[redacted]

Coordinated Science Laboratory

UNIVERSITY OF ILLINOIS - URBANA, ILLINOIS

TOPOLOGICAL STRUCTURES OF INFORMATION RETRIEVAL SYSTEMS

R. T. Chien and F. P. Preparata

REPORT R-325 October, 1966

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.

TOPOLOGICAL STRUCTURES OF INFORMATION RETRIEVAL SYSTEMS

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

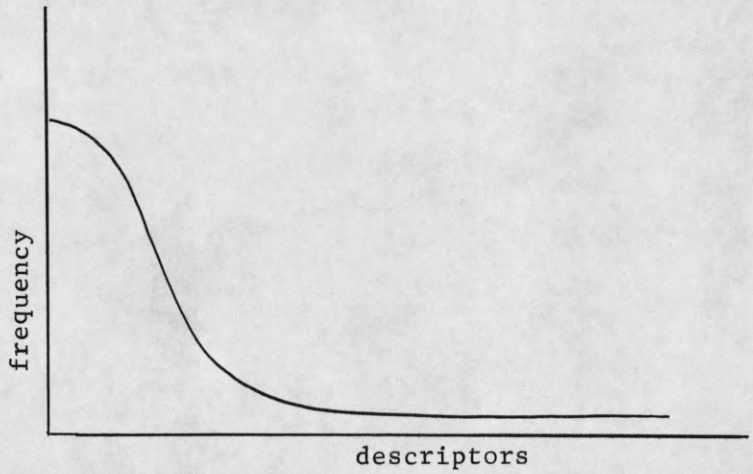


Figure 1. Descriptor Frequency Distribution

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 can not 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 jth descriptor is assigned to the ith document and zero otherwise. We define the mxm 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 compulation 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 $B \equiv \{d_1, d_2, \ldots, d_n\}$ consisting of documents $d_1, d_2, \ldots, d_n$, the directed citation graph $\mathcal{D}$ pertaining to B is entirely described by an nxn matrix $E \equiv \|e_{ij}\|$, 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 ($R \supset Q$) which are subject-related to the documents of Q. In the simplest instance, $Q \equiv d_i$, i.e. it contains a single document d_i. d_i is denoted as the access point.

The determination of the retrieved set R 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 $\mathcal{D}$. 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 $\mathcal{D}$ with the undirected graph $\mathcal{U}$, simply denoted as the citation graph. $\mathcal{U}$ is described by the nxn 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 $\mathcal{U}$: 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

$$\frac{\partial f_{ij}}{\partial t_{hk}} \leq 0 \qquad (1)$$

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 and each 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

$$f_{ij} = f_{ik} + f_{kj} \,. \qquad (2)$$

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 $\mathcal{U}$ would meet our previous requirements (1), (2). The graph $\mathcal{U}$ 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 $\mathcal{U}$ 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 $\mathcal{U}$. We recall that a <u>circuit</u> 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 $G_{ij}^{(s)}$ $(s = 1,2,\ldots,n)$

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

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

$$f_{ij} = \min_{s} \ell[G_{ij}^{(s)}] . \tag{3}$$

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

$$f_{ij} = \sum_{b_{hk} \in \bar{G}_{ij}} \frac{1}{t_{hk}} .$$

It follows that

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

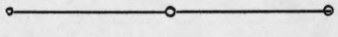
By letting $f_{ii} = 0$ for each i, verification of property (2) follows from the stronger statement that f_{ij}, as given by (3), is a metric function. The proof of this assertion is considerably simplified by the following lemma.

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.

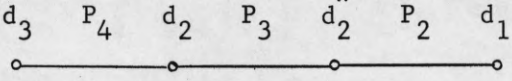
Proof: Let G_1 consist of the two edge-disjoint paths P_1, P_2 and similarly G_2 consist of P_3, P_4. Since $P_1 \cap G_2$ is non-empty, (at least they contain node d_2) starting from d_1 and proceeding on P_1, let d_1^* be the first node of P_1 which also belongs to G_2. Similarly, let d_2^* be the 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 $d_1^* \equiv d_2^*$, it is immaterial which d_j^* ($j = 1,2$) is chosen as the first node reached). Path P_3 is therefore partitioned into paths $d_3P_3d_1^*$, $d_1^*P_3d_2^*$, $d_2^*P_3d_2$, with $d_1^*P_3d_2^*$ possibly empty. We then form the following paths P_1^*, P_2^* :

P_1^* : d_1 P_1 d_1^* P_3 d_3



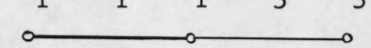
P_2^* : d_3 P_4 d_2 P_3 d_2^* P_2 d_1



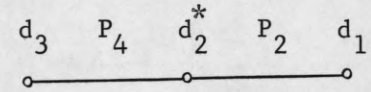
We claim that $G^* = P_1^* \cup P_2^*$ is a circuit. In fact the path $d_1P_1d_1^*$ is edge-disjoint from $d_2^*P_2d_1$ by hypothesis and from $d_3P_4d_2P_3d_2^*$ by construction (since $d_4P_1d_1^*$ contains no edge of G_2). Similarly $d_1^*P_3d_3$ is edge-disjoint from $d_3P_4d_2P_3d_2^*$ by hypothesis 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^* : d_1 P_1 d_1^* P_3 d_3



P_2^* : d_3 P_4 d_2^* P_2 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

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

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,...,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 $\underline{u}^{(1)}$ of M by M and replace the ordinary operation of multiplication with the just defined string product. We obtain the vector

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

We iterate this operation s-1 times and obtain

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

Let us analyze $\underline{u}^{(s)}$ 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 $L = L_i$.

2. Let $\alpha_1, \alpha_2, \ldots, \alpha_{n_r}$ be the last symbols of the terms of L. Set j = 1.

3. Call from the archive list L_{α_j} and form the string product of the term ending with α_j by each term of L_{α_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

$$f_{ij} = a_1 f_{ij}^{(1)} + b_1 f_{ij}^{(2)} \tag{4}$$

$$\ln f_{ij} = a_2 \ln f_{ij}^{(1)} + b_2 \ln f_{ij}^{(2)} \tag{5}$$

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.

Distribution list as of May 1, 1966

1 Dr. Edward M. Reilley
Asst. Director (Research)
Ofc. of Defense Res. & Engrg.
Department of Defense
Washington, D. C. 20301

1 Office of Deputy Director
(Research and Information Rm 3D1037)
Department of Defense
The Pentagon
Washington, D. C. 20301

1 Director
Advanced Research Projects Agency
Department of Defense
Washington, D. C. 20301

1 Director for Materials Sciences
Advanced Research Projects Agency
Department of Defense
Washington, D. C. 20301

1 Headquarters
Defense Communications Agency (333)
The Pentagon
Washington, D. C. 20305

20 Defense Documentation Center
Attn: TISIA
Cameron Station, Building 5
Alexandria, Virginia 22314

1 Director
National Security Agency
Attn: Librarian C-332
Fort George G. Meade, Maryland 20755

1 Weapons Systems Evaluation Group
Attn: Col. Finis G. Johnson
Department of Defense
Washington, D. C. 20305

1 National Security Agency
Attn: R4-James Tippet
Office of Research
Fort George G. Meade, Maryland 20755

1 Central Intelligence Agency
Attn: OCR/DD Publications
Washington, D. C. 20505

1 AFRSTE
Hqs. USAF
Room 1D-429, The Pentagon
Washington, D. C. 20330

1 AUL3T-9663
Maxwell Air Force Base, Alabama 36112

1 AFFTC (FTBPP-2)
Technical Library
Edwards AFB, California 93523

1 Space Systems Division
Air Force Systems Command
Los Angeles Air Force Station
Los Angeles, California 90045
Attn: SSSD

1 SSD(SSTRT/Lt. Starbuck)
AFUPO
Los Angeles, California 90045

1 Det. #6, OAR (LOOAR)
Air Force Unit Post Office
Los Angeles, California 90045

1 Systems Engineering Group (RTD)
Technical Information Reference Branch
Attn. SEPIR
Directorate of Engineering Standards
& Technical Information
Wright-Patterson AFB, Ohio 45433

1 ARL (ARIY)
Wright-Patterson AFB, Ohio 45433

1 AFAL (AVT)
Wright-Patterson AFB, Ohio 45433

1 AFAL (AVTE/R. D. Larson)
Wright-Patterson AFB, Ohio 45433

1 Office of Research Analyses
Attn: Technical Library Branch
Holloman AFB, New Mexico 88330

2 Commanding General
Attn: STEWS-WS-VT
White Sands Missile Range
New Mexico 88002

1 RADC (EMLAL-I)
Griffiss AFB, New York 13442
Attn: Documents Library

1 Academy Library (DFSLB)
U. S. Air Force Academy
Colorado 80840

1 FJSRL
USAF Academy, Colorado 80840

1 APGC (PGBPS-12)
Eglin AFB, Florida 32542

1 AFETR Technical Library
(ETV, MU-135)
Patrick AFB, Florida 32925

1 AFETR (ETLLG-I)
STINFO Officer (for Library)
Patrick AFB, Florida 32925

1 AFCRL (CRMXLR)
AFCRL Research Library, Stop 29
L. G. Hanscom Field
Bedford, Massachusetts 01731

2 ESD (ESTI)
L. G. Hanscom Field
Bedford, Massachusetts 01731

1 AEDC (ARO, INC)
Attn: Library/Documents
Arnold AFS, Tennessee 37389

2 European Office of Aerospace Research
Shell Building
47 Rue Cantersteen
Brussels, Belgium

5 Lt. Col. E. P. Gaines, Jr.
Chief, Electronics Division
Directorate of Engineering Sciences
Air Force Office of Scientific Research
Washington, D. C. 20333

1 U. S. Army Research Office
Attn. Physical Sciences Division
3045 Columbia Pike
Arlington, Virginia 22204

1 Research Plans Office
U. S. Army Research Office
3045 Columbia Pike
Arlington, Virginia 22204

1 Commanding General
U. S. Army Materiel Command
Attn: AMCRD-RS-PE-E
Washington, D. C. 20315

1 Commanding General
U. S. Army Strategic Communications Command
Washington, D. C. 20315

1 Commanding Officer
U. S. Army Materials Research Agency
Watertown Arsenal
Watertown, Massachusetts 02172

1 Commanding Officer
U. S. Army Ballistics Research Laboratory
Attn: V. W. Richards
Aberdeen Proving Ground
Aberdeen, Maryland 21005

1 Commandant
U. S. Army Air Defense School
Attn: Missile Sciences Division C&S Dept.
P. O. Box 9390
Fort Bliss, Texas 79916

1 Commanding General
U. S. Army Missile Command
Attn: Technical Library
Redstone Arsenal, Alabama 35809

1 Commanding General
Frankford Arsenal
Attn: SMUFA-L6000 (Dr. Sidney Ross)
Philadelphia, Pennsylvania 19137

1 U. S. Army Munitions Command
Attn: Technical Information Branch
Picatinney Arsenal
Dover, New Jersey 07801

1 Commanding Officer
Harry Diamond Laboratories
Attn: Mr. Berthold Altman
Connecticut Avenue & Van Ness Street N. W.
Washington, D. C. 20438

1 Commanding Officer
U. S. Army Security Agency
Arlington Hall
Arlington, Virginia 22212

1 Commanding Officer
U. S. Army Limited War Laboratory
Attn: Technical Director
Aberdeen Proving Ground
Aberdeen, Maryland 21005

1 Commanding Officer
Human Engineering Laboratories
Aberdeen Proving Ground, Maryland 21005

1 Director
U. S. Army Engineer Geodesy, Intelligence
and Mapping
Research and Development Agency
Fort Belvoir, Virginia 22060

1 Commandant
U. S. Army Command and General Staff College
Attn: Secretary
Fort Leavenworth, Kansas 66270

1 Dr. H. Robl, Deputy Chief Scientist
U. S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

1 Commanding Officer
U. S. Army Research Office (Durham)
Attn: CRD-AA-IP (Richard O. Ulsh)
Box CM, Duke Station
Durham, North Carolina 27706

1 Superintendent
U. S. Army Military Academy
West Point, New York 10996

1 The Walter Reed Institute of Research
Walter Reed Medical Center
Washington, D. C. 20012

1 Commanding Officer
U. S. Army Electronics R&D Activity
Fort Huachuca, Arizona 85163

1 Commanding Officer
U. S. Army Engineer R&D Laboratory
Attn: STINFO Branch
Fort Belvoir, Virginia 22060

1 Commanding Officer
U. S. Army Electronics R&D Activity
White Sands Missile Range, New Mexico 88002

1 Dr. S. Benedict Levin, Director
Institute for Exploratory Research
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

1 Director
Institute for Exploratory Research
U. S. Army Electronics Command
Attn: Mr. Robert O. Parker, Executive
Secretary, JSTAC (AMSEL-XL-D)
Fort Monmouth, New Jersey 07703

1 Commanding General
U. S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Attn: AMSEL-SC
RD-D
RD-G
RD-GF
RD-MAF-I
RD-MAT
XL-D
XL-E
XL-C
XL-S
HL-D
HL-L
HL-J
HL-P
HL-O
HL-R
NL-D
NL-A
NL-P
NL-R
NL-S
KL-D
KL-E
KL-S
KL-T
VL-D
WL-D

3 Chief of Naval Research
Department of the Navy
Washington, D. C. 20360
Attn: Code 427

4 Chief, Bureau of Ships
Department of the Navy
Washington, D. C. 20360

3 Chief, Bureau of Weapons
Department of the Navy
Washington, D. C. 20360

2 Commanding Officer
Office of Naval Research Branch Office
Box 39, Navy No. 100 F.P.O.
New York, New York 09510

3 Commanding Officer
Office of Naval Research Branch Office
219 South Dearborn Street
Chicago, Illinois 60604

1 Commanding Officer
Office of Naval Research Branch Office
1030 East Green Street
Pasadena, California

1 Commanding Officer
Office of Naval Research Branch Office
207 West 24th Street
New York, New York 10011

Distribution list as of May 1, 1966 (cont'd.)

1 Commanding Officer
Office of Naval Research Branch Office
495 Summer Street
Boston, Massachusetts 02210

8 Director, Naval Research Laboratory
Technical Information Officer
Washington, D. C.
Attn: Code 2000

1 Commander
Naval Air Development and Material Center
Johnsville, Pennsylvania 18974

2 Librarian
U. S. Naval Electronics Laboratory
San Diego, California 95152

1 Commanding Officer and Director
U. S. Naval Underwater Sound Laboratory
Fort Trumbull
New London, Connecticut 06840

1 Librarian
U. S. Navy Post Graduate School
Monterey, California

1 Commander
U. S. Naval Air Missile Test Center
Point Magu, California

1 Director
U. S. Naval Observatory
Washington, D. C.

2 Chief of Naval Operations
OP-07
Washington, D. C.

1 Director, U. S. Naval Security Group
Attn: G43
3801 Nebraska Avenue
Washington, D. C.

2 Commanding Officer
Naval Ordnance Laboratory
White Oak, Maryland

1 Commanding Officer
Naval Ordnance Laboratory
Corona, California

1 Commanding Officer
Naval Ordnance Test Station
China Lake, California

1 Commanding Officer
Naval Avionics Facility
Indianapolis, Indiana

1 Commanding Officer
Naval Training Device Center
Orlando, Florida

1 U. S. Naval Weapons Laboratory
Dahlgren, Virginia

1 Weapons Systems Test Division
Naval Air Test Center
Patuxtent River, Maryland
Attn: Library

1 Mr. Charles F. Yost
Special Assistant to the Director of Research
National Aeronautics and Space Administration
Washington, D. C. 20546

1 Dr. H. Harrison, Code RRE
Chief, Electrophysics Branch
National Aeronautics and Space Administration
Washington, D. C. 20546

1 Goddard Space Flight Center
National Aeronautics and Space Administration
Attn: Library, Documents Section Code 252
Greenbelt, Maryland 20771

1 NASA Lewis Research Center
Attn: Library
21000 Brookpark Road
Cleveland, Ohio 44135

1 National Science Foundation
Attn: Dr. John R. Lehmann
Division of Engineering
1800 G Street, N. W.
Washington, D. C. 20550

1 U. S. Atomic Energy Commission
Division of Technical Information Extension
P. O. Box 62
Oak Ridge, Tennessee 37831

1 Los Alamos Scientific Laboratory
Attn: Reports Library
P. O. Box 1663
Los Alamos, New Mexico 87544

2 NASA Scientific & Technical Information Facility
Attn: Acquisitions Branch (S/AK/DL)
P. O. Box 33
College Park, Maryland 20740

1 Director
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

1 Polytechnic Institute of Brooklyn
55 Johnson Street
Brooklyn, New York 11201
Attn: Mr. Jerome Fox
Research Coordinator

1 Director
Columbia Radiation Laboratory
Columbia University
538 West 120th Street
New York, New York 10027

1 Director
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

1 Director
Stanford Electronics Laboratories
Stanford University
Stanford, California

1 Director
Electronics Research Laboratory
University of California
Berkeley 4, California

1 Director
Electronic Sciences Laboratory
University of Southern California
Los Angeles, California 90007

1 Professor A. A. Dougal, Director
Laboratories for Electronics and
Related Sciences Research
University of Texas
Austin, Texas 78712

1 Division of Engineering and Applied Physics
210 Pierce Hall
Harvard University
Cambridge, Massachusetts 02138

1 Aerospace Corporation
P. O. Box 95085
Los Angeles, California 90045
Attn: Library Acquisitions Group

1 Professor Nicholas George
California Institute of Technology
Pasadena, California

1 Aeronautics Library
Graduate Aeronautical Laboratories
California Institute of Technology
1201 E. California Boulevard
Pasadena, California 91109

1 Director, USAF Project RAND
Via: Air Force Liaison Office
The RAND Corporation
1700 Main Street
Santa Monica, California 90406
Attn: Library

1 The Johns Hopkins University
Applied Physics Laboratory
8621 Georgia Avenue
Silver Spring, Maryland
Attn: Boris W. Kuvshinoff
Document Librarian

1 Hunt Library
Carnegie Institute of Technology
Schenley Park
Pittsburgh, Pennsylvania 15213

1 Dr. Leo Young
Stanford Research Institute
Menlo Park, California

1 Mr. Henry L. Bachmann
Assistant Chief Engineer
Wheeler Laboratories
122 Cuttermill Road
Great Neck, New York

1 University of Liege
Electronic Department
Mathmatics Institute
15, Avenue Des Tilleuls
Val-Benoit, Liege
Belgium

1 School of Engineering Sciences
Arizona State University
Tempe, Arizona

1 University of California at Los Angeles
Department of Engineering
Los Angeles, California

1 California Institute of Technology
Pasadena, California
Attn: Documents Library

1 University of California
Santa Barbara, California
Attn: Library

1 Carnegie Institute of Technology
Electrical Engineering Department
Pittsburgh, Pennsylvania

1 University of Michigan
Electrical Engineering Department
Ann Arbor, Michigan

1 New York University
College of Engineering
New York, New York

1 Syracuse University
Department of Electrical Engineering
Syracuse, New York

1 Yale University
Engineering Department
New Haven, Connecticut

1 Airborne Instruments Laboratory
Deerpark, New York

1 Bendix Pacific Division
11600 Sherman Way
North Hollywood, California

1 General Electric Company
Research Laboratories
Schenectady, New York

1 Lockheed Aircraft Corporation
P. O. Box 504
Sunnyvale, California

1 Raytheon Company
Bedford, Massachusetts
Attn: Librarian

Security Classification

DOCUMENT CONTROL DATA R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
University of Illinois
Coordinated Science Laboratory
Urbana, Illinois 61801

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

TOPOLOGICAL STRUCTURES OF INFORMATION RETRIEVAL SYSTEMS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (Last name, first name, initial)
Chien, R.T. & Preparata, F.P.

6. REPORT DATE	7a TOTAL NO. OF PAGES	7b. NO. OF REFS.
October, 1966	15	

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DA 28 043 AMC 00073(E) b. PROJECT NO. 20014501B31F: also NSF GK-690. c. d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this report is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Joint Services Electronics Program thru U.S. Army Electronics Command Ft. Monmouth, New Jersey 07703

13. 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.

DD FORM 1 JAN 64 1473

Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Graph theory						
Information retrieval						
Digital systems						
Computers						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (corporate author) issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (either by the originator or by the sponsor), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (paying for) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473