Cleveland State Law Review

Volume 21
Issue 2 *Legal Profession Problems Symposium*

Article 6

1972

Information Science Techniques for Legal Searching

Deborah C. Goshien

Information Science Techniques For Legal Searching

*Deborah C. Goshien**

A MAJOR BREAKTHROUGH IN INFORMATION SCIENCE came in 1968 when Goffman published a description of a mathematical model for information retrieval that takes into account the real-life practice of assessing the usefulness of a new item of information on the basis of knowledge already known, and knowledge gained from the previous item of information retrieved.[1] Goffman's mathematical model uses the computer to establish automatically communication chains of documents related by their own word content. Goffman's paper demonstrates the usefulness of mathematical techniques for information science and proves that it is possible to go beyond the word-by-word searches of the entire file to newer methods which take less computer time, by requiring only parts of the file to be searched. Information scientific methods can be combined with current legal searching techniques to improve the usefulness and cost-effectiveness of computerized legal research. By combining methods from several disciplines, the lawyer-user may be enabled to locate relevant material that might be missed in either a manual or a straight word-by-word computer search.

This first study, completed in the Fall of 1971, is based on Goffman's research using mathematical models.[2] It is also based in part on Booth's demonstration that the relevant words in a document occur with an unusual frequency and can be determined automatically by computer.[3]

In 1965, Booth reported a simple, yet automatic method to derive the significant words in a text.[4] The method begins with a word frequency analysis. The computer makes a word frequency list of the entire document in descending frequency of occurrence. Then the frequency of the words is compared with a standard frequency either

---

[*Editor's Note*: This paper describes the first in a series of studies in legal information retrieval being performed at Case Western Reserve University School of Library and Information Science.]

*B.A., University of Chicago; J.D., Cleveland State University; M.S. in L.S., Case Western Reserve University; Member of the Ohio Bar; Enrolled in the Ph.D. program in Library and Information Science at Case Western Reserve University.

[1] Goffman, *An Indirect Method of Information Retrieval*, 4 INFO. STOR. & RET. 73-38 (1968).

[2] *Id.*

[3] Booth, *Characterizing Documents—A Trial of an Automatic Method*, 14 COMPUTERS & AUTOMATION 32-33 (1965); The Canadian project, QUIC/LAW, headed by Prof. Hugh Lawford, is also experimenting with complex programs employing word frequency, statistical analysis and assignment of relative weights to words contained in a document. Prof. Lawford, Project Director, Queen's University at Kingston QUIC/LAW Project: Final at 6 (Mar. 1971) and Interim at 37 (Dec. 1970), Progress Reports to the Honourable John N. Turner, Minister of Justice and Attorney-General of Canada.

[4] Booth, *supra* note 3.

known or compiled from a random sampling of the entire file. Next, a test is applied to determine unusual frequency. Instead of comparing with a standard frequency list, a word can be considered significant if it appears more than a certain arbitrarily set percentage of the time within its own text. Or Booth suggests that a "criterion of relevance" could be established by using as the standard list the words from a paper already known to be relevant.[5] The measure of the amount of difference between the word frequencies can be determined automatically by computer, and the amount of difference between the papers would indicate their degree of content relationship. Using this approach, the questioner could enter the system with a known relevant case without the need for formulating a question and without the need to continually refine his terms to find relevant documents.

If the user has no relevant case, he could enter a system such as Mead Data Central System (OBAR) using words, as is done at present. The user would continue to take advantage of presently available Boolean techniques, and could be aided by a synonym thesaurus, if one is available. The documents containing the words with which the user entered would be retrieved in the usual manner employing direct matching procedures. After the initial retrieval of documents in which the desired words appear, an automatic frequency analysis would indicate the relative frequency with which those words occurred in each document that was retrieved. The user is thus enabled to make a more useful estimate of relevance to his needs.

In addition to the words requested by the user, the computer could retrieve other words not initially requested, which occur with unusual frequency in any document containing the original word or words. These new words, initially unknown, are linked to the initial query by their significant appearance in the set of documents retrieved in response to the original query. Retrieval of such words would improve the usefulness of the response by giving an indication to the user of the entire content of a document.

The user can set up his own critical threshholds by asking for documents in which a particular word occurred a certain percentage of the time or more, and gain added information from additional words retrieved, if so requested. He can then raise or lower the parameters until a desired response is obtained. In this way, with the addition of simple word frequency counts, the user can gain more useful information from his query. The words appearing with unusual frequency will be selected automatically by simple computer techniques.

The use of complex mathematical models for extensive law searching is beyond the scope of this paper. Mention is made here

[5] *Id.* at 33.

for the purpose of introducing new, potentially useful concepts that will add new dimensions to present searching techniques, beginning with this study.

The experiment is described in two parts. Part I demonstrates the use of simple word counting techniques to determine words which occur with unusual frequency and would therefore be considered relevant descriptive terms. Part II describes the beginning of more complicated mathematical procedures which would eliminate the need for an expensive word-by-word search of the entire file in response to each query.

Part I: The Experiment

OBAR, Ohio's computerized legal research service operated by Mead Data Central System, was asked to choose twenty-five recent Ohio Supreme Court cases at random and to make a word count of each case separately. Words were listed in descending frequency of occurrence. A merge was then made of the entire file of 25 cases, listing all of the words in descending frequency of occurrence. For the purpose of this experiment, this merge became the standard list, consisting of 16,458 words. In a considerably enlarged file, the most useful standard list would be a random sample selected from all documents in the entire file, and updated at regular intervals.

"Noise words" (articles and prepositions) were already deleted by OBAR when the documents were originally put into OBAR's computer system. Further, certain common or significant words of little informative value, such as "being" "simply," "next," "none," "laws," "later," were arbitrarily excluded. Also excluded were all words appearing more than .002 times in the merge, such as "court," "v.," "appeal," "state," "case," "action," etc. Insignificant words were in fact removed at the end rather than at the beginning of the experiment in order to double-check their significance. In almost every instance, words such as "court," "defendant," "state," etc., which were predicted to be not descriptive, appeared less than five times as frequently in the individual cases than in the merge and therefore would not have appeared in the results anyway. The insignificant words would not be counted at all in an actual run. The words that are uncertain, which might be significant or not in combination with other words in the context of an individual case, would be counted and stored. These would be available for use at the discretion of the questioner, thus broadening or narrowing the parameters of the search.

For purposes of this study, a word was considered significant if it appeared five or more times as frequently in the context of its own case than in the standard list consisting of the merge of the entire file of 25 cases. This *threshhold of relevance* was determined arbitrarily, and can be raised or lowered to suit the user.

The division into percentages was carried to three decimal places and rounded off. This excluded as not significant any word appearing .004 of the time or less within the context of its own case. This threshhold could be raised or lowered by the user if the desired results are not obtained. For example, in case number 15, *In Re Estate of Duiguid*,[6] the words "probate" and "passbook" each appeared four times in 1009 words, or .004 of the words in the case. Consequently, these two words did not appear in the results. However, these two words might be useful to a user desiring a broader search. Such a user might lower the threshhold and/or carry the division to four or more decimal places.

The significant words were determined in the above-described manner for each of the 25 cases. These words were then examined and compared with the full text of each case and with a summary of each case which had been written by an attorney before the experiment began. In every instance, the words retrieved automatically by computer appeared to be descriptive of the case, and would appear to be of help to the user in assessing relevance. The results seemed to be descriptive for each case regardless of topic or length in words.

The first five cases in the file typify the results obtained for all 25 cases. These five cases concern various subjects and are of varying lengths, ranging from 179 words to 1484 words. An attorney's brief exposition of the important facts and issues in each case is given followed by a listing of the words considered significant as obtained automatically by the method described above. There were 16,458 words in the standard list. The words considered significant (appearing five times more frequently in the case than in the standard list), excluding words considered to convey no information, are listed in order of frequency as they appeared in each case.

Case 1. State v. Stephens[7]

In this criminal case, the defendant was convicted of using a false name and a false and forged prescription in attempting to purchase cocaine (HCL) from two drug stores. Defendant was silent at the time of his arrest, but later testified at his trial. In his final argument to the jury, the prosecutor implied defendant's guilty knowledge that the prescription was forged by commenting upon defendant's previously asserted silence. Defendant appealed the conviction, claiming that these direct references were prejudicially erroneous and were in violation of his constitutional rights against self-incrimination. Here, the court ordered a new trial, holding that drawing implications from his silence previously asserted is not permitted unless the defendant has waived the privilege against self-incrimination previously invoked. The court declared that the right to remain silent does not require that the accused's silence must be

[6] In Re Estate of Duiguid, 24 Ohio St. 2d 137, 265 N.E.2d 287 (1970).
[7] State v. Stephens, 24 Ohio St. 2d 76, 263 N.E.2d 773 (1970).

total from its original invocation until the jury's final verdict. Also, the court cautioned the prosecution that comments of personal opinion of defendant's guilt stated in the final argument may taint an otherwise error-free record.

Total words in Case 1: 1484. The merge is the standard list.

Significant Words	Times in Case	Freq. in Case	Times in Merge	Freq. in Merge
prosecutor	21	.014	26	.001
statement	16	.011	28	.001
silence	14	.009	14	.001
accused	13	.009	24	.001
privilege	13	.009	14	.001
comment	12	.007	15	.001
united	11	.007	23	.001
error	9	.006	20	.001
incrimination	8	.005	8	.001
waived	8	.005	10	.001
officer	8	.005	15	.001
constitutional	8	.005	23	.001
silent	7	.005	8	.001
interrogation	7	.005	8	.001
custody	7	.005	11	.001

Case 2. State v. Gribble[8]

Defendant trailer truck operator was convicted of operating an overloaded vehicle on a state highway. The conviction was overturned here because the state failed to establish its *prima facie* case in that it offered no proof of the axle spacing on the vehicle, and no proof that the scales used were properly sealed. The court also commented that venue need not be proved in express terms, provided it be established by all the facts and circumstances beyond a reasonable doubt.

Total words in Case 2: 893. The merge is the standard list.

Significant Words	Times in Case	Freq. in Case	Times in Merge	Freq. in Merge
scale	20	.022	20	.001
vehicle	19	.021	42	.002
axle	18	.020	18	.001
sealed	11	.012	12	.001
weight	9	.010	11	.001
McChesney	8	.009	8	.000
load	8	.008	8	.000
seal	8	.008	8	.000
venue	7	.008	7	.000

[8] State v. Gribble, 24 Ohio St. 2d 85, N.E.2d 904 (1970).

weighing	7	.008	7	.000
truck	6	.007	6	.000
prima	6	.007	7	.000
highway	6	.007	10	.001
patrolman	5	.006	5	.000
crime	5	.006	18	.001
axles	5	.006	5	.000
sealer	5	.006	5	.000
spacing	5	.006	5	.000

Case 3. Davenport v. Tehan[9]

Habeas corpus. Petitioner was held on two charges: aiding and abetting in the shooting of a police officer, and shooting with intent to kill when police tried to arrest him. Bail was set at $25,000 on each charge. Held: the amount of bail is within the discretion of the trial court. Here there is no indication that it is excessive or that the judge abused his discretion.

Total words in Case 3: 179. The merge is the standard list.

Significant Words	Times in Case	Freq. in Case	Times in Merge	Freq. in Merge
bail	7	.039	7	.000
habeas	4	.022	10	.001
feet	3	.017	8	.000
official	3	.017	9	.000
pound	3	.017	3	.000
accordance	3	.017	7	.000
bearing	3	.017	3	.000
customarily	3	.017	4	.000
committed	3	.017	8	.000

Case 4. State, ex. rel. Scanlan v. Court of Common Pleas[10]

Mandamus. In 1966, relator pleaded guilty to and was convicted unlawful entry of a financial institution and of shooting with intent to kill. In 1969, relator filed a post-conviction remedy petition which was dismissed. This dismissal was affirmed by the Court of Appeals. Here relator seeks to compel the Court of Common Pleas to act on a second post-conviction remedy petition which relator mailed to the judge instead of filing it in the proper court. The Supreme Court dismissed relator's request and held here that *mandamus* will not be issued because relator has shown no clear legal duty on respondent judge to act. The Ohio Supreme Court declared that merely sending

[9] Davenport v. Tehan, 24 Ohio St.2d 91, 264 N.E.2d 642 (1970).

[10] State, ex rel. Scanlan v. Court of Common Pleas, 24 Ohio St. 2d 92, 264 N.E.2d 644 (1970).

a petition to a judge does not constitute a filing nor is the judge receiving the petition under any duty to file it.

Total words in Case 4: 191. The merge is the standard list.

Significant Words	Times in Case	Freq. in Case	Times in Merge	Freq. in Merge
petition	11	.058	38	.002
mandamus	4	.021	31	.002
postconviction	4	.021	9	.000
duty	4	.021	26	.001
excessive	3	.016	3	.000
sheriff	3	.016	4	.000

Case 5. State, ex. rel. Foreman v. Court of Appeals[11]

Here relator filed a *mandamus* action to compel the Court of Appeals to set forth properly its reasons for reversal and to set forth the elements of *res judicata.* Relator had filed an action to declare a zoning ordinance invalid. The Court of Common Pleas ruled in favor of Relator. The Court of Common Pleas refused to apply the defense of *res judicata* even though it recognized its existence. Instead, the Court of Common Pleas declared the zoning ordinance to be invalid. The Court of Appeals reversed, applying the doctrine of *res judicata.* The Court of Appeals stated generally that the same issue had been decided between the parties in 1962 and that all elements of *res judicata* were present. Here the Supreme Court of Ohio affirmed the reversal by the lower appellate court, stating that the comprehensiveness of the decision in relation to the analysis of the facts or law upon which it is based is within the sound discretion of the court.

Relator challenged the Court of Appeal's refusal to certify this cause as a conflicts case. Here the Supreme Court stated that *mandamus* does not lie to review the refusal of a Court of Appeals to certify a cause as a conflicts case.

Here, the Supreme Court stated that *mandamus* is not the proper vehicle for granting a motion for separate findings of fact and law, since there is an adequate remedy at law.

The motion to dismiss was sustained and final judgment entered for respondents.

Total words in Case 5: 377. The merge is the standard list.

[11] State, ex. rel. Foreman v. Court of Appeals, 24 Ohio St. 2d 93, 264 N.E.2d 642 (1970).

Significant Words	Times in Case	Freq. in Case	Times in Merge	Freq. in Merge
mandamus	6	.016	11	.001
conflict	5	.013	8	.000
certify	5	.013	17	.001
res	5	.013	8	.000
finding	5	.013	28	.001
judicata	5	.013	8	.000
entry	4	.011	9	.000

Used in combination with the present straight word-matching techniques, this addition of significant words determined automatically could be of immediate aid to the user in assessing relevance of legal materials.

Part II: New Dimensions

In direct-matching procedures, a word or combination of words is fed to the computer, and the computer retrieves all cases containing that word or combination of words. However, in every instance the entire file is checked word-by-word, and the only words that are searched for are the ones directly named in the query by the user. It should be possible to go further and develop matches between words and patterns of words that appear in a document without compelling the user to pick and choose, thus perhaps missing a relevant document.

If a significant word appears in more than one case, the assumption is that the cases in which that word appears may be related, and that this relationship could perhaps be measured and used to determine the relevance of one case to another.

Goffman devised and tested a mathematical model which automatically separated documents into equivalence classes computed on the number of citations shared by different documents in the file. Chains of linked documents were formed and relationships computed automatically using machine methods with no necessity for subjective analysis of terms.[12]

In a more simplified fashion, using the actual words contained in a case rather than its citations, this first experiment in legal searching at C.W.R.U. S.L.S. attempted to test relationships between the 25 cases in the file.

A list was made of all of the words in the file that were considered to be significant on the basis of the test described in Part I of this paper. The words defined as significant which appear in more than one case are:

---

[12] Goffman, *supra* note 1.

Significant Word	File numbers of the cases in which the word appears
account	14, 15
board	6, 22
bank	6, 15
bar	6, 16
constitution	11, 17
code	9, 18
constitutional	1, 11, 17
crime	2, 14
county	18, 19
commission	21, 24
highway	2, 25
habea	3, 16
mandamus	4, 5, 8, 10, 20
official	3, 19
office	14, 23
postconviction	4, 11
public	17, 18, 21
robbery	11, 23
tax	17, 22
vehicle	2, 25
witness	11, 16, 23
member	13, 16
property	12, 18

Mandamus, for example, appears in five cases; 4, 5, 8, 10, 20. There are no other words shared by any of these five cases. This lack of shared words would appear to indicate that there may be very little or no factual relationship between the cases. Therefore the user would not find these cases to be significantly related to each other in spite of the fact that all five concern actions requesting relief in the form of *mandamus.* A comparison of the texts of these cases bears out the prediction that these cases are essentially unrelated:

Case 4 is an action to compel the county Court of Common Pleas to act on a second postconviction remedy petition.[13]

Case 5 is an action to require the county Court of Appeals to state its finding of *res judicata* more comprehensively.[14]

Case 8 is an action to prevent the Secretary of State from counting the votes cast for a certain candidate in an election for U. S. Senator.[15]

---

[13] State, ex. rel. Scanlan v. Court of Common Pleas, 24 Ohio St. 2d 92, 264 N.E.2d 644 (1970).

[14] State, ex. rel. Foreman v. Court of Appeals, 24 Ohio St. 2d 93, 264 N.E.2d 642 (1970).

[15] State, ex. rel. Kay v. Brown, 24 Ohio St. 2d 105, 264 N.E.2d 908 (1970).

Case 10 is an action to require certification by the county Court of Common Pleas of copies of certain proceedings against the relator.[16]

Case 20 is an action to compel the State to dismiss a criminal indictment. The writ was denied because an adequate legal remedy is available.[17]

The results are encouraging but inconclusive because of the small number of cases in the file. These results warrant further experiments with word frequencies and indicate that future research using information scientific methods for law searching may yield significant benefits.

Subsequent experiments will be done with a greatly enlarged file. Significant words appearing in more than one case will be mapped onto a matrix, linking cases on the basis of the number of words shared in relation to the total number of words in each case. It will then be possible to establish communication chains between related documents and to improve the quality of retrieval by establishing flexible threshholds of relevance and by eliminating the need to search the entire file for each query. The user may continue to enter with a word or words as is presently done, but the effective rate of retrieval or relevant material should be greatly improved at reduced cost to the user.

Future plans include a relevance study of citations. When a relevant case has been located, the significant words in the cases cited by this relevant case will be compared with the significant words in the relevant citing case. The degree of relationship will be computed statistically by taking the number of words in common and dividing by the total number of words in the citing case. A threshhold of association be set by the user and raised or lowered to fit the needs of the search.

---

[16] Bradley v. Shannon, 24 Ohio St. 2d 115, 265, N.E.2d 260 (1970).

[17] State, ex. rel. Bowling v. Court of Common Pleas, 24 Ohio St. 2d 158, 265 N.E.2d 284 (1970).