_{ij}^{(1)}$ and $f_{ij}^{(2)}$. The two simplest expressions of $F_{ij}$ which we propose are

$$f_{ij} = a_1 f_{ij}^{(1)} + b_1 f_{ij}^{(2)}$$

$$\ln F_{ij} = a_2 \ln f_{ij}^{(1)} + b_2 \ln f_{ij}^{(2)}$$

where $a_1, b_1, a_2, b_2$ are positive constants. We remark that function (4) corresponds to the set theoretical operation of union when applied to the two graphs, while (5) corresponds to the set theoretical operation of intersection.

No insight has so far been obtained into the possible values of the constants $a_1, a_2, b_1, b_2$. An extensive experiment has been planned which should shed light on this aspect of the proposed scheme, as well as on further theoretical developments.

Acknowledgment

The authors wish to acknowledge the many stimulating discussions they have had on information retrieval with their colleagues Drs. Dewey Carroll, Sylvian Ray and Paul Weston.
This paper considers the problem of information retrieval from the point of view of graph theory. In this formulation documents are represented as nodes and relationships among the documents are represented by edges. Two types of graphs are introduced, namely the similarity graph which is based on subject-content correlation and the citation graph, which is derived from direct citation linkages among documents. Several distance measures are considered and evaluated with regard to retrieval operations.
Graph theory
Information retrieval
Digital systems
Computers