Essays of an Information Scientist, Vol:1, p.83, 1962-73
Current Contents, #9, March 4, 1970 Also see vo1p084-1962-73.pdf

"Current Comments"

Can Criticism and Documentation of Research Papers Be Automated?

Click for Reprint: "Can Citation Indexing be automated?" Essays, Vol:1, p.84

March 4, 1970

Scientific tradition requires that when a reputable scientist or technologist publishes an article, he should reference earlier articles which relate to his theme. These references are supposed to indicate those earlier researchers whose concepts, methods, apparatus, etc. inspired or were used by the author in developing his own article. Such references state the precedents for a current activity and provide the reader with bibliographic information needed to find an article if he wants to know more about some aspect of the subject.

Some scientific articles have hundreds of references; others have none at all. Most contain from ten to twenty. Part of the reason for such disparity in the number of references is the great difference in quality, not only in articles, but also in the journals that publish them. Many authors, editors, and referees are quite meticulous in ensuring that an article includes a comprehensive set of references. On the other hand, especially in the less scholarly journals, the references may be inadequate or non-existent.

As the publisher of the *Science Citation Index*®, the inconsistency in the application of the reference tradition has been a constant source of concern to me. Long before the *SCI*® became a reality, I and my colleagues were hopeful that an automatic system might be devised in which the computer could read an article and determine not only if the references provided were appropriate, but also what references were missing.

Reprinted in this issue is an article of mine discussing the possibility of achieving such an automatic referencing system. Ever since I presented this paper, I have somewhat regretted the choice of the title. Too many people have believed that the article deals primarily with the mechanics of publishing a citation index. Most emphatically, it does not. Therefore, with the benefit of hindsight, I would like to suggest that you read "Can Citation Indexing be Automated?", but with the new title "Can Criticism and Documentation of Research Papers Be Automated?"

Reprinted from: Mary Elizabeth Stevens, Vincent E. Giuliano, and Laurence B. Heilprin, Eds., *Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington 1964.* (National Bureau of Standards Miscellaneous Publication 269, December 15, 1965), pp. 189-192.

Click for Introduction on p.83, Essays, Vol:1

Can Citation Indexing Be Automated?

by
Eugene Garfield
Institute for Scientific Information
Philadelphia, Pennsylvania 19106

The main characteristics of conventional language-oriented indexing systems are itemized and compared to the characteristics of citation indexes. The advantages and disadvantages are discussed in relation to the capability of the computer automatically to simulate human critical processes reflected in the act of citation. It is shown that a considerable standardization of document presentations will be necessary and probably not achievable for many years if we are to achieve automatic referencing. On the other hand, many citations, now fortuitously or otherwise omitted, might be supplied by computer analyses of text.

This paper considers whether, by man or machine, we can simulate the process of "documenting," the process by which authors provide reference citations to pertinent and usually earlier documents. My paper does not concern the manipulative or mechanical problems of automatically compiling or printing citation indexes. The existence of the *Science Citation Index®* is adequate testimony to the ability of the computer rapidly to sort, edit, and print large-scale citation indexes.[1]

My paper also does not consider the problem of automatically recognizing (reading) and/or extracting explicit citations appearing in published documents by use of character-recognition devices. Programming such a device will require the resolution of fantastic syntactic problems even if the machine has a universal multifont reading capability. For example, in the citation, "*J. Chem. Soc.* 1964, 1963," which number is the year and which the page number? These are not trivial problems. To handle the vagaries of bibliographic syntax we "pre-edit" all documents before keypunching the citation data needed for the *Science Citation Index*. We also "post-edit" both by computer and human editing procedures. Do not confuse the "automatic" or "routine" nature of citation indexing with a syntactically

intelligent automation. Our citation indexers do not require subject-matter competence, but they do require considerable bibliographic training. The diverse and unstandardized citation practices in the world's literature make this necessary. In addition, there are linguistic variations in names and publication titles which must be handled. Our citation indexers essentially must be trained in descriptive cataloging.

My paper does concern the ability of an artificially intelligent machine to deal with, among other things, the *implicit* reference citation as distinguished from the *explicit* reference citation. Such might be the case in a paper where the author, for one reason or another, has neglected to provide a pertinent bibliography. The editor of a scientific journal would ask such an automaton to supply all "pertinent" references, if for no other reason than to make certain the research was original. Citations are generally used to provide "documentation" or support for specific statements. However, reference citations are also provided in papers for numerous reasons including, among others:

1. Paying homage to pioneers
2. Giving credit for related work (homage to peers)
3. Identifying methodology, equipment, etc.
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticizing previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of fact—physical constants, etc.
12. Identifying original publications in which an idea or concept was discussed.
13. Identifying original publication or other work describing an eponymic concept or term as, e.g., Hodgkin's Disease, Pareto's Law, Friedel-Crafts Reaction, etc.
14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage)

The problem of identifying all "pertinent" references, to support implicit citations, is a special case of the general problem of automatic indexing. It has previously been reported that machines can index or abstract by use of key words in context taken from titles,[2] by use of statistically significant sentences,[3] kernels,[4] etc. O'Connor has recently reviewed these methods,[5] as has Artandi.[6] Associative methods have been widely discussed by Stiles,[7] Maron,[8] Giuliano,[9] etc. All of these systems, however, are concerned with indexing by use of the text only. Bibliographic citations are regarded as metalinguistic elements.

Recently, however, Salton[10] has discussed the use of bibliographic citations as indicators of document content. Es-

sentially he proposes to treat citations as descriptors, which may seem strange to those who think in terms of conventional indexing. Indexers do not ordinarily think of citations (addresses of cited documents) as descriptions of the citing document. However, that does not alter the fact that they are.[11]

Citations (document addresses) are brief representations of the documents they identify. As one sacrifices compactness, such as is found in serial numbers for patents,[12] and expands to full titles and then to abstracts, one sees the gradual enlargement of the document description toward the complete text. In this transition from "citation" to "document," redundancy is introduced as well as additional information content. Indeed, a document and a citation approach equality as the depth of indexing decreases (from the full text) and the length of the citation increases. This corresponds to my earlier definition of the document as the set of descriptors which describe it.[13] In an information retrieval system, information content can be measured only on the basis of indexed information that is supplied in the indexing process. By this definition a document is a unique combination of descriptors not assigned to any other document in the collection. In most thesaurus-based collections indexing is not sufficiently deep to achieve such uniqueness. However, the combination of conventional subject headings or descriptors with the bibliographic citations used as references increases our ability to describe documents uniquely and specifically. Indeed, those who have studied citation indexes and so-called bibliographic coupling are well aware that only a small number of reference citations are needed to isolate uniquely a particular document in the collection from all others.[11] That is why a search of a citation index generally produces a highly selective and useful search result.

In discussing citation indexing it is frequently stated that weaknesses of the method include under-citation (the deliberate or unwitting failure to cite pertinent literature) and over-citation (the excesive reference to presumably nonpertinent literature). Under-citation is illustrated by the patent literature, since there is an economic motivation to cloud rather than clarify the information disclosed in a patent. However, the patent examiner, otherwise motivated, attempts to clarify the prior art by providing a list of "references cited".[14] Suppose, however, the patent examiner, or a journal editor, wishes to examine a document quite critically and asks that the "machine" provide all the pertinent documentation or prior art. This brings me once again to the main theme of my paper.

To answer the question "Can citation indexing be automated," as we have seen, obviously entails a discussion of the entire range of question-answering problems encountered in designing any information retrieval system. Consideration of the automatic procedure for supplying reference cita-

tions, when they are missing, merely focuses attention on the complex indexing task performed by the author when he does give pertinent reference citations. Such considerations help us focus attention on the significant differences between *a priori* and *a posteriori* indexing.[15] Since each person may interpret the meaning or significance of words and documents differently, the problem we are dealing with inevitably involves the human ability to create novelty, to invent, to discover, and to be critical.

Are machines, or machinelike people, capable of imitating or simulating the human process of being critical? What are the peculiarly "human" earmarks of certain sentences containing citations? When do such sentences contain implicit citations that could be supplied by an intelligent machine and when would this appear to be difficult or impossible?

Consider the following example: "Mr. X, an impossible idiot, has recently published a paper on gobbledegook. The conclusions reported in his paper are wrong as are the data on which the conclusions are based. The recommendations made by Mr. X, on the basis of his conclusions, will be a calamity for mankind."

In polite circles, this is called the critical review. Obviously, "intelligent" machines are not yet ready to generate such criticism. Or at least programmers are not yet able to program machines to prepare such critiques. If they were, then the paper by Mr. X would probably never have appeared because the same artificial intelligence would have been available to tell him that his data were wrong before he published and why! (If he persisted in publishing, we probably would have identified a quality common to humans, but invariably attributed to machines—stupidity.)

The first sentence in the example illustrates the case for an implicit citation that our machine ought to be able to provide. What could be more simple than the kernel sentence "Mr. X has published," which one would hope could be the result of a transformational analysis[4] when such methods are perfected. Such an analysis combined with a complete computer listing of the papers by Mr. X is a good starting point. Since we know that this is not sufficiently specific we must then expect of the linguistic analysis "Mr. X has published on gobbledegook" and then we have reduced the computer search to the "simple" task of identifying the one paper out of the thousands by men named X to those which concern gobbledegook. Alas, this simple task alone requires the resolution of all the linguistic and semantic problems associated with matching the word "gobbledegook" with the possibly different words in the title of the implicitly cited paper or book. Indeed, there is no reason at all to assume the same word has occurred either in the title or the text of the "cited" work. If these problems were not sufficient, keep in mind that the word "recently" is quite significant in the example chosen because it stresses

the possibility that Mr. X may have written extensively on gobbledegook and it is only one particular, or a few recent papers, that is the target for discussion.

Fortunately authors usually do provide, explicitly, the citations needed to support such sentences. As a consequence the citation index, created by human indexers, does correlate the cited work with the critical statements which appear in the second and third sentences of the example paragraph. This feature of the citation index alone would have justified its creation. However, it is interesting to speculate whether transformational or any other automatic analysis of such a paragraph could produce a useful additional "marker" which would describe briefly the kind of relationship that exists between the citing and cited documents.

These "markers" would appear in the published citation index along with the usual citation data. In the case of the paragraph above, for example, "critique" or one of several other terse statements like "Mr. X is wrong," "data spurious," "conclusions wrong," "calamity for mankind," etc., might be appropriate. The "intelligent" machine would examine a new document and generate a critical statement such as "rather poor paper." As we have seen above, a less intelligent machine might analyze the paragraph and conclude that a bibliographic citation to the work of Mr. X is missing and needed. The machine might also conclude that the cited work was under "critical" discussion because of certain syntactic or vocabulary characteristics associated with "critical." Presumably they would be identified by transformational or other sophisticated analyses not yet available. This would be no mean accomplishment. Among other nontrivial problems is the fact that the information needed to assign the marker can be spread throughout, not in a single sentence of, the source paper.

O'Connor's studies on the term "toxicity" are quite pertinent to this problem because the problems have in common the need to discover methods for assigning descriptions of documents which are subject to considerable variation.[16] What is toxic to one man may be euphoric to another!

To examine a document from the "citation" point of view, to determine what reference citations could or should be provided which link the sentence, phrase, or word in question to man's prior recorded knowledge, is to say the least a formidable challenge. The task is an excellent exercise for new journal editors. To follow the "citation" method of appraising a paper is in essence to challenge rigorously each statement in that paper. If an author does not provide documentation for statements it does not mean that they are false. However, they should ideally be supported by a "reference" to some prior document, conversation, etc.

It would appear that in the "ideally" documented paper almost every sen-

tence or phrase could be interpreted to require reference to the past. While one can accept intuitively the notion that there are novel sentences that one can express in English, novel concepts appear to be comparatively rare. Most novel combinations of words, punctuation, etc. could be transformed into concepts that had appeared before. Indeed, patent examiners like to remind inventors of this when disclosing generic concepts, alone or in combination, which anticipate specific embodiments.

I recently did an experiment with a group of my students at the University of Pennsylvania in which I asked them to read a paper published in the *Journal of Chemical Documentation*[13] which contained no bibliographic citations. The reason this paper did not have a bibliography is simple. Many published papers don't have bibliographies for similar reasons. The paper was originally presented at a meeting. The editor of the journal asked for a copy, but it was published without the bibliography which obviously was not needed in the oral presentation.

Each student was asked to supply the missing bibliography for this paper. Twelve students were involved in the experiment. One student assigned 12 references while another assigned 75. The average was about 40. This is not surprising, as a considerable amount of literature was reviewed in the paper. The bibliography could have been expanded to hundreds of items if the common German practice were adopted of giving a complete list of papers every time a topic is mentioned. Thus, in a discussion of information theory where I felt one citation was sufficient, someone else might have cited numerous related works.

The comments above are intended to give you a feeling for the problem we face in automating citation indexing. It is a wide open area of research and it will take us into every fundamental area of textual analysis—something comparable to exegesis.[17] It is apparent that each author restricts his use of reference citations according to the importance he places on the statements involved. From our knowledge of quantitative citation data, a doubling or trebling of the number of citations in the average paper would not overload the system from the user's viewpoint. The average paper that was cited in 1961 was cited about 1.5 times.[18] To double the amount of citation would not even double this figure, because not the exact same set of papers would be cited. However, even if we did significantly increase the average number of references to a particular work, we would then give consideration to a more specific approach to citations. This is well illustrated in the citations to books where one finds the list of sources subdivided by the page cited. This only adds an additional dimension in the specificity of citation indexing. There is no reason why this same principle cannot be extended to the paragraph, sentence, or word. Indeed, this is exactly what happens in exegesis.

REFERENCES

1. **Garfield E and Sher I H.** *Science Citation Index*, 2672 pp. (Institute for Scientific Information®, Philadelphia, Pa. 1963).
2. **Luhn H P.** Keyword-in-context index for technical literature (KWIC Index), ASDD Rept. RC-127 (IBM, Yorktown Heights, N.Y., Aug. 31, 1959).
3. **Luhn, H P.** The automatic creation of literature abstracts. **IBM J. Res. and Devel.** 2:159-65, 1958.
4. **Harris Z S.** Linguistic transformation for information retrieval, Proc. Intern. Conf. Sci. Inform, 1958, vol. 2, 937-950 (Natl. Acad. Sci., Washington, D.C., 1959).
5. **O'Connor J.** Mechanical indexing methods and their testing. AD #409, 276, **J. Assoc. Comp. Mach.** 11:437-49, 1964.
6. **Artandi J.** A selective bibliographic survey of automatic indexing methods, **Special Libraries** 54:630-34, 1963.
7. **Stiles H E.** The association factor in information retrieval, **J. Assoc. Comp. Mach.** 8: 271-79, 1961.
8. **Maron M E.** Automatic indexing: an experimental inquiry, **J. Assoc. Comp. Mach.** 8: 404-17, 1961.
9. **Giuliano V E.** Analog networks for word association, **IEEE Trans. Mil. Elec.** MIL-7, 221-34, 1963.
10. **Salton G.** Associative document retrieval techniques using bibliographic information, **J. Assoc. Comp. Mach.** 10:440-57, 1963.
11. **Garfield E.** The *Science Citation Index*—a new dimension in indexing. **Sci.** 144:649-54, 1964.
12. **Garfield E.** Forms for literature citations, **Sci.** 120:1030-40, 1954.
13. **Garfield E.** Information theory and other quantitative factors in code design for document card systems, **J. Chem. Doc.** 1:70-75, 1961.
14. **Garfield E.** Breaking the subject index barrier—a citation index for chemical patents, **J. Patent Office Soc.** 39:583-95, 1957.
15. **Garfield E.** Citation indexes—new paths to scientific knowledge, **Chem Bull. (Chicago)** 43(4):11-12, 1956.
16. **O'Connor J.** Mechanical indexing studies of MSD, toxicity (DDC No. not yet assigned. Contact author for copies c/o Institute for Scientific Information).
17. **Garfield E.** Citation indexes to the Old Testament. **Am. Documentation Inst. (Nov. 1955).**
18. **Garfield E.** Citation indexes in sociological and historical research, **Am. Documentation** 14:289-91, 1963.