Associative Document Retrieval Techniques Using Bibliographic Information*

GERARD SALTON

Harvard University,† Cambridge, Massachusetts

Abstract. Automatic documentation systems which use the words contained in the individual documents as a principal source of document identifications may not perform satisfactorily under all circumstances. Methods have therefore been devised within the last few years for computing association measures between words and between documents, and for using such associated words, or information contained in associated documents, to supplement and refine the original document identifications. It is suggested in this study that bibliographic citations may provide a simple means for obtaining associated documents to be incorporated in an automatic documentation system.

The standard associative retrieval techniques are first briefly reviewed. A computer experiment is then described which tends to confirm the hypothesis that documents exhibiting similar citation sets also deal with similar subject matter. Finally, a fully automatic document retrieval system is proposed which uses bibliographic information in addition to other standard criteria for the identification of document content, and for the detection of relevant information.

1. *Introduction*

In recent years considerable attention has been devoted to the design of automatic documentation systems. If the system is to operate fully automatically, the intervention of human experts for the analysis of document content and for the preparation of document identifications ought to be eliminated. Under these circumstances the retrieval system must of necessity be based primarily on the words occurring in the individual texts, and on the terms used to formulate the search requests.

It has been suggested [1] that an acceptable system can be generated by extracting from the texts and from the information requests those linguistic units which are believed to be representative of document content, and by defining a standard of comparison between words extracted from documents and words used in the requests for documents. To determine which words are particularly significant as an indication of document content a variety of criteria may be used, including the position of the words in the texts, the word types, the vocabulary size, and most importantly the frequency of occurrence of the individual words. The most significant words are then used as "index terms" to characterize the documents, and the most significant sentences, that is, those containing a large number of significant words, are used as abstracts for the documents.

A typical automatic indexing and abstracting system based on word frequency

* Received July, 1962; revised March, 1963. This study was supported in part by the Air Force Cambridge Research Laboratories and in part by Sylvania Electric Products, Inc.

† Computation Laboratory.

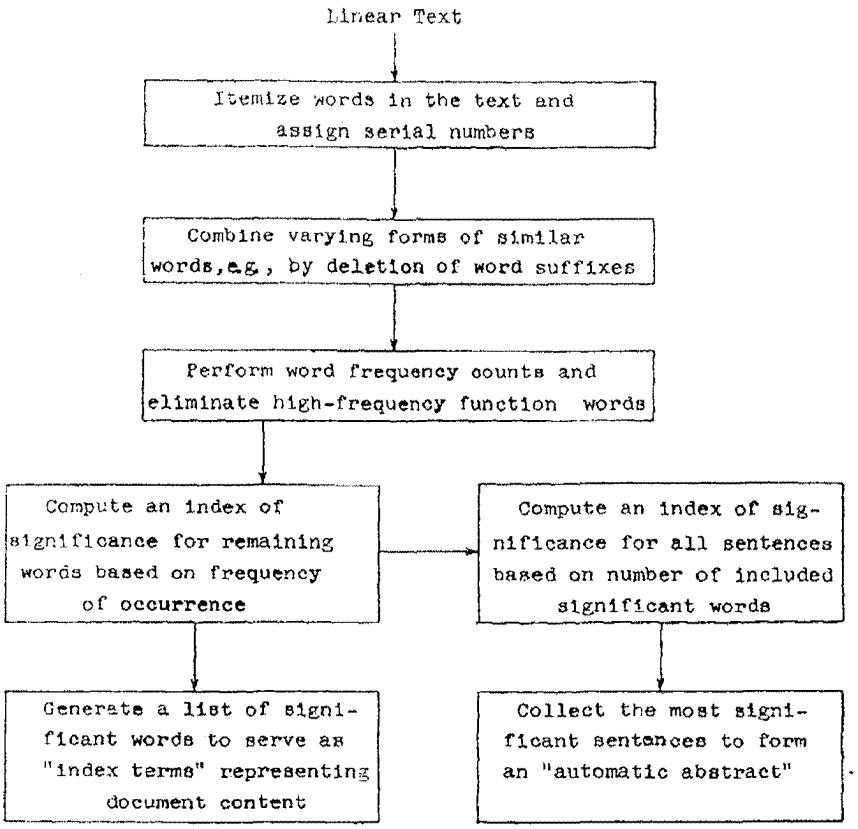

FIG. 1. Typical automatic indexing and abstracting system based on word frequency counts.

counts is shown in Figure 1. The principal drawback of the system outlined in Figure 1 is the lack of any normalization procedure designed to take into account differences between individual authors or between individual document types. Thus, a given set of documents covering some homogeneous subject area may quite possibly give rise to many different index sets. Similarly, completely different document sets may be obtained in answer to only slightly differing search requests.

In order to reduce the importance attached to the individual words and to their frequencies of occurrence, the introduction of a synonym dictionary, or thesaurus, is often proposed. All words extracted from documents or search requests could then be replaced by standard thesaurus forms before being used. This solution, while attractive in theory, is difficult to implement because no definite criteria exist for the construction of good or useful thesauruses, and because the generation of any thesaurus is a complex and time-consuming undertaking. For this reason, several workers [2, 3, 4, 5] have been interested in automatic procedures designed to supplement the original terms extracted from the documents with

new terms related to the old ones in various ways. Indexing techniques which make use of such "associated" terms have come to be known as "associative indexing," and the corresponding retrieval operations are known as "associative retrieval."

The present report suggests an extension of the usual associative retrieval techniques by taking into account bibliographic citations and other information peculiar to the author of a given document. It is suggested, specifically, that the set of identifying words extracted from the documents be supplemented by new words obtained in part from the bibliographic information provided with the documents; these new expanded sets of index terms may then give a more accurate representation of document content than the original ones and may thus provide a more effective retrieval mechanism.

The standard associative indexing techniques are first briefly reviewed. Thereafter, some properties of bibliographic citations are described, and the role of bibliographic information as an indication of document content is evaluated. A small computer experiment using citations is then summarized and the significance of the numeric results is discussed. Finally, a proposed fully automatic document retrieval system using bibliographic information in addition to other criteria is described.

2. *Associative Information Retrieval*

Most associative retrieval systems are based on the statistical word frequency counting procedures previously illustrated in Figure 1. Thus, given a document collection, it is possible to extract a set of n distinct high-frequency words $W_1, W_2, \cdots, W_n$, such that each document within the collection is initially identified by some subset of the set of n given words.

In practical retrieval systems, it becomes useful to provide for some additional flexibility. For example, given a search request expressed in terms of words in the natural language, it may be convenient to alter somewhat the original request, either by making it more specific and thus presumably reducing the size of the document set which fulfils the request or, alternatively, by making it more general. In the same way, given a set of terms identifying a specified document, it may be useful to alter somewhat the original set by deletion of old terms or addition of new ones in such a way that documents dealing with similar subject matter are identified by similar sets of index terms.

An analogous problem arises in connection with the document sets which are obtained in answer to certain search requests. It is often useful to alter these document sets by addition of further documents which may also have some relevance or, alternatively, by deletion of documents which are not directly relevant. Both questions can be treated by determining a *measure of association* between words or index terms on the one hand and between documents on the other, and by using this association measure for the alteration of the corresponding index term and document subsets.

Consider first the problem of word associations. Words may be related in

Terms	Documents D_1	D_2	⋯		⋯	D_m
W_1	C_1^1	C_2^1	⋯	C_j^1	⋯	C_m^1
W_2	C_1^2	C_2^2	⋯	C_j^2	⋯	C_m^2
⋮			⋮			
W_n	C_1^n	C_2^n	⋯	C_j^n	⋯	C_m^n

$= \mathbf{C}$

(a) Typical term-document incidence matrix $\mathbf{C}$ ($C_j^i = n \leftrightarrow$ document D_j contains term W_i exactly n times)

Terms	Terms W_1	W_2	⋯	W_n
W_1	R_1^1	R_2^1	⋯	R_n^1
W_2	R_1^2	R_2^2	⋯	R_n^2
⋮			⋮	
W_n	R_1^n	R_2^n	⋯	R_n^n

$= \mathbf{R}$

(b) Typical term-term similarity matrix $\mathbf{R}$

$$\left(\mathbf{R}_i^j = \mathbf{R}_i^j = \sum_{k=1}^{m} \mathbf{C}_k^i \mathbf{C}_k^j \Big/ \sqrt{\left(\sum_{k=1}^{m} (\mathbf{C}_k^i)^2 \sum_{k=1}^{m} (\mathbf{C}_k^j)^2\right)}\right)$$

FIG. 2. Matrices used for the generation of term associations

many different ways: for example, they may exhibit the same word stems, or they may have similar syntactic properties, or they may be usable in the same contexts, and so on. The criteria of association used in most automatic programs do not normally require a determination of syntactic or semantic properties. Rather, they are based on simple co-occurrence of words in the same texts or sentences, or on co-occurrence with individual or joint frequencies greater than some given threshold value.

Given a set of m documents and a set of n index terms, a typical procedure for the generation of term associations is as follows:

(a) a term-document *incidence matrix* C is constructed which lists index terms against documents; matrix element $\mathbf{C}_j^i$ is defined to be equal to k if and only if document j contains term i exactly k times;
(b) a coefficient of similarity between terms is then defined based on the frequency of co-occurrence of pairs of terms in the individual documents;
(c) a term-term *similarity matrix* R is then generated which exhibits all similarity coefficients between pairs of index terms;
(d) term associations are defined for those pairs whose associated similarity coefficient is greater than some stated threshold value.

A sample term-document incidence matrix C is shown in Figure 2(a). To obtain a coefficient of similarity between two terms based on the frequency of co-occurrence in the documents of a given collection, it is only necessary to perform a pairwise comparison of the corresponding rows of C. Many different types of similarity coefficients have been suggested in the literature [2, 3, 4, 5]; a simple coefficient of similarity between rows of a numeric matrix, and one which may be as meaningful as any of the others, is the cosine of the angle between the

corresponding m-dimensional vectors [6]. The similarity coefficients can be displayed in an $n \times n$ symmetric term-similarity matrix $\mathbf{R}$, where the coefficient of similarity $\mathbf{R}_j^i$ between term W_i and term W_j is

$$\mathbf{R}_j^i = \mathbf{R}_i^j = \frac{\sum_{k=1}^{m} \mathbf{C}_k^i \mathbf{C}_k^j}{\sqrt{\left(\sum_{k=1}^{m} (\mathbf{C}_k^i)^2 \sum_{k=1}^{m} (\mathbf{C}_k^j)^2\right)}}.$$

The term-similarity matrix $\mathbf{R}$ corresponding to the term-document matrix $\mathbf{C}$ of Figure 2 (a) is shown in Figure 2 (b). Since $\mathbf{R}$ is symmetric, only the right (or left) triangular part of $\mathbf{R}$ must be scanned in order to detect pairs of terms with large similarity coefficients.

To generate document associations instead of term associations the same procedures can be used, since the strength of association between documents may be conveniently assumed to be a function of the number and frequencies of the shared terms in their respective term lists. Document similarities are therefore obtained by comparing pairs of columns (instead of rows) of the term-document matrix $\mathbf{C}$, and a document-document similarity matrix is constructed and used in the same way as the previously described term-term matrix $\mathbf{R}$.

Consider now a typical system for document retrieval using term and document associations as shown in Figure 3. A list of high-frequency terms is first generated for each document by word frequency counting procedures. Normalization may or may not be effected by thesaurus lookup. A term-term similarity matrix is then constructed by using co-occurrence of terms within sentences, rather than within documents, as a criterion. It should be noted that as new term associations are defined, the original incidence matrix can be revised by inclusion in some of the matrix columns of new, associated terms which are not originally contained in the respective sentences or documents. The revised incidence matrix then gives rise to a new term-term similarity matrix, incorporating second-order associations, and so on. This feedback process is represented by an upward-pointing arrow in Figure 3.

To retrieve documents in answer to search requests, the programs already available can be used by adding to the term-document matrix $\mathbf{C}$ a new column $\mathbf{C}_{m+1}$, representing the request terms. Specifically, element $\mathbf{C}_{m+1}^k$ is set equal to w if term W_k is used in the search request with weight w; if word W_k is not used in the given search request $\mathbf{C}_{m+1}^k$ is set equal to 0. If no weights are specified by the requestor the values of the elements of column $\mathbf{C}_{m+1}$ are restricted to 0 and 1. An estimate of document relevance is then obtained by computing for each document the similarity coefficient between the request column $\mathbf{C}_{m+1}$ and the respective document column. The documents can be arranged in decreasing order of similarity coefficients, and all documents with a sufficiently large coefficient can be judged to be relevant to the given request. Clearly, the final relevance criterion depends not only on the terms assigned to the various documents or on the words used in the documents and search requests, but also on other terms associated with the original ones through co-occurrence in a given document collection.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.