A COMPUTERIZED INTERACTIVE SYSTEM FOR THE RETRIEVAL OF CASE LAW

by

ANATOLE JARMOLYCH

S.B., M.I.T.
(1969)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF
SCIENCE
at the
MASSACHUSETTS INSTITUTE OF
TECHNOLOGY
February, 1973

Signature of Author ....
Alfred P. Sloan School of Management, January 30, 1973

Certified by ....
Thesis Supervisor

Accepted by ....
Chairman, Departmental Committee on Graduate Students

ABSTRACT

A Computerized Interactive System for the Retrieval of Case Law

by Anatole Jarmolych

Submitted to the Alfred P. Sloan School of Management on January 30, 1973, in partial fulfillment of the requirements for the degree of Master of Science in Management.

Because the judicial system of the United States operates under the common law tradition, judicial opinions arising from disputes litigated in the courts constitute a major body of law. This is known as case law. In order to advise a client as to the nature of the law and to predict how the law will be applied to resolve a controversy, an attorney must search through the collection of reported case decisions to find a precedent similar to the dispute at hand. Since there are over two and a half million reported cases at the appellate level, case law research is a difficult and time-consuming process.

The most popular tool used by attorneys in locating relevant cases is the West Digesting system. The West system, which was started over ninety years ago, consists of a number of different indices and digests which are used to trace a legal issue back to case opinions. The size and the highly interpretive nature of case law severely limit the effectiveness of the West system. Hierarchial indexing of cases by legal concept results in many kinds of errors and inefficiencies. There exists an obvious need for faster and more accurate case law research methods. This thesis investigates the application of computers to the retrieval of case law and proposes an automated case law retrieval system.

Attempts to use computers for legal research date back over fifteen years. Early experiments resulted in systems which were little more than mechanizations of the West Digesting system. These "Point-of-Law" systems contained many of the errors and deficiencies of the West system. The Key-Words-In-Combination (KWIC) approach was the first computerized case law retrieval algorithm which did not depend on hierarchial indices. By matching key search words to the full text of cases, KWIC is able to circumvent the major problems associated with manual indexing. However, KWIC's reliance on exact word matching makes the structuring of key words very difficult. Several variations on the basic KWIC algorithm have attempted to minimize this problem. The most promising experiments in this field have involved the use of probabilities to expand search word requests to include related words, and to rank retrieved cases in order of probable relevancy to the original request. This "association factor" approach uses statistical analysis of the co-occurrences of words in cases to measure the relationship between words.

The potential of the association factor for overcoming the index and language problems inherent in other manual and automated systems has led to its selection as the basis for a proposed computerized case law retrieval system. A number of improvements to the basic technique are suggested. These include the computer analysis of case text to extract informing words and extensive lawyer interaction in the computation of retrieved case relevancy. A preliminary systems hardware investigation reveals that a time-shared system consisting of a large central computer and many remote terminals located in the offices of law firms would be a technically feasible implementation of the algorithm. Although a complete analysis cannot be made until the system is designed in greater detail, informal cost-benefit calculations indicate that the benefits of such a system could more than offset the operating costs. Several implementation problems have been identified in this study. Technical problems include initial system set-up and data base generation; legal problems involve questions of copyright and unauthorized practice of law; behavioral problems center on the acceptance of the system by practicing attorneys.

The need for better case law research methods is obvious. The system proposed in this thesis appears to be technically and economically feasible. It is recommended that a thorough investigation be made of the various technical, legal, and behavioral issues raised in this study.

Thesis Advisor: Stanley M. Jacks
Title: Senior Lecturer, Sloan School of Management

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

LIST OF FIGURES

"...I could never be more convinced than now that once we have opened up the box that lets mathematicians, and other scientists examine what the lawyer does, the legal profession, legal research, the entire administration of justice will never be the same again."[1]

- John F. Horty

[1] Quoted by Reed C. Lawlor in "Computers and Automation in Law," California State Bar Journal, XL (Jan. - Feb. 1965), p. 34.

INTRODUCTION

The objective of this thesis is to examine the application of computers to the retrieval of case law. Although the study culminates with a suggested systems design, it is not meant to an optimal solution to the problem. Rather, it is a reasonable approach based on an analysis of the problem and previous efforts, which is used to determine technical and economic feasibility. It is recommended that additional research be undertaken to pursue a computerized case law retrieval system to its practical conclusion.

Chapter 1 begins by addressing the generic problem of document retrieval. A model is developed to account for the major sources of errors in a manual index, manual search system. Issues such as index rigidity, semantic noise, and user feedback are discussed. Next, the nature of case law is described. Case decisions arising from disputes litigated in the courts constitutes a major body of law in the U.S. Because of the importance of case law and the large number of reported cases, lawyers and judges need an efficient retrieval system by which they may search through the reported cases and retrieve only those cases which are relevant to the controversy at hand. The West Digesting System, which is the most popular method for performing case research, is described next. In addition to the errors and deficiencies common to all manual retrieval methods, West's system is further hampered by the size and highly interpretive and changing nature of case law. The need for faster and more accurate case research methods makes case law retrieval a likely candidate for computerization.

Chapter 2 presents a brief analysis of retrieval system performance measures. The three measures of speed, recall (completeness), and precision (accuracy) appear to be the germane parameters needed to evaluate case law retrieval systems.

The search for a suitable computer algorithm begins with an examination of previous efforts in this area. The earliest systems were little more than mechanizations of the West Digesting System and as such, embodied the errors attributable to manual indexing. The first major breakthrough was the development of the Key-Word-In-Combination approach. Since it relies on matching search words and the full text of cases, the KWIC algorithm is able to circumvent the problems associated with manual index systems. The major drawback of KWIC is that its dependence on exact match requires the careful selection of search words. Attempts to minimize the problems of exact match range from providing the user with a thesaurus of words contained in the data base to encoding the case text into a unique symbolic language. The most promising approach has resulted from experiments with statistical association. Through statistical calculations, the original search words are augmented with related words whose association factor is greater than some threshold. This expanded list is then compared with words which index the cases in the data base. Based on word matches and the value of the association factor, a document relevance number is computed by which the retrieved cases may be ranked in order of probable relevance to the original request.

Because it is not dependent on exact language and because of its potential to rank cases in order of relevancy, the association factor was chosen as the basis for the proposed case law retrieval system. A number of refinements to the basic association factor algorithm are itemized in Chapter 3. These include computer analysis of text to extract index words and interaction on the part of the user in calculating document relevance numbers. User interaction is necessary in order for the user to steer the system onto the right path.

An informal implementation analysis is presented in Chapter 4. A time-shared approach utilizing a central computer and remote terminals located in subscribing law firms is suggested for the system organization. The high cost associated with the initial generation of the system's data files appears to be the major implementation problem. Although a thorough economic investigation cannot be conducted until the system has been designed in greater detail, a preliminary cost-benefit analysis was performed in Chapter 5. Using the simplifying assumption that the only benefits which would accrue from such a system is a savings in lawyers' time, the analysis indicates that the benefits will exceed the system operating costs. Potential legal and behavioral problems associated with a computerized case law retrieval system are discussed in Chapter 6.

Chapter 1 - CASE LAW RETRIEVAL

1.1 DOCUMENT RETRIEVAL

Although the principal focus of this study is an investigation of the use of computers in the retrieval of case law[2], the concepts, evaluations, and algorithms described herein are applicable to nearly all facets of document retrieval. Similar studies and analyses may be undertaken to investigate the practical problems in applying the same algorithms to any large document collection such as an engineering library, a corporate data base, the statutory laws of a state, and the Library of Congress.

Before proceeding to a more detailed discussion of the problems involved in case law, it is appropriate at this point to analyze some of the generic problems of document retrieval. First of all, why is document retrieval a problem? Perhaps two hundred years ago when a lawyer or a physician could store all of his references in a small wheelbarrow, document retrieval was not really a problem. But today, the advancement of technology and the expansion of knowledge has brought with it a proliferation of information. It has been estimated that over half of what has been written

[2] A functional description of case law is contained in the next section.

in the history of civilization has been written in the past ten years.[3] Both the Harvard Law Library and the law section of the Library of Congress contain over one million volumes each. The law library of Harvard needs over four and a half miles of new shelf space every year for new acquisitions.[4] Professionals, be they lawyers, physicians, engineers, cannot hope to keep abreast of all of the literature being generated in their respective fields of endeavor. Out of necessity, various indexing and cataloging procedures have evolved to aid the researcher. Indexing systems, periodical indexes, card catalogs, the Dewey Decimal System, ideally should act as filters. This indexing filter should screen the entire collection of documents and allow the researcher to examine only those documents directly related to his request.

Unfortunately, manual indexing mechanisms have not kept pace with the information explosion. Actual indexing systems are a far cry from ideal filters. The common complaint heard again and again is that literature searches are too time consuming and yield either a pitifully small amount of data or so much information that it becomes unmanagable.

[3] Conlin F. H. Tapper, "Research and Legal Information By Computer," Chicago Bar Record, XLVIII (June-July, 1967), p. 227.

[4] *Ibid.*, p. 228.

The major shortcomings of manual index/document retrieval systems are as follows:

(1) the researcher does not know what he is seeking

(2) searches usually involve multiple subjects

(3) manual indexing systems are rigid

(4) indices are language oriented, and

(5) the researcher does not know when to stop searching.

The fact that a researcher is performing a document search means that he is looking for some information. If he knew what the information was, it would certainly be much easier to structure the question and the search such that a reasonably well organized indexing system would lead him to the answer. Since he doesn't know exactly what he is seeking, he must wade through the indexing system in a haphazard manner, hoping to stumble across the right combination. Suppose a literature search yielded nothing at all. This could either mean that the library contained no documents pertinent to the request or that the indexing system prevented the researcher from obtaining the desired information. Unless the entire document collection is examined, one is never quite sure which one it is.

Many searches involve multiple subjects and cross disciplines. Unfortunately, the traditional indexing systems allow the researcher to zero in on only one subject at a time.

How does one research the topic of "the economic impact of nuclear reactors in underdeveloped nations"? Does one look under economics, nuclear reactors, or underdeveloped nations? Or, all three? The time necessary to perform an adequate search is multiplied by three or four when multiple subjects are involved.

The rigidity of an indexing system is the inverse of the problem cited above. How would one classify a document on "the economic impact of nuclear reactors in underdeveloped countries" in, for example, the Dewey Decimal System? Once an indexing system is established, it remains fairly inflexible. Any new documents or articles are forced into existing pigeonholes.

Since manual indicies are typically based on subjects which are words in the English language, the occurrence of "semantic noise"[5] tends to corrupt the indicies. The term semantic noise describes all of the linguistic problems associated with indexing. This includes synonyms, spelling variations, homographs, and root variances. As an example, the word "exposure" could relate to exposure to the elements, indecent exposure, photographic exposure, or radio-active

[5] See M. E. Maron and J. L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval, "Journal of The Association for Computing Machinery, VII (1960), p. 216.

exposure. Similarly, an article on nuclear reactors could be indexed under nuclear, reactor, thermo-nuclear, electricity, power, generators, atomic, electrical production, radioactive, fission, uranium, or energy. This semantic noise hinders both the indexor and the researcher.

The last major problem in conventional document retrieval is that one never really knows when to terminate a search. At what point does a search cease? Typically, this question is answered by practical considerations of time. As more and more time is expended in the search for information, the search enters into a region of diminishing returns. This diminishing returns is both an actual and a perceived effect. The actual effect may be explained as follows: Starting with little or no knowledge on a particular topic, the first few relevant articles retrieved by a researcher in a literature search provide him with a great deal of basic knowledge concerning the topic of interest. As the search continues and additional documents are retrieved, many of these documents will repeat the same concepts presented in the first documents, but will also add a fair amount of new knowledge. Continuing the search will produce articles whose incremental contribution to the total store of knowledge on the topic is small. Thus, after a fairly exhaustive search, the researcher may find that by continuing his search, perhaps only one out of ten or one out of twenty articles found will yield some small piece of information which has not already been uncovered. This actual effect of diminishing returns, however, is only valid if the researcher has conducted his search in an optimal manner. That is, the retrieved articles were somehow ranked in order of relevancy, with the most relevant articles examined first. It is safe to say that most searches are suboptimal. Typically, the researcher will enter a region of perceived diminishing returns. After searching for information in some document library by means of some indexing system, he will reach a point where the retrieved documents are either not contributing or contributing an incremental amount to his knowledge on the searched topic. The researcher will use this result to generate an apriori prediction of the value of the information to be obtained from continuing the search (expected value). When either the library is completely examined, or more likely, when the expected value of continuing the search drops below the researcher's perception of the value of his time, the search is terminated. Since searches are seldom conducted in an optimal fashion, there is a danger that the researcher's expected value of continuing a search may be either too high or too low. If it is too low, the researcher will not obtain valuable information which could be had through a marginal expenditure of time. If it is too high, the researcher will be wasting his time.

Figure 1 illustrates a generalized manual index/ searching system and highlights the major sources of errors. Documents to be indexed are filtered through human indexors who "decide" what are the key elements and concepts. The key elements are then corrupted by semantic noise and pigeonholed into some existing rigid indexing system. In structuring a search request, the researcher is first hampered by the fact that he doesn't know exactly what it is that he is looking for. He then has to "second-guess" the indexor in deciding where to look for relevant information. The search output becomes feedback for the researcher. Based on the retrieved documents, he may intensify the search, expand the search into other areas, or terminate the search altogether.

It is clear that document retrieval systems have not kept pace with the information explosion. Manual indexing systems are inefficient, time consuming, prone to many kinds of errors, and leave one with an uneasy feeling as to when to stop searching.

1.2 CASE LAW

The importance and complexities of document retrieval become compounded when analyzed in the context of the legal profession. The lawyer, more than any other professional, is dependent on information contained in printed documents. Printed information is the very life of law. In order to

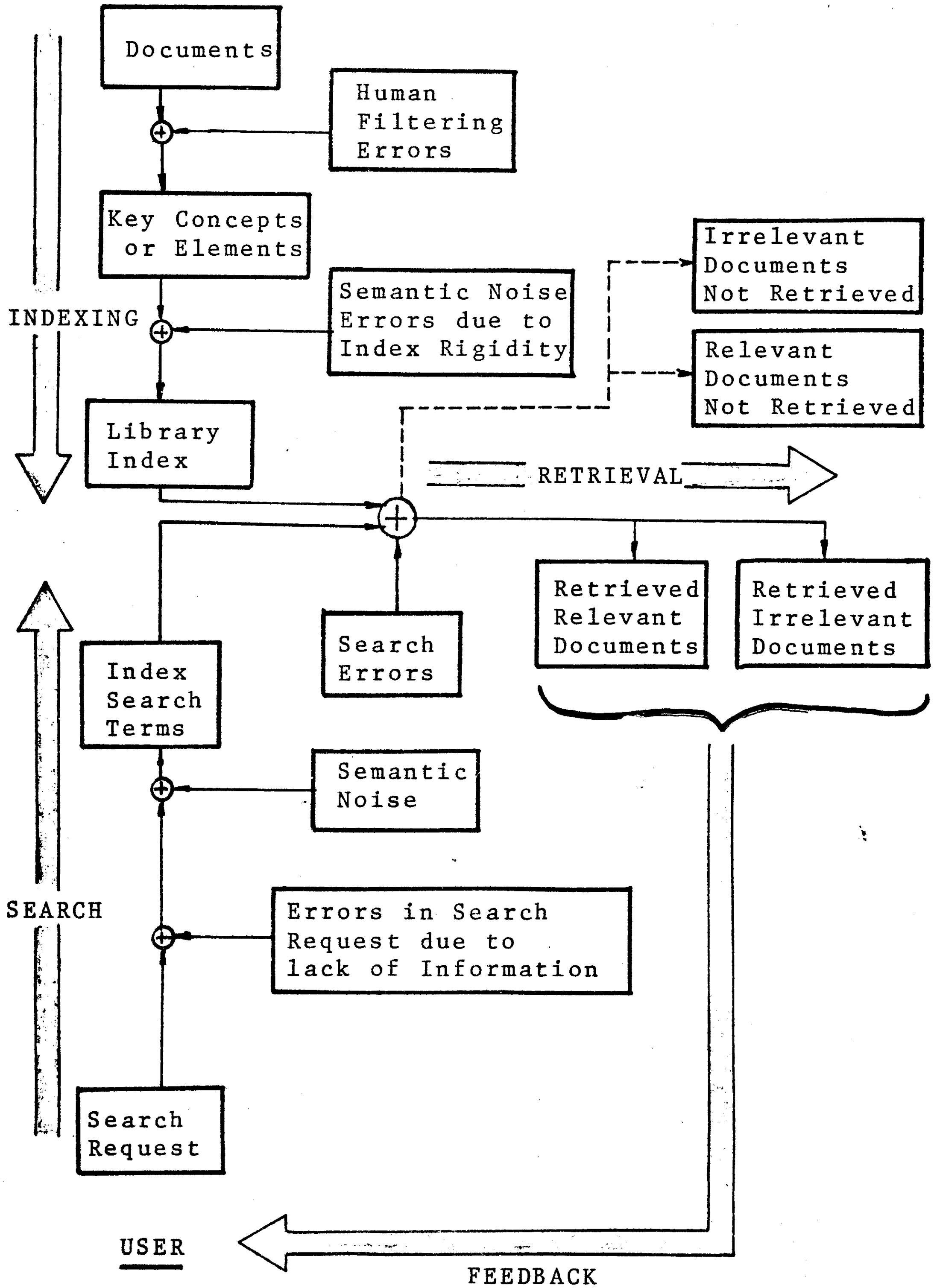

Documents
Human Filtering Errors
Key Concepts or Elements
Semantic Noise Errors due to Index Rigidity
INDEXING
Library Index
Irrelevant Documents Not Retrieved
Relevant Documents Not Retrieved
RETRIEVAL
Retrieved Relevant Documents
Retrieved Irrelevant Documents
Search Errors
Index Search Terms
Semantic Noise
SEARCH
Errors in Search Request due to lack of Information
Search Request
USER
FEEDBACK

Fig. 1 Generalized Index/Search System

appreciate the importance of legal documents, it is necessary to digress for a moment to examine the nature of case law.

Most laymen view the state and Federal legislatures as the sources of law in the United States. But since the U. S. operates under a common law system[6], the decisions of the courts of the land provide another major source of law. Court decisions arise from disputes between litigants. In settling these disputes, the courts interpret the law, decide how the law applies in the case before the bench, and sometime even "make" law. When the United States Supreme Court stated that an arresting officer must inform the person to be arrested of his constitutional rights, (Miranda v. Arizona, 1966), that became the law of the land. The judicial system serves as a dynamic, interpretive instrument of change which transformed a two-hundred-year-old constitution into the most valuable document governing a highly-technical, rapidly-changing society. However, this power to change is not taken lightly. For a society to remain stable, law must have continuity and serve as a basis for predicting the outcomes of future social interactions. The maxim of not changing law except when deemed absolutely necessary is known as "judicial restraint." In order to preserve equity and to

[6] A discussion of the differences between common law and civil law traditions may be found in most introductory textbooks on law.

make the law a viable mechanism by which society can base future actions, the courts are tradition bound to render "like" judgments in "like" cases. Once a principle of law has been established in resolving the dispute between two parties, the same principle is to be applied in future cases whose elements are essentially the same. This is knows as stare decisis. Stare Decisis is defined in Black's Law Dictionary as "to abide by, or adhere to, decided cases." The adherence to previous cases is by no means absolute. One needs only to look at the recent legal history of civil rights to see how previous case law has been overturned, expanded in scope or otherwise modified in the course of settling more recent disputes. However, it is judicial restraint which requires good cause to be shown in order to deviate from a previous decision.

The function of an attorney at law is not so much to inform a client of what the law is, but rather to use his legal training and expertise to advise a client of how the law will be applied with regard to the client's particular problem. This involves much more than just a pronouncement of the appropriate statutory laws on the books. The attorney must search through previous court decisions and find a case whose content is similar to the dispute at hand. Because of the principle of stare decisis, the attorney and his client have good reason to believe that the law will be interpreted in the same manner in the present case if the facts are substantially the same. This process is known as finding a precedent.

In a dispute between two litigants, counsel for each side will attempt to find precedents which support their client's position. The attorneys will argue before the bench that the substantive facts in the dispute are more similar to the precedents favoring their respective clients. The judge must now analyze the previous decisions, extract the general principle of law, and apply it to resolve the present dispute. In rare instances, precedents will be overturned and a new law will be created.

The decision of the court may be appealed to a higher court. The court systems of all fifty states as well as the federal courts have a hierarchial structure. They start with the lowest trial courts and go all the way up to supreme courts. The United States Supreme Court is the supreme court of the land. Its decisions and judgments are binding on all other federal and state courts. The decisions of the supreme court of each state are binding on all lower courts of that state.

Thus, the record of all past court disputes--the events which took place, the nature of the controversy, the judgment rendered, and the remedies provided--is an essential body of law in the United States. This is known as case law.

1.3 CASE LAW RETRIEVAL

In order to accurately advise a client of how the law will be applied with regard to a particular problem, the attorney must be thoroughly familiar with all of the past cases which are in some way relevant to the problem. Unless the attorney can confidently state that all relevant cases have been considered or that there are no relevant precedents, the client is susceptible of being misinformed. Action based on inaccurate or incomplete legal advice could result in the loss of large sums of money, personal liberty, happiness, etc. Courts have held that attorneys who fail to acquaint themselves with relevant precedents are breaching their duty to their clients.[7] This serves to point out the need for completeness in the retrieval of previous case decisions.

A lawyer's search for all of the relevant previous court decisions is seriously hampered by the sheer number of reported cases. Unlike many other document collections, case law is cumulative. Engineering documents on tube circuits become obsolete and are replaced with new articles on transistor circuits. Case law, on the other hand, has a virtually infinite lifetime. New cases may be added, but old ones are

[7] Joseph J. Beard, "Information Systems Application in Law," Annual Review of Information Science and Technology, Vol. 6, ed. Carlos A Cuadra (Chicago: Encyclopedia Britannica, Inc. 1971), p. 372.

not discarded. As C. H. Tapper remarked, "In how many professions can a practical problem of the present be solved by reference to a document drafted five hundred years ago."[8] Of course, not all court decisions are useful in determining the application of the law. Trial and other inferior courts usually render judgments based on factual disputes involving sufficiency of evidence and do not decide on points of law. Substantive law arising out of litigated disputes is generated by appellate and supreme courts.

The Electronic Data Retrieval Committee of the American Bar Association has estimated that there are over two and a half million reported cases on the appellate level in the United States since the seventeenth century. This number is growing at the rate of 25,000 new cases a year (nearly seventy cases a day). A lawyer cannot hope to keep track of all of the cases and decisions constantly being generated by the judicial system. Again, out of necessity, some kind of indexing system must evolve to act as an information filter. By using the indexing system, the attorney should be able to quickly sort through all of the two and a half million reported cases and examine only those cases which are relevant to the problem under consideration in order to find a legal precedent.

[8] Tapper, op. cit., p. 227.

1.4 THE WEST SYSTEM

To aid the attorney in the retrieval of case law, a number of different reporting and indexing systems have evolved. One of the oldest and certainly the most commonly used is the West Digesting System. Started by the West Publishing Company in 1879, the West system is a valiant attempt by a private concern to collect and classify all of the case decisions generated by the U. S. and state appellate courts. For lack of a better indexing mechanism, the West system has become the standard tool of case law research and the use of the West system is taught in most law schools.

Since the West system is the standard by which other case law retrieval systems will be judged, it is appropriate to describe the basic elements of the West system. Decisions of the appellate courts are published by West and bound together in volumes known as "Reporters." The U. S. is broken down into seven geographic districts (Pacific, North Western, South Western, North Eastern, Atlantic, South Eastern, and Southern). Each geographic district has its own set of reporters. For example, the appellate decisions of Florida, Alabama, Mississippi, and Louisiana may be found in the Southern Reporter. In addition to the text of the case, the Reporters will also include a brief summary of the decision and a series of headnotes which highlight particular points of law discussed in the dicision. The cases published in the Reporters appear in approximately chronological order. Grouping cases by subject or by point of law would prove to be an impossible task. Any single case might cover a large variety of subjects and legal issues. Since chronological volumes of case decisions are not especially useful to the attorney trying to uncover cases on a particular point of law, West developed the Key Number Digests to index all of the reported cases. The idea behind the key number system is that each reported case contains a number of legal issues. By identifying these issues and assigning to each a key number, an attorney could sort through a digest of legal issues and obtain a list of all of the cases in which that particular principle of law was applied. The points of law itemized in the Key Number Digests almost serves as an outline of law. In order to generate the key number system, West sectioned law into seven main divisions: Persons, Property, Contracts, Torts, Crimes, Remedies, and Government. The main divisions are broken down into 34 subdivisions which are in turn divided into over 400 digest topics. Each digest topic is further sectioned and divided into many hundred fairly narrow issues each of which is assigned a key number. As an example, references to cases involving the formation of a holding company may be found under "purchasing and holding stock in other corporations" which is under "purposes of incorporation" which is under "corporations" which is under "associated and artificial persons" which finally comes under the main division of "persons." Because of the difficulty involved in tracing a legal issue through the hierarchial structure of divisions, subdivisions, topics, and subtopics to reach a key number, West chose to publish the Descriptive Word Index. This index contains thousands of words, phrases, facts, etc., in alphabetical order to allow the lawyer to go from a description of the case directly to the key numbers. For example, under "Basketball" in the Descriptive Word Index is the subheading "injuries to boy participating in game in defendent's back yard" with the reference to key number 32(4) under Negligence (one of the over 400 digest topics).

Figure 2 illustrates the means by which new cases are entered into the West system. Each case decision is analyzed by a team of legal editors who identify the points of law contained in that case. These points are reduced to key numbers, and the case reference is posted in the Key Number Digest. The actual decision along with a summary, headnote, and key numbers is entered into the current reporter volume. The name of the case itself is added to the table of cases and certain key words may be added to the Descriptive Word Index.

The West Publishing Company advises the lawyer attempting to use the West system for the retrieval of case law to begin by dissecting the facts of the case at hand into the

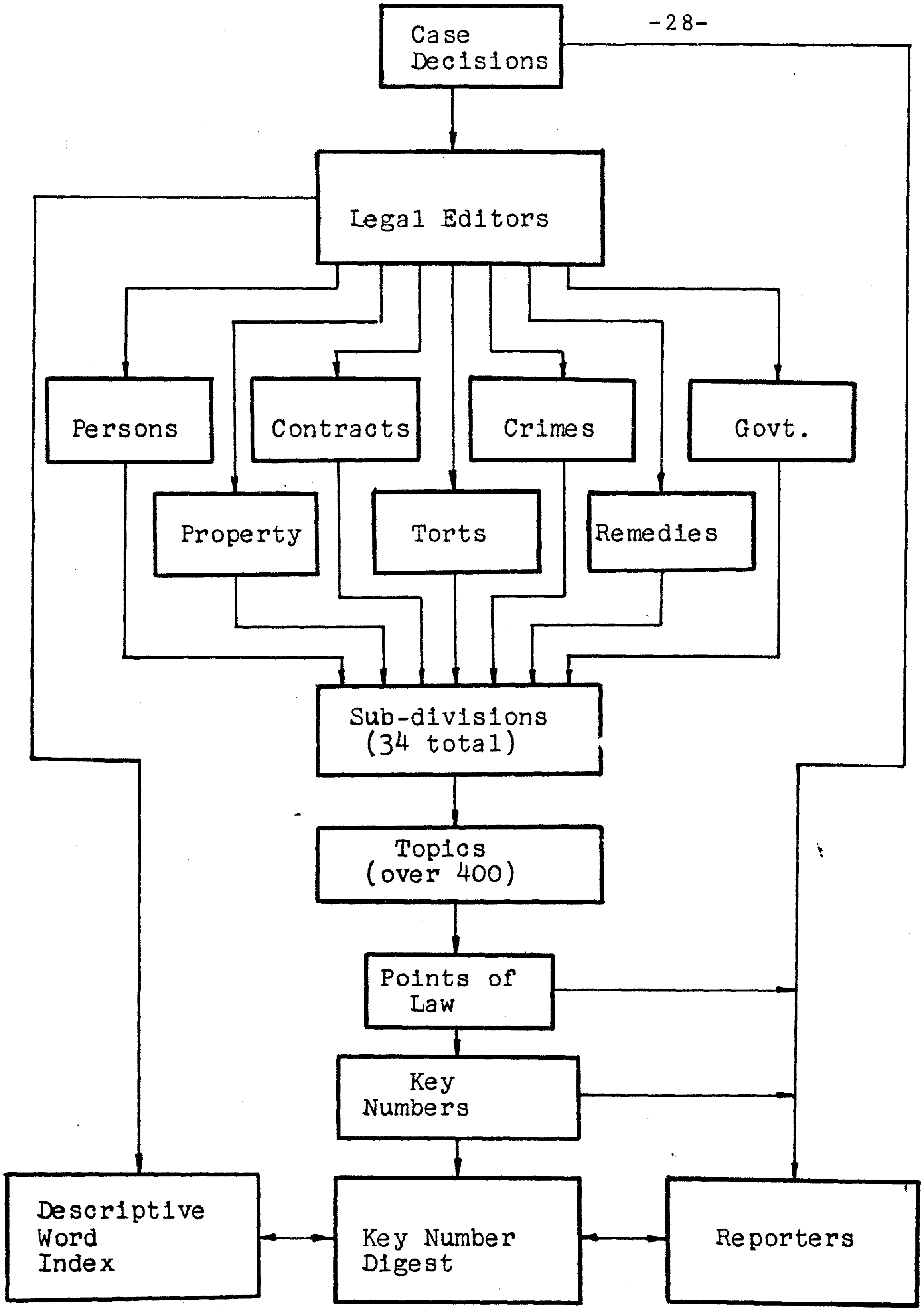


Fig. 2 West's Digesting System

following five groupings:

(1) the party or parties concerned

(2) the subject matter

(3) the cause of action or defense

(4) the object of action

(5) the points of controversy other than cause of action

These divisions will provide descriptive words which, through the Descriptive Word Index, can lead to key number references, or specific subjects by which the Key Number Digests may be referenced directly. Under each of the key numbers in the digest, the attorney may obtain citations referring to those cases in which that principle of law was applied. Armed with these citations[9], the attorney can read the actual case decision in the appropriate reporter volume. If the name of a relevant case is known, the lawyer may alternatively reference the Table of Cases to obtain the topic and key numbers of propositions of law contained in that case. The entire process is represented schematically in Figure 3.

[9] Case citations refer to the location where the case opinion may be found. For example, 275 NE 2d 33 (1971), refers to a case in volume 275 of the second series of the Northeastern Reporter, page 33, which was decided in 1971.

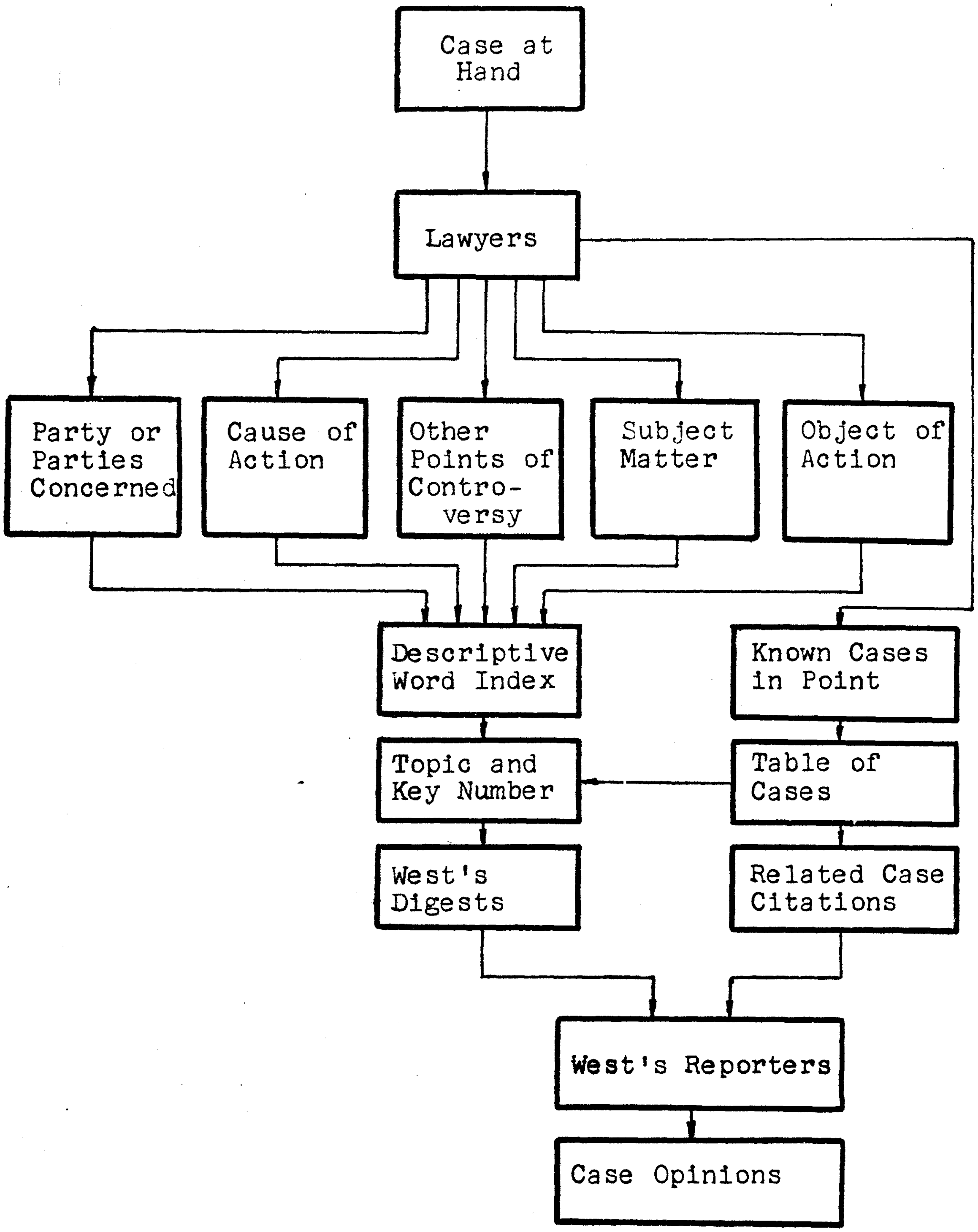


Fig. 3 Use of the West System

1.5 CRITIQUE OF THE WEST SYSTEM

Major criticisms of the West system fall into the following five categories:

(1) It is difficult and timeconsuming to use

(2) West's indexing system is rigid

(3) Cases are indexed only one time

(4) The bulk of the system makes it difficult to access, and

(5) It is expensive to use

It is apparent even from the above cursory description that using the West system for the retrieval of case law is a very difficult and timeconsuming procedure. Although West's is the best and most widely used system, it is so only by default for the lack of any other comprehensive case law retrieval mechanism. Because of the timeconsuming, frustrating, and clerical nature of using the West system to retrieve relevant cases, a senior lawyer will typically hire someone to actually perform the legal research for him. Junior members of law firms are invariably delegated this tedious task. "The hours of time which must be spent in consulting indexes, jotting down references, locating and reading them, and finally discarding most of them as 'not in point' constitute a great wast of valuable time and highly skilled brainpower."[10] Tracing a legal point through the

[10] Robert A. Wilson, "Computer Retrieval of Case Law," Southwestern Law Journal, XVI (Sept. 1962), p. 409.

winding hierarchial structure of the West system is in itself a cumbersome chore. Unfortunately, the indexing system only allows the attorney to search for one legal point at a time. Since most disputes involve a number of legal issues, the lawyer must repeat the process a number of times to obtain all of the relevant information.

The rigidities of the West system stem from the fact that once a stratified classification/indexing system has been established, it is difficult to make additions or changes without a complete reorganization. This rigidity has resulted in a number of different kinds of distortions. Because the law of our society is changing more rapidly than the classification structure, the indexing system is most accurate and reliable in the relatively stagnant areas of law and least accurate in the dynamic, changing areas. Thus, one can more readily obtain the pertinent case law with regard to "livery stable keepers" (a West digest topic) than laws regarding more recent issues such as electronic surveillance. Since the number of key number classifications remains fixed in the short run, West's legal editors are often faced with the dilemma of squeezing a case into a classification where it really does not belong. Once this happens, the case is virtually "hidden" from the attorney attempting to locate cases on a given point of law. In West's Fifth Decennial Digest (spanning the years 1937 thru 1946) all cases regarding

Social Security were indexed by key number 78.2 under the digest topic of "Master and Servant."[11] In the Sixth Decennial Digest (1947 thru 1956), the new digest topic of "Social Security" was introduced with 751 assigned key numbers. This process clearly adds to the dilemma of the attorney performing legal research across the time frame of the two digests. Professor Irving Kayton points out even a more serious flaw resulting from the rigidity of the West system.[12] New points of law brought out in cases for which there exists no classification in the West system may simply not be indexed at all. As an example, Kayton cites Section 103 of the Patent Act of 1952. This section describes "non-obviousness" as a condition for patentability. Since the passage of this act, several cases have come before the courts in which this condition of patentability was discussed at great length. In Kayton's words,

> Despite all this case law and other significant legal literature not to mention the fact of the passage of the statute itself, no legal index includes the terms "non-obvious" or "obvious" in or near the hierarchial generic headings of "patentability" and "invention," or anywhere else. The single most significant legal issue in this field of law for the past fourteen years

[11] Ibid., p. 411

[12] Irving Kayton, "Retrieving Case Law by Computers: Fact, Fiction, and Future," George Washington Law Review, XXXV (October, 1966), p. 1.

> has not succeeded in breaking the mold or pattern of the preconceived legal index structures...[13]

Another major problem with the West system is that cases are digested and indexed only once. There is nothing wrong with this as long as the meaning of the cases does not change with time. But change is the very hallmark of our judicial system. With the passage of time, previous case dicisions are reinterpreted to take on completely different meanings. Unless the legal editors can achieve the unattainable goal of extracting all of the significant points of law from a case for all time, some salient issues will be lost forever.

The sheer size and bulk of the West system make it inaccessable and unmanageable. West's National Reporter System contains well over four thousand large volumes. Only a few of the largest law firms in the nation can afford to maintain a complete West system. Many law firms cannot even maintain a complete set of Reporters for their own geographic region. As a result, lawyers must commute to large law libraries in major cities in order to perform extensive legal research. This, of course, involves a great deal of wasted time and money.

lastly, as with most document retrieval systems, the lawyer is faced with the problem of not knowing when to stop

[13] Ibid.

searching. He is never sure that he has seen all of the relevant data that he ought to have seen. Unlike the general researcher who mentally formulates the expected value from continuing a search, the attorney solves this perpetual dilemma by billing his time to the client. The quality and intensity of the case law research thus becomes a function of how much the client is willing to spend. (Clients are hardly aware that they are spending "top dollar" for the lawyer to perform what is essentially clerical gymnastics with an archaic indexing system.) This of course results in very complete and thorough legal research in five-million-dollar anti-trust suits and a passing examination of the law in a five-thousand-dollar personal injury claim. Certainly, one is struck by the resulting unfairness of the fact that access to a major body of law in this nation is hampered by an economic barrier due to the lack of an adequate document retrieval mechanism.

Perhaps the remarks made by Robert A. Wilson over a decade ago best serve to summarize the current state of legal research.

> ...the present day legal research picture consists of legal problems that are becoming more and more complex and yet in need of more rapid answers; an unwieldy accumulation of cases and statutes inherited from the past; a great yearly outpouring of new materials which must be added to the present accumulation; and indexing systems which are no longer precise enough to give access to pertinent

> precedents with sufficient speed or accuracy. As a consequence, legal research in important cases is unnecessarily slow and expensive. It results in delays in litigation, frustration for clients, and an inordinate expenditure of time and money on the part of lawyers. Thus, the time has arrived to look into the capabilities of modern scientific instruments, such as the electronic computers, to see if they can assume some of the research burden.[14]

1.6 POSSIBILITIES OF AUTOMATION

Although Wilson made reference to the electronic computer as a "modern scientific instrument," the past ten years have witnessed the transformation of computers from laboratory instruments to everyday business machines affecting all of us. The range of computer activities runs the gamut from relatively simple operations such as processing weekly payrolls and reserving seats on airline flights to very complex functions such as analyzing seismic waves in oil exploration or controlling the missile guidance system for a lunar landing. The wizardry of the electronic computer lies not in the premise that it is "smarter" than a man. On the contrary, a computer is quite "dumb." It does exactly what it is told to do, it can only do one thing at a time, and the set of operations it can perform is rather limited. The real power of the computer is that it is exceedingly fast. Present digital computers can add two large numbers in a matter of nano-seconds (10^{-9} seconds). To appreciate this speed. the following observation is made: a nano-second

[14] Wilson, <u>op. cit.</u>, p. 412.

is to a second as a second is to thirty years. Thus, the computer can perform millions of operations in a second. Computers are most profitably applied to problems in which very large amounts of data must be processed as rapidly as possible. As Wilson and other researchers point out, the retrieval of case law is just such a problem. Literally millions of case decisions must be "processed" in order to extract those few cases which are relevant to a dispute under consideration.

Before one can proceed to automate any kind of system, the following issues must be examined:

(1) The establishment of output measures

(2) The formation of an algorithm

(3) Practical implementation of an algorithm, and

(4) Economic analysis of the solution

The establishment of output measures is another way of saying "what is it that we are trying to improve?" There needs to be some kind of measurement methodology which allows one to determine whether in fact the automated system did or did not solve the problem it was designed to solve. The formation of an algorithm is just the formularization of the method by which the computer will perform the job. Just because the job was performed in a certain way manually does not necessarily mean that the computer will have to go through the same procedure. Because of its "stupidity" and its incredible speed, the computer might be programmed to perform a job in what would be a manually inefficient manner--and yet it might do the job a thousand times faster and with greater accuracy than a man. These algorithms must then be incorporated into a practical hardware design to assure its technical feasibility. Lastly, an economic analysis must be performed on the design of the computerized solution to determine whether the benefits of an automated system exceed its costs.

The remainder of this study focuses on the above four issues with regard to the automatic retrieval of case law.

Chapter 2 MEASURES OF PERFORMANCE

2.1 VENN DIAGRAMS

In asking the question of how an automated system can improve case law retrieval as is currently being performed with the West system, the answer which immediately comes to mind is that an automated system should be faster and provide more relevant cases than a manual system. In comparing the productivity (as measured by the number of relevant cases retrieved) of two systems, the first approach taken was to make use of Venn diagrams (see Figure 4). The number of relevant cases retrieved by each of two different systems is depicted by a circle. The larger the number of relevant cases retrieved, the larger the circle. The intersection of the two circles represents that set of cases found by both systems. If in comparing a manual and an automated system, Venn diagram "A" of Figure 4 is produced, one may state that the automated search is clearly superior to the manual search. It not only retrieved all of the relevant cases found by the man, but it also retrieved a number of cases which were not uncovered in the manual search. If Venn diagram "B" is produced, one might argue that an automated system might be useful in supplementing a manual search. It found a significant number of cases which would have been otherwise missed; however, it was not able to retrieve most of the cases found

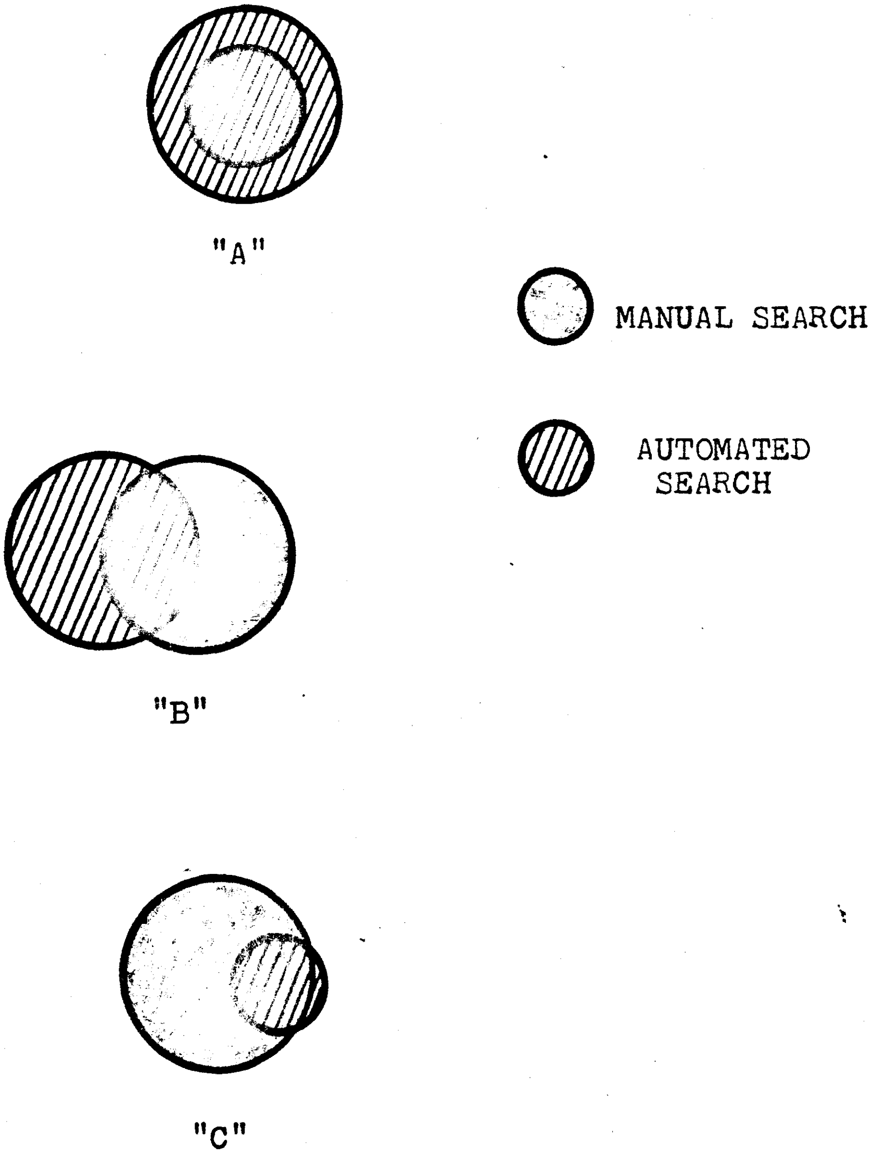


Fig. 4 Venn Diagrams

by the man. Diagram "C" depicts a situation in which the automated search did not contribute significantly to the retrieval of relevant cases. For such a system, one would have to question its cost-effectiveness given that only a fraction of the relevant cases found by the man were retrieved automatically.

Despite their intuitive appeal and visual simplicity, Venn diagrams have several serious drawbacks when used to quantify productivity of document retrieval systems. First of all, there is no way to represent the "intensity" of the manual search. A lawyer arguing a routine case in district court might not conduct the same intensive and exhaustive search as a lawyer arguing in front of the United States Supreme Court. Secondly, there is no way to determine how many relevant cases were missed by both the manual and automated search (i.e., a measure of completeness). Lastly, there is no indication of how many irrelevant cases were retrieved in order to obtain that subset of relevant cases. An extreme example of this is the following: suppose in searching some point of law, an automated system retrieved all two and a half million reported cases. A Venn diagram would show that the machine obtained all of the relevant cases found manually and probably some relevant cases missed by the researcher. Such a system would clearly be worthless even though it would have a favorable Venn diagram.

2.2 RECALL, PRECISION

In order to overcome some of the difficulties involved in using Venn diagrams, alternative measures of productivity have been developed.[15] One standard measure is "recall." Recall is a measure of completeness. It is expressed mathematically as the ratio of relevant cases retrieved to the total number of relevant cases in the collection. However, this measure must be complemented with an expression for the accuracy of the system. This measure is called "precision." Precision is the ratio of relevant cases retrieved to the total number of cases retrieved.

$$\text{RECALL} = 100 \left(\frac{R}{C}\right)$$

$$\text{PRECISION} = 100 \left(\frac{R}{L}\right)$$

C is the number of relevant documents in the collection
R is the number of documents in "C" found by the search
L is the total number of documents retrieved in the search

An ideal system would produce 100% recall and 100% precision (i.e., all of the relevant cases would be found with no false drops on irrelevant cases). Researchers in the field of information systems found that the ideal is seldom realized and that typically, there is a trade-off between recall and precision. A convenient way to represent the two

[15] For an excellent analysis of productivity measures, see S. E. Robertson, "The Parametric Description of Retrieval Tests," Journal of Documentation, XXV (March, 1969), p. 1.

measures is to plot recall on one axis and precision on another axis on a graph (R-P curves). A typical R-P curve for a manual search is shown in Figure 5. This curve can be explained in the following way. Suppose a lawyer wants to retrieve all cases pertinent to some point of law. He may remember a few important citations and immediately retrieve several cases. Since those cases would be directly related to the point of law, he would have 100% precision. However, these few cases would represent only a small fraction of the total cases concerning that particular subject; therefore, the recall is low. In trying to uncover more relevant cases, the lawyer would begin a general library search using an index system such as West's. As he uncovers more and more relevant cases, he has to wade through an increasing number of irrelevant cases. Thus, the recall is increased at the expense of precision.

A quantitative comparison of automated retrieval systems and a manual search can be made by plotting recall/precision curves. The "best" system would have an R-P curve which is closest to the ideal of 100% recall and 100% precision. An example of two curves is shown in Figure 6. Curve AB is clearly superior to curve CD. Notice that at any point on the CD curve such as (R,P), one can more to the AB curve and obtain either better recall at the same precision (R',P), better precision at the same recall (R,P'), or a combination

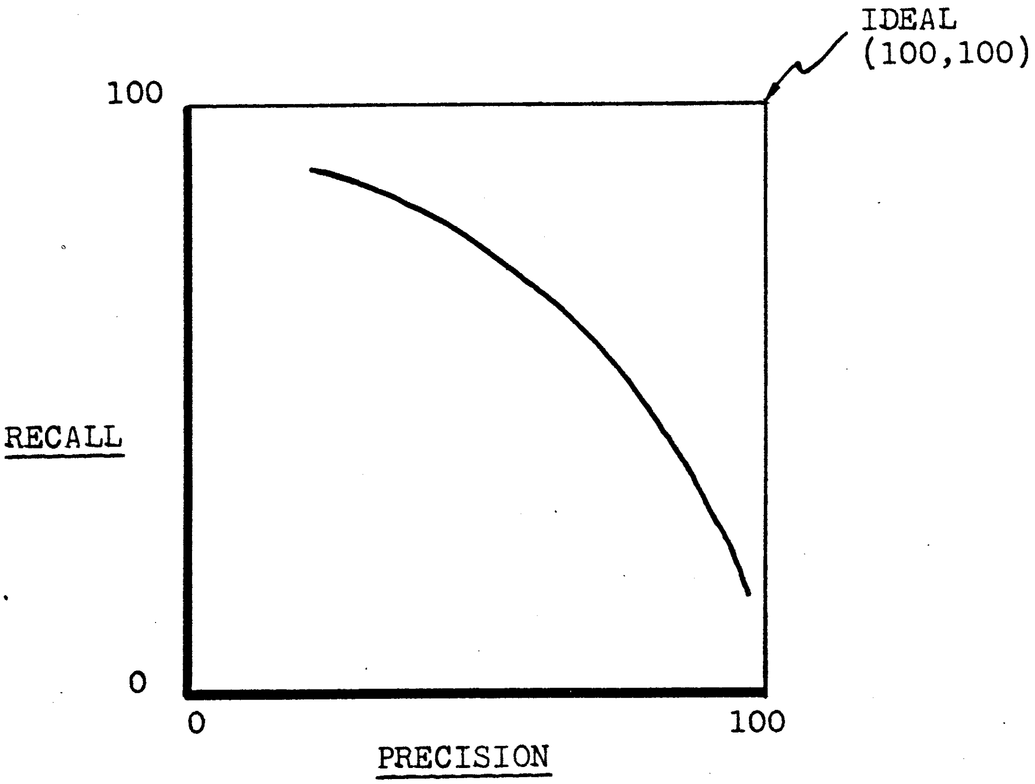


Fig. 5 Typical Recall-Precision Curve

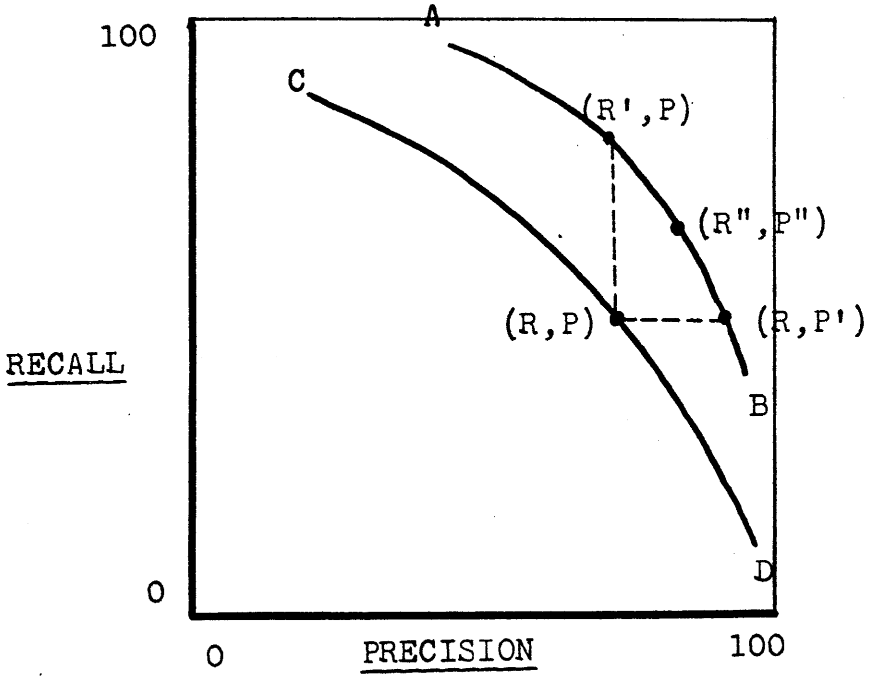


Figure 6 R-P Curves

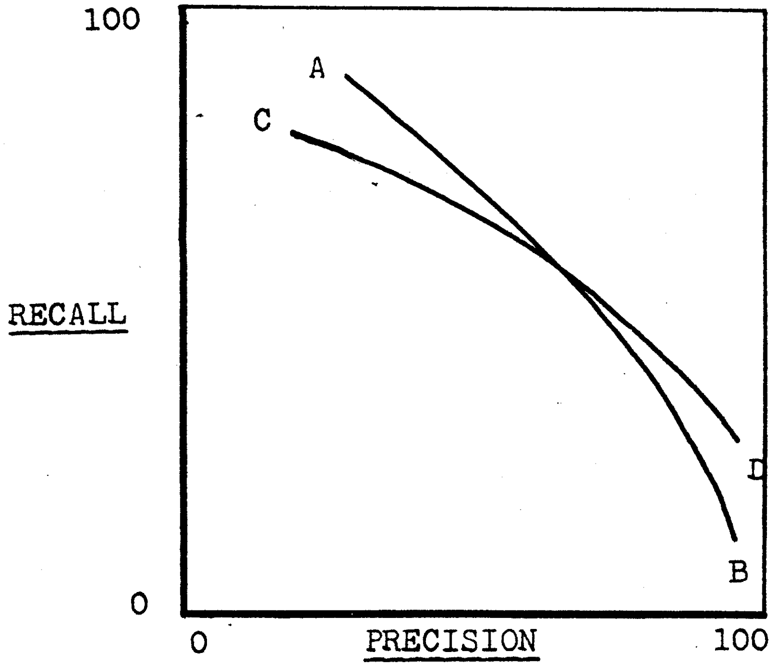


Figure 7 R-P Curves

of both (R",P"). Figure 7 illustrates a situation in which neither curve is absolutely superior to the other. If the researcher desires maximum recall, then curve AB is better than curve CD. However, if a lower level of recall is sufficient, then curve CD would provide that level of recall with greater precision. Thus, in this situation the "best" curve becomes a function of the researcher's operating point with regard to recall and precision.

The last measure of performance is speed of the system. Most certainly, an automated case law retrieval system would perform legal research many times faster than a man. If an automated system produced better recall and precision, then this coupled with the increased speed would be sufficient evidence to categorically state that the automated case law retrieval system is superior to manual search. If, however, the automated system generated either less recall, less precision, or both, one would have to weigh this degradation against the savings in time to determine which system is better for that particular application.

Chapter 3 FORMATION OF AN ALGORITHM

3.1 DOCUMENT VERSUS INFORMATION RETRIEVAL

What is it that the lawyer is seeking through case law research? The actual case decision is not the end. Rather, the case decision is means by which the lawyer may obtain pertinent legal information in order to advise his client. One might argue that what is needed is not an automated document retrieval system, but an automated information retrieval system which would lead the attorney directly to the relevant issues under consideration. Layman E. Allen puts forth the following propositions:

> IF 1. the written materials used in the tax field are more systematically drafted
>
> THEN 2. human beings will be able to "read" and "work with" those materials "better", and
>
> 3. automated devices will be able to "read" and "work with" those materials "better."[16]

Allen proposes that a systematic organization be adopted in the drafting of all legal documents. By organizing legal material into precise sentence patterns and statement structures, researchers (be they human or "machines") could rapidly access these materials and quickly extract the exact information desired.

[16] Layman E. Allen, "Beyond Document Retrieval Toward Information Retrieval," Minnesota Law Review, XLVII (April, 1963), p. 714.

Allen's comments certainly have a great deal of merit, especially when applied to statutory law and administrative codes. However, they fall short of their mark when viewed in the context of case law. Case decisions are not just factual statements of the issue and the judgment rendered. Judicial opinions tend to be literary works of art filled with analogies and metaphors. As Goldblum points out, "...some readers of cases are surprised at the volume of folk wisdom, general philosophy, elaborate rationalizations, and out-and-out hot air with which many opinions seem to be filled..."[17] Literary exposition on the parts of court judges is not without its reasons. A judge presiding over a dispute before the bench is fully aware that the ruling he will render will be used by other courts in future litigations of a similar nature. Because he cannot possibly be aware of all of the future consequences of issuing some point of law, many judges will render a decision to specifically settle the dispute at hand and cloud the general principle in a factual description of the case. Thus, the actual general principle of law applied in the case is left "flexible." Typically, the principle of law does not become completely evident until judges in future cases reflect on previous cases and make judgments as to what facts were

[17] Edward J. Goldblum, "Application of Computers to Retrieval of Case Law," (Unpublished Master's thesis, Sloan School of Management, M.I.T., 1968), p. 11.

sufficient and necessary for that particular decision. Herein lies the distinction between holding and dictum. The holding of a case is the factual situation which prompted the rendered judgment. Dictum is all other facts and information which, although pertinent to that particular case, were not actually necessary in order for the court to reach a decision.

Systematically organizing the structure of case opinions as Allen suggests is contrary to the operation of the judiciary. It would place a heavy burden on judges who would have to structure general principles of law without the buffering effect of future reinterpretation of holding and dictum. Thus, it would be infeasible to design an automated case law information retrieval system. The best that one can hope for is an automated document retrieval system which will present to the attorney the actual relevant case decisions in the words of the judge. From there, it is up to the attorney to use his own skill, training, and legal expertise to extract the general proposition of law and predict how it will apply to his client's problem

3.2 ANALYSIS OF PREVIOUS SYSTEMS[18]

Efforts to use computers to retrieve case law date back over fifteen years. Some attempts were nothing more

[18] General material for this section was gathered from a large number of survey articles listed in the bibliography, and as such will not be footnoted except to reference specific information.

than small experiments which have long since been abandoned, while other attempts have yielded operational systems which are in use today. By analyzing the algorithms, implementation, and results of these previous and current efforts, it is hoped to determine whether these systems, in fact, satisfy the lawyers' requirements of increased speed, recall and precision in the retrieval of case law. If not, the analysis will provide guidance for the design of an alternative document retrieval system.

3.2.1 Manual Index, Automation Retrieval

The earliest documented experiment of applying computers to case law retrieval was undertaken by the late Professor Robert T. Morgan of Oklahoma State University in 1957. The technique employed, which has become known as the "Point-of-Law" approach, has been described by many as nothing more than an automation of the West system. Legal researchers analyze each case and extract the pertinent legal issues contained therein. Each legal concept is assigned a unique numeric code, much in the same way West uses key numbers. The citations, titles, headnotes, numeric codes, and other relevant data for each case analyzed is stored into the computer data bank. Once a sizeable collection of cases has been digested, an alphabetic listing of all of the legal concepts with their corresponding numeric codes is generated and published. This listing has been likened to a telephone directory of points of law. A lawyer or researcher can thumb through this listing and obtain a set of numeric codes pertinent to his problem. Once these codes are entered into the system, the computer can scan its data base to identify and output those cases which are indexed under the requested numeric codes.

With its first public showing in December of 1960, the "Point-of-Law" system demonstrated without a doubt that computers can be used to relieve some of the burden of performing case law research. Subsequent critics of Morgan's system often fail to appreciate the fact that this effort marked the first time that state of the art technology was applied to legal research--an effort which created widespread public attention and which sparked the imaginations of numerout scientists, engineers, and attorneys.

Although "Point-of-Law" research at Oklahoma State University was cut short by the untimely death of Professor Morgan in 1962, the technique was adopted and used in two other systems. The Federal Trade Commission's "Concepts of Decision" system and a private commercial system called "Law Research Services, Inc." both rely on the "Point-of-Law" algorithm.

In the FTC system, decisions of the commission itself, circuit courts, and the U.S. Supreme Court are digested and the principal "concepts of decisions" are extracted. As with the Morgan system, each concept is assigned a unique numeric code. Researchers specify the code numbers of interest and the computer will search its data base and print out the relevant citations. No published results of the system's performance are available.

Law Research Services, Inc. was founded in 1964 by Ellias C. Hoppenfeld. Officially, the operation of Law research Services, Inc. is a "trade secret." However, most observers believe that the LRS system is based on the "Point-of-Law" technique. The promotional literature of LRS, Inc. claims that over one million case abstracts are stored in the system's data base. An attorney may perform case research on the LRS system in one of two ways. First, LRS provides a set of legal glossaries or "thesauri." Each of the nine thesauri covers one of the following fields of law: Corporations; Contracts and Business Law; Criminal Law; Domestic Relations; Estates and Wills; Evidence and Procedure; Negligence; Public Law and State Taxation; and Real and Personal Property. Beside each term or descriptor in the thesauri is a ten-digit identification number. An attorney will use the thesauri to locate terms which describe the legal problem he is researching. After the corresponding term numbers are entered into the system via a Western Union Telex terminal, a computer (Univac III) will search its data base and print out up to ten relevant case citations, along with an indication of whether additional relevant citations are contained in the data base. (These additional citations may be obtained for an additional fee.) Citations are presented in reverse chronological order beginning with the decisions of the highest courts. Robins points out that the attorney "...would get a very old Supreme Court case before yesterday's court of appeals case."[19] Full text printout of the case decisions may be ordered through the remote terminal, but are not remotely outputted. This has led observers to believe that the text itself is not stored in the computer. Apparently, requested decisions are typed by the LRS staff from West's Reporters and mailed to the attorney.

As an alternative to using the computer-generated thesauri, LRS offers a "Special Evaluation Query" form. An attorney may fill out one of these forms by specifying the exact legal research question and itemizing the factual content of the problem. Once this "Special Evaluation Query" is submitted to LRS, a trained staff member presumably analyzes the questionnaire and translates its contents into the machine readable numeric codes in order to obtain the computer-generated citations. The cost of a "Special Evaluation Query" is, of course, greater than a search based on descriptor codes supplied by the attorney.

[19] W. Ronald Robins, "Automated Legal Information Retrieval," <u>Houston Law Review</u>, V (March, 1968), p. 691.

Because it is a private, profit-seeking concern, it is difficult to obtain good, scientific data with regard to measures of productivity of the LRS system. The only information that is available are the highly optimistic claims put forth in LRS' promotional literature. These claims are severely cast into doubt inasmuch as there has been a legal suit filed against LRS arising from the "lack of correlation between legal questions submitted and the case citations produced."[20]

Even without hard evidence as to speed, recall, and precision, certain general evaluation statements can be made about the three "Point-of-Law" systems described above. Although the "Point-of-Law" technique is very similar to the West system, these early attempts at computerizing case law retrieval appear to have a number of significant operational advantages. Besides being much faster than manual search, "Point-of-Law" systems enable the attorney to search for several different legal concepts at the same time. Also, the user may specify the type of output to be generated by the computer. He may, for example, request that only the title and citation be printed, or title, citation, and headnotes. The speed and convenience of these systems make them "better" in many respects than manual search. However, these systems did not go far enough in restructuring the methodology

[20] Beard, op. cit., p. 377.

of case law search and retrieval. The basic "Point-of-Law" depends on the human extraction of legal issues from case opinions and the classification of these legal issues into some fairly rigid, hierarchial indexing system. This is exactly what West does with its key number system. As such, all of the criticisms of West with regard to human digesting errors, rigidities of index classification, etc. are equally applicable to the "Point-of-Law" systems. As with West, cases in "Point-of-Law" systems digested only once (even though their meaning might change with time) and legal issues are pigeonholed into some limited set of numeric descriptors.

3.2.2 KWIC Systems

"Point-of-Law" systems have served to "break the ice", so to speak, in applying computer technology to legal research, but have not solved the single-most important drawback of the West system which is rigid hierarchial indexing based on legal issues. In order to overcome the problems of indexing and human extraction of legal concepts, a large number of "key-word-systems" have been developed. The technical approach of most key word systems can be described as follows: the entire text of all of the documents in a collection is stored in a computer data bank. A "vocabulary" file is then internally created specifying all of the different words used in the document collection and referencing for each word all of the documents which contain that word. The user structures his request by entering a number of key words which describe his legal problem. The system returns by outputting citations to those documents which contain the desired key words.

John F. Horty, the then Director of University of Pittsburgh's Health Law Center, pioneered the use of key word systems in the retrieval of law in 1959. His KWIC--Key Words In Combination--system has served as the basis for dozens of subsequent law retrieval systems and references to Horty and KWIC are found in literally all publications concerned with computerized law research.

Horty's early experiments were centered around computer retrieval of Pennsylvania statutory law in the field of health. Later efforts expanded the system to include all of the Pennsylvania statutes and some case law. Statutory law in a relatively narrow area was initially chosen for two reasons. First of all, the vocabulary used in statutory law is rather precise and to the point. One doesn't find the literary metaphors and analogies contained in case opinions. It was thought that if the Key-Words-In-Combination approach were to work at all, it would work on a data base which contained fairly exacting language. Secondly, each statutory section (which was considered to be a separate document) was relatively short (typically less than 500 words). One of the major problems with storing full text is that it requires a great deal of computer memory. By confining the data base to statutory law, a large number of "documents" may be stored without an enormous data bank.

The operation of Horty's system begins with the input of documents. Keypunch operators transcribe the text of the statutory laws to be used in the data base onto punched cards. These cards are entered into an IBM 7070 computer which transfers the contents of the cards onto magnetic tapes positioned on the ten computer tape drives in the system. Once all of the text has been entered, the system can begin to construct a vocabulary file. The vocabulary file consists of a list of each of the different words used in the data base text with code numbers which indicate where in the collection of documents these terms are found. The 112 most commonly used words (such as "a," "the," "of," "but," "this," "and," etc.) which account for roughly 40% of normal text but contain very little informational value are not included in this file. (The Pennsylvania statutes contained approximately 15,000 different words). The four-part "word locator" number contains the document number of the document in which the word appears, the sentence number, word position in the sentence, and type of sentence. Beside each word in the vocabulary file is a "word locator" number for each time that word is used in the collection of documents. A researcher wishing to use Horty's KWIC system begins by itemizing key words which describe his problem. These key words may be combined (hence the name, Key-Words-In-Combination) through use of the logical (Boolean) operators of "and" and "or." An "and" requires that all terms controlled by the operator be present in the document in order for that document to be retrieved while an "or" operator requires that any of the terms be present. As an example, a request for documents on the subject of child beating might be structured in the following way:

[(FATHER) or (MOTHER) or (PARENT)] AND [(BEAT) or
(INJURE)] AND [(CHILD) or (SON) or (DAUGHTER)]

The computer will now search through the vocabulary file and identify those documents which contain the desired combination of key words.

Early experimentation showed that the KWIC system had a great deal of promise, but it also uncovered some operational problems as well. Perhaps the most serious problem involved the specification of the key word search terms. Unless the exact words in their exact form were contained in both the request and in the text, the document would not be retrieved. In the simple child beating example cited above, the concept of child beating can be expressed in eighteen different ways using various key word combinations. Yet there are hundreds of other ways to specify the same concept. Child could be substituted with offspring, or juvenile, of foundling, or baby. The same is true of parents and beat. Furthermore, all of the grammatical variations must be taken into account such as children, babies, and parents. Another major problem was the retrieval of documents which, although containing all of the desired key words, were not relevant to the user's request. The example request cited above could result in the retrieval of a document describing parents and children being injured in automobile accidents.

To overcome some of these problems, a number of procedural and operational changes were made to the KWIC system. To aid the user in specifying his key words, a thesaurus containing all of the different words in the document collection was generated. The thesaurus allows the researcher to frame his request with key words actually used in the data base. To minimize the retrieval of irrelevant documents, a two-fold approach was taken. First of all, the user was given the ability to specify the "nearness" of the key words within the documents. For example, the user could request that two key words appear in the same sentence, or within so many words, or one following the other. Although this would certainly reduce the number of irrelevant documents, a too restrictive key word structure would result in the missing of relevant documents (hence a tradeoff between recall and precision). The second approach involved the generation of a Key-Word-in-Context output. Key-Word-in-Context (which is also known as KWIC and is often confused with Key-Words-in-Combination) is a line of text (usually 70 characters) in which the key word is embedded. An example of such an output is illustrated in Figure 8.[21] This kind of KWIC output does not actually prevent false drops, but rather allows the user to quickly determine in many cases whether the document is irrelevant.

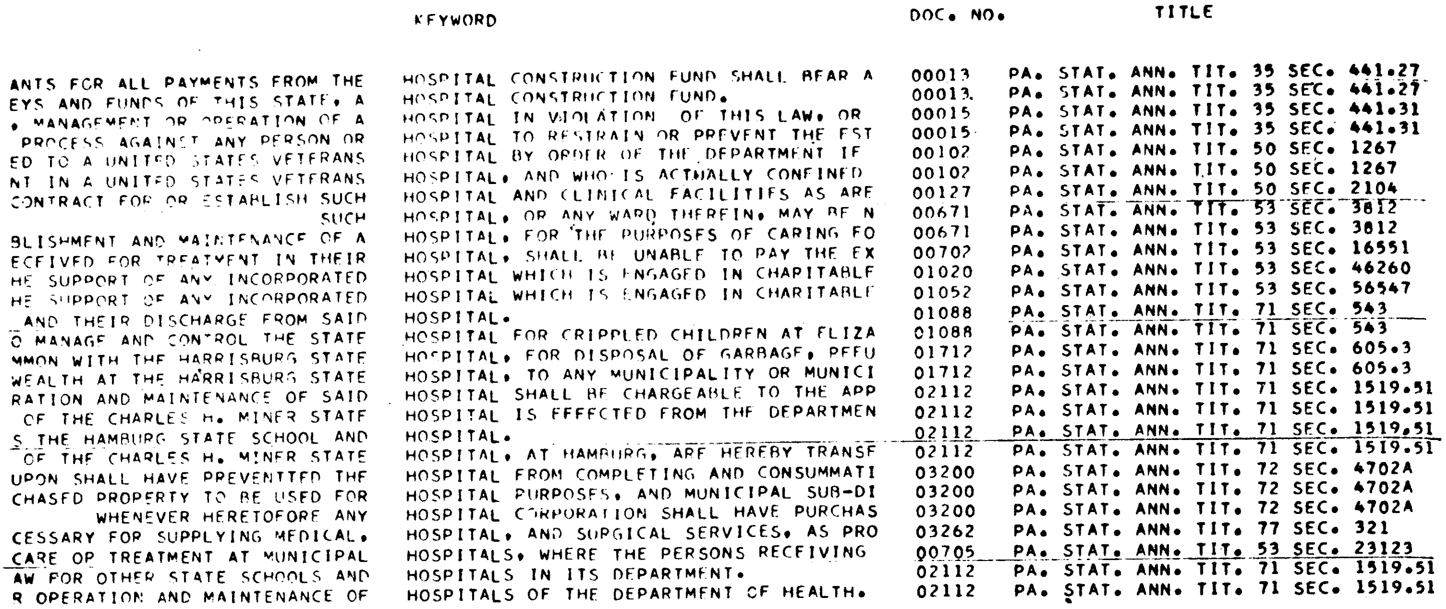
The success of the Key-Words-in-Combination at the University of Pittsburgh has prompted its commercial exploitation by the Aspen Systems Corporation of which John F. Horty is the president. Aspen provides KWIC-type searches of the full text of the statutes of all fifty states. According to Horty, the principal users of the Aspen system are legislative bodies and government agencies.[22]

How effective is KWIC? The following is typical of published experimental results:

> The results of a comparison of a [Health Law] Center search with a manual search on six different legal research problems by University

[21] William B. Kehl, John F. Horty, Charles R. T. Bacon, and Dennis S. Mitchell, "An Information Retrieval Language for Legal Studies," Communications of the Association for Computing Machinery, IV (Sept. 1961), p. 388.

[22] Beard, *op. cit.*, p. 379.



```
                                   KEYWORD                                   DOC. NO.            TITLE

ANTS FOR ALL PAYMENTS FROM THE   HOSPITAL CONSTRUCTION FUND SHALL BEAR A     00013    PA. STAT. ANN. TIT. 35 SEC. 441.27
EYS AND FUNDS OF THIS STATE. A   HOSPITAL CONSTRUCTION FUND.                 00013    PA. STAT. ANN. TIT. 35 SEC. 441.27
. MANAGEMENT OR OPERATION OF A   HOSPITAL IN VIOLATION  OF THIS LAW. OR      00015    PA. STAT. ANN. TIT. 35 SEC. 441.31
 PROCESS AGAINST ANY PERSON OR   HOSPITAL TO RESTRAIN OR PREVENT THE EST     00015    PA. STAT. ANN. TIT. 35 SEC. 441.31
ED TO A UNITED STATES VETERANS   HOSPITAL BY ORDER OF THE DEPARTMENT IF      00102    PA. STAT. ANN. TIT. 50 SEC. 1267
NT IN A UNITED STATES VETERANS   HOSPITAL. AND WHO IS ACTUALLY CONFINED      00102    PA. STAT. ANN. TIT. 50 SEC. 1267
CONTRACT FOR OR ESTABLISH SUCH   HOSPITAL AND CLINICAL FACILITIES AS ARE     00127    PA. STAT. ANN. TIT. 50 SEC. 2104
                          SUCH   HOSPITAL. OR ANY WARD THEREIN. MAY BE N     00671    PA. STAT. ANN. TIT. 53 SEC. 3812
BLISHMENT AND MAINTENANCE OF A   HOSPITAL. FOR THE PURPOSES OF CARING FO     00671    PA. STAT. ANN. TIT. 53 SEC. 3812
ECEIVED FOR TREATMENT IN THEIR   HOSPITAL. SHALL BE UNABLE TO PAY THE EX     00702    PA. STAT. ANN. TIT. 53 SEC. 16551
HE SUPPORT OF ANY INCORPORATED   HOSPITAL WHICH IS ENGAGED IN CHARITABLE     01020    PA. STAT. ANN. TIT. 53 SEC. 46260
HE SUPPORT OF ANY INCORPORATED   HOSPITAL WHICH IS ENGAGED IN CHARITABLE     01052    PA. STAT. ANN. TIT. 53 SEC. 56547
 AND THEIR DISCHARGE FROM SAID   HOSPITAL.                                   01088    PA. STAT. ANN. TIT. 71 SEC. 543
O MANAGE AND CONTROL THE STATE   HOSPITAL FOR CRIPPLED CHILDREN AT ELIZA     01088    PA. STAT. ANN. TIT. 71 SEC. 543
MMON WITH THE HARRISBURG STATE   HOSPITAL. FOR DISPOSAL OF GARBAGE. REFU     01712    PA. STAT. ANN. TIT. 71 SEC. 605.3
WEALTH AT THE HARRISBURG STATE   HOSPITAL. TO ANY MUNICIPALITY OR MUNICI     01712    PA. STAT. ANN. TIT. 71 SEC. 605.3
RATION AND MAINTENANCE OF SAID   HOSPITAL SHALL BE CHARGEABLE TO THE APP     02112    PA. STAT. ANN. TIT. 71 SEC. 1519.51
 OF THE CHARLES H. MINER STATE   HOSPITAL IS EFFECTED FROM THE DEPARTMEN     02112    PA. STAT. ANN. TIT. 71 SEC. 1519.51
S THE HAMBURG STATE SCHOOL AND   HOSPITAL.                                   02112    PA. STAT. ANN. TIT. 71 SEC. 1519.51
 OF THE CHARLES H. MINER STATE   HOSPITAL. AT HAMBURG. ARE HEREBY TRANSF     02112    PA. STAT. ANN. TIT. 71 SEC. 1519.51
UPON SHALL HAVE PREVENTTED THE   HOSPITAL FROM COMPLETING AND CONSUMMATI     03200    PA. STAT. ANN. TIT. 72 SEC. 4702A
CHASED PROPERTY TO BE USED FOR   HOSPITAL PURPOSES. AND MUNICIPAL SUB-DI     03200    PA. STAT. ANN. TIT. 72 SEC. 4702A
        WHENEVER HERETOFORE ANY   HOSPITAL CORPORATION SHALL HAVE PURCHAS     03200    PA. STAT. ANN. TIT. 72 SEC. 4702A
CESSARY FOR SUPPLYING MEDICAL.   HOSPITAL. AND SURGICAL SERVICES. AS PRO     03262    PA. STAT. ANN. TIT. 77 SEC. 321
CARE OR TREATMENT AT MUNICIPAL   HOSPITALS. WHERE THE PERSONS RECEIVING      00705    PA. STAT. ANN. TIT. 53 SEC. 23123
AW FOR OTHER STATE SCHOOLS AND   HOSPITALS IN ITS DEPARTMENT.                02112    PA. STAT. ANN. TIT. 71 SEC. 1519.51
R OPERATION AND MAINTENANCE OF   HOSPITALS OF THE DEPARTMENT OF HEALTH.      02112    PA. STAT. ANN. TIT. 71 SEC. 1519.51
```

Fig. 8 Key-Word-In-Context Output

> of Pennsylvania Law faculty members were promising. The computer did turn up 4½ times as many irrelevant statutes as the professors had, and about the same number of medium-relevant statutes, and missed two clearly relevant and two medium-relevant statutes found by the professors. But, it found 2½ times as many clearly relevant statutes as the professors had.[23]

In the framework of productivity measures, the above results indicate greater recall but less precision. Unfortunately, none of the systems examined by this author have been extensively tested with regard to recall-precision measures. At best, some experiments have produced enough data to approximate one point on an R-P curve. It is believed that that cost is the main obstacle in obtaining accurate statistics. In order to adequately test a system via R-P curves, a large data base must first be established such that the test will be statistically significant. Secondly, many manual searches by different people at different levels of exhaustivity must be conducted to generate sets of curves. Similarly, many automated searches are needed to generate sets of curves. Lastly, a qualified panel of appraisers must wade through all of the material retrieved by both manual and automated searches in order to judge the material's relevancy with respect to the search request. Clearly, this would be very expensive and time consuming.

[23] S. Mermin, "Computers, Law, and Justice: An Introductory Lecture," Wisconsin Law Review, (Winter, 1967), p. 64.

Getting back to evaluating Horty's KWIC system, the limited success of the University of Pittsburgh experiments is partially attributable to the choice of statutory law as a data base. Key word systems lend themselves to applications such as statutory law in which the vocabulary tends to be exacting, precise, and to the point. For the sake of consistency, the same language is often used to characterize a particular subject or action throughout a complete set of statutes. As Eldridge and Dennis point out, "An adequate selection of search terms will not be so easy in case law which is not always 'carefully framed in words chosen for clarity rather than literary quality!"[24] Aside from selecting search terms, the user must also enter all of the possible synonyms and grammatical variants to satisfy the "exact match" requirements of KWIC. The test results indicate that KWIC produces a large amount of irrelevant data. This may be reduced by specifying additional key words, or by restricting the relative nearness or position of the words, but this will also result in the reduction of the number of relevant documents retrieved. There is no good way to "throttle" the output of the KWIC system based on some measure of relevancy. On the positive side, KWIC was the first system to demonstrate

[24] William B. Eldridge and Sally F. Dennis, "The Computer as a Tool for Legal Research," Law and Contemporary Problems, XXVIII (1963), p. 89.

improved recall over manual search, which was not dependent on an indexing system. Since the full text of documents was stored, the researcher in structuring his search was in no way bound by the limitations of a rigid, hierarchial index. This in itself is a notable advancement in document retrieval.

Overall, one is left with mixed feelings with regard to KWIC. Although it does represent a significant advancement over manual systems such as West's, its dependence on matching exact words with all of its associated problems leaves one to suspect that there "ought to be a better way." Before discussing other algorithms, it is important to point out that the advantages of the KWIC technique have led to its widespread use in a number of different systems. The two major KWIC systems beside the Aspen system already mentioned, are LITE and OBAR, and will be described briefly. The general criticisms of KWIC apply to both of these implementations.

LITE is an acronym for "Legal Information Through Electronics." In 1961, the Office of the Staff Judge Advocate of the Air Force Center initiated an investigation as to the applicability of computers in the retrieval of legal information. After examining the results of the KWIC system at the University of Pittsburgh, the Air Force let out a contract to IBM to develop a similar system for the Air Force. The results of the effort was a KWIC-type system called LITE. The LITE's system data base contains the full text of the following documents:

(1) All titles of the U.S. code

(2) All published decisions of the Comptroller General of the U.S.

(3) All unpublished decisions of the Comptroller General of the U.S.

(4) The Armed Services Procurement Regulations, and

(5) The Defense Contract Audit Agency Manual as well as other regulatory material.

All in all, there is in excess of forty million words of text in the data base.

Richard Davis, director of the LITE project has reported experimental results as follows:

> During the test period (six months), 215 separate search questions, processed by the LITE System and researched manually, were evaluated by the test User activities.
> ...The overall analysis revealed:
>
> 1. In 16 of these 215 searches (7.5% of the total), the computer was less efficient than human research. LITE retrieved less relevant citations than were discovered by manual research.
>
> 2. In 95 of these 215 searches (44.1% of the total), the computer equaled human effort. LITE retrieved the same number of relevant citations as were discovered by manual research.
>
> 3. In 104 of these 215 searches (48.4% of the total), the computer (LITE) was more efficient and retrieved more relevant citations than were discovered manually.[25]

[25] Richard P. Davis, "The LITE System," JAG Law Review, XVIII (November-December, 1966), p. 9.

Thus in 48.4% of the searches, LITE produced greater recall as compared with manual search. In a different experiment, the LITE system was tasked to retrieve information for the Bureau of the Budget on a question which had already been manually researched. Of the total 137 relevant statutory provisions found by either the manual or automated search, LITE retrieved 128 citations as compared to 85 found manually.[26] In neither of the two experiments were there published statistics with regard to the number of irrelevant documents retrieved by the LITE system.

Perhaps the most successful application of computers to the retrieval of case law to date is the OBAR system. In 1967, the Ohio Bar Association, the Ohio Legal Center Institute, and other concerned organizations incorporated Ohio Bar Automated Research (OBAR), whose purpose was to develop a practical automated legal research system. OBAR awarded a contract to the Data Corporation (which is now a part of Mead Data Central, Inc.) to implement and test a KWIC system. The initial experiments used a test data base of fifty of the most recent volumes of Ohio Supreme Court Reports. The OBAR system was tasked to retrieve relevant cases on a test question regarding state sales tax and give-away promotions by oil companies (an actual question in a then-pending Ohio dispute). A manual search found five

[26] Ibid.

cases while the OBAR system yielded ten cases of which one was irrelevant.

OBAR's data base now contains the full text of the Ohio Constitution and the code of statutes as well as the full text of decisions of the Supreme Court and Court of Appeals of Ohio. OBAR's commercial success is partially based on its use of relatively advanced computer equipment. The OBAR system is time shared--that is to say, many users may access the central computing system at the same time. Remote access into the system is achieved through TV-like video display and keyboard terminals. These remote terminals allow the attorney to interact with the system in real time. If too many or too few citations appear as the result of a search, the attorney can immediately modify his search request to correct the situation. Attorney/system interaction greatly enhances the flexibility of systems operations. Optional peripheral equipment for making paper copies of the information on the video screen is also available. A lawyer may be trained to use the OBAR system in a matter of one to two days.

Over twenty Ohio law firms have installed OBAR terminals in their offices. The Mead Corporation anticipates that as many as 300 Ohio law firms will subscribe to the system. Other than limited experiments, such as the one cited above, very little data is available on the productivity of OBAR.

The fact that OBAR is operationally used in a number of practicing law firms is a notable tribute to its success. However, there has been some reported dissatisfaction with OBAR and at least one subscriber has cancelled service.[27]

Since Horty's early work at the University of Pittsburgh, two experimental systems have been developed which, although they rely on a KWIC approach, attempt to circumvent the exact match requirements. The first such system was developed under the direction of Robert A. Wilson at the Southwestern Legal Foundation. The project was dubbed OGRE which stood for Oil and Gas Reports - Electronic. Like the Horty system, OGRE stored the full text of documents. The initial data base contained 250 federal decisions pertaining to tax problems in the oil industry. The principal refinement of the OGRE system was that the vocabulary file was edited by human intervention to only include the roots of all the words. Each root was then assigned a numeric code. Thus, "injure," "injured," "injuries," and "injuring" would be classified as the same word. This innovation permits the user to specify key word requests without including all of the possible grammatical variations. Work on the OGRE system was discontinued in 1963 because of a lack of funds.

[27] W. G. Harrington, "Computers and Legal Research," American Bar Association Journal, LVI (December, 1970), p. 1147.

The second such system represents perhaps the most complicated variation on KWIC. This system is knows as the "Semantic Coded Abstract" approach and was developed at the Western Reserve University Center for Documentation and Communication. Unlike most KWIC systems, the Semantic Coded Abstract system only stores the abstract of the document. The unique aspect of the system is that the abstract is not stored in natural English text. Rather, it is transformed into a scientifically coded symbolic language with rigid rules in which each term has a rather specific meaning. The basis of this symbolic language is similar to the generation of English words from Greek and Latin roots. Starting with the roots "meter," "graph," and "stat," one may add rather standard prefixes to obtain a variety of other words: thermometer, hydrometer, photometer, photograph, phonograph, photostat, thermostat, hydrostat, etc. In a similar manner, rigid semantic rules transform the word "blueprint" into the symbolic code CVNS DACM RUGL 3002, while the word "specifications" (which is close in meaning but not identical to blueprint) is encoded as CVNS DACM RUGL 3001.

A user specifies his request as number of key words in natural English. An analyst then takes these words and encodes them into the semantic code. The coded key words are compared to the coded vocabulary file of the data base. Documents which contain the desired combination of coded key words are presented to the user. The main advantage of the Semantic Coded Abstract approach is that the user does not need to list all of the synonyms in his key word request. All synonyms will presumably be characterized by the same semantic code. The obvious disadvantage to the system is the large amount of skilled human interaction which is necessary to encode both the data base document collection and the user-search requests. Limited experiments of the Semantic Coded Abstract system on the sales portion of the Uniform Commercial Code yielded inconclusive results.[28] The expense of semantic encoding has prevented the system from advancing beyond the experimental stage.

3.2.3 Probabilistic Search Systems

The most innovative and promising use of computers in document retrieval has been the application of probabilistic statistics in the structuring of search term requests. This algorithm, which is known as the "Association Factor Technique," was pioneered by Dr. H. Edmund Stiles of the Department of Defense in 1958.[29] The Association Factor

[28] "Proceedings of the Special Committee on Electronic Data Retrieval," Modern Uses of Logic in Law, March, 1962, p. 50.

[29] H. Edmund Stiles, "The Association Factor in Information Retrieval," Journal of the Association for Computing Machinery, VIII (1961), p. 271.

Technique will now be briefly described (a detailed explanation and analysis of the technique appears in the next chapter). The full text of all of the documents in a collection is analyzed to create a vocabulary file which contains all of the different words in the text and the "degree of association" between each term and all other terms. That is to say, for any pair of terms in the document collection, the file would contain the degree of associativity of these terms with each other. This associativity is a measure of the relationship between two words based not on the meanings of the words or the semantic structure of the words, but rather on the number of times the two words co-occurred in the same document. If two words were completely independent, there is an expected number of times that the two words would co-occur in any particular document based on the relative frequencies of the words in the English language. However, if those two words co-occur more than the expected number of times within some set of documents, this would be piece of probabilistic evidence that the two words are somehow related to each other. This relationship can be mathematically quantified a number of different ways based on the comparison of expected and actual co-occurrences. Thus, one may compute an "association factor" between two words. Words which are highly "related" within the collection of documents will have a greater association factor than those words which are unrelated. An example of associated terms is given by Stiles.[30] In identifying words associated with "friction" within a limited set of documents, the following terms had the greatest associativity: wear, thin, lubrication, and belt. Clearly, none of these terms are synonyms for friction, nor are they in any way semantically related to friction. However, within the context of those documents dealing with friction, these words occurred more often than expected. Thus, they are associated with friction.

This piece of information may be used in the following way: Suppose a researcher were interested in retrieving documents about friction in some kind of machinery. Using a KWIC system, he would specify the kind of machinery and the word friction. KWIC, however, would not retrieve an article on how a particular lubricant would prevent excessive wear in the machinery unless the word friction were explicitly stated. Such an article might be totally relevant to the researcher's interests, but because all of the possible terms were not expressed in the KWIC request, the document would not be retrieved. The most powerful aspect of the Association Factor Technique is the ability of the system to "expand" the

[30] _Ibid._, p. 273.

researcher's original request terms to include words which, although not synonyms, are somhow related to the search terms. In this manner, the words wear, thin, lubricant, and others would be added to the original search terms and relevant documents which did not explicitly use the word friction would be retrieved. The expansion of request terms may be performed more than once. A secondary expansion would produce all of the terms associated with wear, thin, lubricant, and belt.

The data base of Stiles' and subsequent Association Factor systems did not contain the full text of the documents, but rather, manually extracted key words which describe each document. Since user requests are expanded to include a variety of related terms, not much is lost by not storing the full text. The use of a limited number of key words to describe each document has the advantage that the amount of storage necessary to maintain a data base of a large number of documents is much less, and the time necessary to perform a full data base search is greatly reduced.

The operation of the system is described as follows: The user structures his requests in the same manner as the KWIC system - he specifies a number of key words and may impose the logical relationships of "and" or "or" between the words. The system now takes the request terms and creates an expanded list of words which includes the original key words and all words whose association factor with the key words is greater than some threshold. This list may be expanded a second time to obtain even greater depth. "No word in this list could be substituted for the request because each has its own variety of meanings and uses, yet it would be hard to use a group of them without touching on the subject of the request."[31] Each word in this list is given a weight equal to the normalized association factor between that word and all others in the list. Taking into account the specified logical relationships, this expanded list is compared to the file of words which index the document collection. "Whenever the terms match, the weight of the requested term is assigned to the corresponding document index term. The sum of these weights for each document is called the document relevance number. This number should indicate the degree of fit between the request and the contents of the document."[32] Herein lies the other great advantage of the association factor technique. Documents may be ranked according to their probabilistic relevance to the researcher's request. Such a system holds the promise that documents might be examined in an optimal fashion; i.e., the most relevant document will be

[31] Ibid., p. 277.

[32] Ibid.

examined first. This would clearly enable the researcher to make efficient decisions with regard to search intensity and search termination.

Although Stiles did not publish system results on recall and precision, the table illustrated in Figure 9 testifies to the system's ability to rank retrieved documents in order of relevancy.[33] In the table, "Document Relevance Number" is the computer-assigned rank while the "Degree of Association" is a human judgment as to the relevance of the document with respect to an experimental search for documents related to thin films. A subsequent government contract to investigate associative techniques was awarded to Arthur D. Little, Inc.[34] As part of this study, a number of experiments were conducted which involved automated search. The data base was the entire collection of NASA documents (nearly 100,000) which contained 18,000 different words. In comparing a manual and an automated search for documents on "rendezvous and docking," the automated system found all but 36 of the 179 unclassified documents found by an analyst, and in addition, found 61 relevant documents which were not retrieved manually.[35] In assessing the success of the system, the

[33] *Ibid.*, p. 276-277.

[34] Reported in "Application of Statistical Association Techniques for NASA Document Collection," *NASA Contractor Report CR-1020*, prepared by Paul E. Jones, Robert M. Curtice, Vincent E. Giuliano, and Murry E. Sherry (Cambridge: Arthur D. Little, Inc., 1968).

[35] *Ibid.*, p. 39.

Document Relevance Number	Degree of Association	Document Relevance Number	Degree of Association
24.32	YES	7.85	NO
24.22	YES	7.27	NO
24.22	YES	7.16	P
24.22	YES	7.13	M
22.47	YES	7.13	M
19.87	YES	6.93	P
15.59	YES	6.94	NO
15.30	P	6.94	NO
14.83	YES	6.94	NO
12.54	NO	6.91	NO
11.81	M	6.40	P
11.20	M	6.40	P
10.72	M	6.36	NO
10.14	P	6.03	M
10.08	M	6.01	NO
10.05	YES	5.82	P
9.83	YES	5.33	NO
9.66	M	5.23	NO
9.50	M	5.23	NO
9.38	M	4.65	NO
9.38	NO	4.41	P
9.30	P	4.41	P
9.30	P	4.41	P
9.30	P	4.26	NO
9.30	P	2.70	NO
9.18	P	2.70	NO
9.12	P		
8.66	M		
8.03	P		
7.95	M		

YES -- Document does contain relevant information.

M -- May be useful background information.

P -- Possibly contains useful background information.

NO -- Does not contain relevant information.

Fig. 9 Relevancy Ranking

Arthur D. Little study reports "Our observations during trials of the system's operation have reinforced our view that the use of term association holds great promise for improving the cost/effectiveness of retrieval searching."[36]

The first application of the associative technique to the retrieval of law was performed in the early 1960's at George Washington University Graduate School of Public Law under the direction of John C. Lyons and in conjunction with the Datarol Corporation.[37] Using IBM 1401 and 7090 computers, associated searches were conducted on a data base of 350 anti-trust documents. Although this limited set of experiments could not produce statistically defensible results, Lyons has stated that the association factor is the "best technique known to date for document retrieval."[38] The departure of Mr. Lyons and a shortage of funds terminated further association work at George Washington University. Mr. Lyons has subsequently founded Autocomp, Inc., which primarily performs codification and computerized photo-composition of legal documents. He hopes to begin further experimentation with

[36] Ibid., p. 6.

[37] John C. Lyons, "New Frontiers of the Legal Technique," Modern Uses of Logic in Law, December, 1962, p. 256.

[38] Personal interview with Mr. Lyons, November 20, 1972.

the association factor technique at Autocomp sometime during this next year.

The only other reported experimentation of applying the association factor technique to the retrieval of legal documents was a research project sponsored by the American Bar Foundation. In 1961, the ABF initiated a joint research project with IBM called "Legal Research Methods and Materials." The objective of this project was to investigate the use of computers as applied to legal problems of searching and indexing. The association factor was used as the basis of a search system with data base of nearly three thousand cases taken directly from West's Northeastern Reporter. Extensive disagreement among a qualified panel of evaluators as to the relevance of retrieved cases prevented accurate quantitative assessment of the system's performance. Perhaps the most important result of the ABF-IBM project was the development of a methodology by which document index words may be automatically separated in preparation for association factor searches. This process will be discussed in detail in the following chapter. When the project was terminated in 1965, the following conclusions were made:

> "it seems likely that additional research will refine the system to the point of operational adequacy" ...searching with the amplified word-lists seemed twice as efficient as straight word

> matching..the system was fairly insensitive to different phrasings of the same search question.. "the chances are good that output quality can be brought to the vicinity of 95% complete retrieval with no more than 25% inapplicable citations ."[39]

3.2.4 Other Systems

Many other document/information retrieval systems have been developed both in the U.S. and in other countries.[40] These efforts range from limited experiments to operational systems. However, most of these systems rely on some variation of three generic retrieval algorithms described above and, as such, will not be analyzed here.

3.3 ALGORITHM SELECTION

In analyzing the various approaches to automating case law retrieval, a number of systems have claimed to have "solved the problem" of automatically retrieving cases relevant to an attorney's request. Indeed, some of these systems are fully operational and in use in offices of a number of practicing law firms. Almost no one will disagree with the claim that these systems are an improvement over manual systems such as West. Given the archaic, inefficient, error-prone, time-consuming nature of manual case law indices, even

[39] Mermin, op. cit., p. 62.

[40] For an analysis of foreign efforts, see "Computerized Legal Research in Countries Outside North America," Jurimetrics Journal, XII (March, 1972), p. 119 or Beard, op. cit., p. 388.

a relatively unsophisticated application of computers such as the "Point-of-Law" approach would certainly be welcomed as a vast improvement. This leads to the more difficult questions of how much improvement does a particular system produce and can additional improvement be obtained through an alternative approach?

KWIC-type systems are the most successful case law retrieval systems in operation today. Even without accurate recall-precision curves, it is obvious that these systems allow the attorney to obtain many more relevant case citations than he would be able to otherwise obtain. However, the KWIC system brings with it its own set of problems for the researcher. In structuring a request, the researcher must take into account all of the synonyms and grammatical variations of the key words to meet the exact match requirements of KWIC. He can increase or reduce the number of documents retrieved by altering the search request terms, but this does not guarantee that the most relevant documents will be retrieved. Although it is more accurate than manual search, KWIC systems do retrieve an abundance of irrelevant citations. Clearly, there is a great deal of room for improvement.

The remainder of this paper will be devoted to a system design of a practical computerized system for the retrieval of case law. Despite the fact that there are no operational case law retrieval systems based on probabilistic indexing, the experiments conducted in this area indicate that association factor techniques have the greatest potential for overcoming the problems of KWIC and significantly improving case law retrieval. The main advantages of the association factor technique may be itemized as follows:

(1) it is not dependent on a hierarchial indexing scheme

(2) it is a subject- rather than a language-oriented search

(3) the system output may be ranked in order or relevance

Statistical association does not rely on a rigid hierarchial indexing system such as the West system or Morgan's "Point-of-Law." Therefore, it remains flexible with regard to new issues and controversies which might arise. A case whose meaning has changed with time is not hidden from the researcher by outdated indexing. Furthermore, errors and distortions created by the "pigeonholing" of cases into an itemized list of legal issues is eliminated. Because the researcher's request terms are associatively expanded to include a variety of words germane to the attorney's problem, the search becomes subject rather than language oriented. Unlike the KWIC system, the researcher does not need to guess the exact words used by the judge writing the opinion of a given case. Lastly, the association technique will rank cases in the order of probabilistic relevancy. This will allow the attorney to conduct case research at the intensity appropriate to the problem. Since he will examine the most relevant cases first, his analysis will be performed in an optimal manner. Thus, the attorney can chose to look, for example, at only the 5 most relevant cases, or the 100 most relevant cases depending on his needs.

Because of all the advantages cited above, the association factor technique is chosen as the basis for the proposed improved system. A number of changes to the basic algorithm used in the early experiments will be suggested in order to enhance the system's ability to accurately retrieve those cases most relevant to the user's needs.

3.3.1 Association Factor

Although the association factor was described generally in the previous section, a more detailed description is appropriate at this junction. In short, the association factor is a measure of the degree to which two terms are related within a particular collection of documents. This might be best illustrated by an example. Suppose the term "search" appeared in two thousand cases out of a total collection of a million cases. The probability of finding the word "search" in a case chosen at random would be one in five hundred. Similarly, if the word "seizure" appeared in a thousand cases, the probability of finding the word "seizure" in a case chosen at random would be one out of a thousand. If the two words were independent, the probability of finding both words (joint probability) in a randomly selected case would be the product of the two probabilities or one chance in 500,000. Thus, the expected number of co-occurrences of "search" and "seizure" would be equal to the product of the number of total cases and the joint probability or $1{,}000{,}000 \times \frac{1}{500{,}000} = 2$. Suppose, however, that "search" and "seizure" co-occurred in two hundred cases. It could, therefore, be concluded that the two words co-occur a hundred times more often than expected had they been independent. Thus, they are somehow related. This relationship is purely statistical--"search" and "seizure" are certainly not synonyms and do not have a common semantic root. This particular measure of associativity is expressed as the ratio of observed co-occurrences to the expected number of co-occurrences. The actual formula is derived mathematically in Figure 10. Many different measures of associativity have been proposed. For example, one could measure how much more frequently two terms co-occurred than expected by the use of standard deviations rather than simple ratios. Figure 11 illustrates four other common measures of association.[41] Formula I is the original

[41] *NASA Contractor Report*, *op. cit.*, p. 55.

$N =$ Number of documents (cases) in collection

$f_a =$ Number of documents containing term a

$f_b =$ Number of documents containing term b

$f_{ab} =$ Number of documents containing both term a and b

Probability a document chosen at random contains term a

$$P_a = \frac{f_a}{N}$$

Probability a document chosen at random contains term b

$$P_b = \frac{f_b}{N}$$

Probability a document chosen at random contains both term a and term b assuming independence

$$P_{(a \cdot b)} = \left(\frac{f_a}{N}\right)\left(\frac{f_b}{N}\right)$$

Expected co-occurrences of a and b if independent

$$N\,P_{(a \cdot b)} = \left(\frac{f_a}{N}\right)\left(\frac{f_b}{N}\right) N = \frac{f_a f_b}{N}$$

$$\frac{\text{Observed co-occurrences}}{\text{Expected co-occurrences}} = \frac{f_{ab}}{f_a f_b} N$$

$$\text{Association}_{ab} = A_{ab} = \frac{f_{ab}}{f_a f_b}$$

$A_{ab} > 1$ Terms a and b are associated.

$A_{ab} = 1$ Terms a and b are not associated.

$A_{ab} < 1$ Terms a and b are negatively associated.

Fig. 10 Statistical Association

(I) $$A_{ab} = \log_{10} \frac{\left(\left| f_{ab} N - f_a f_b \right| - \frac{N}{2} \right)^2 N}{f_a f_b \, (N - f_a) \, (N - f_b)}$$

(II) $$A_{ab} = f_{ab} - \frac{f_a f_b}{N}$$

(III) $$A_{ab} = \frac{f_{ab}}{f_a + f_b - f_{ab}}$$

(IV) $$A_{ab} = \frac{\left(f_{ab} - \frac{f_a f_b}{N} \right)}{\sqrt{\frac{f_a f_b}{N}}}$$

Fig. 11 Other Measures of Association

association factor used by Stiles. It is a form of the Chi square formula using the marginal values of the 2 x 2 contingency table and the Yate's correction for small samples. Formula II is just the difference between the observed and the expected number of co-occurrences. Formula III is the number of observed co-occurrences normalized by the number of documents indexed by only one of the terms. Formula IV is the number of standard deviations the observed co-occurrence falls to the right of the expected co-occurrence. To date, there has been no accurate assessment as to which one of the formulas provides "better" measures of relationships between words in a document collection. (It is difficult to imagine how the formulas could be compared quantitatively). However, each of the formulas is dependent on f_a, f_b, f_{ab}, and N as defined in Figure 10. Once a data base has been established with all of these four variables, a computer could quickly run through all of the association formulas mentioned and compute relationships based on each of the measures. Perhaps a team of evaluators could analyze the various measures and come to some agreement as to which formula is "best."

The ability of the association factor to identify words related to an initial set of words, is used in the system to augment or "expand" the initial set of key words which the attorney specifies to describe his problem. Stiles suggests that the user's request terms be expanded two times via the association factor. The reason given is that the second generation expansion would include a number of synonyms or near-synonyms which might not be included in the initial expansion. Suppose the original request terms included the word "nuclear." An initial expansion of the request terms might yield associated words such as "reactor," "power," and "fission." Another expansion of this new list of terms might produce the result that "atomic" (which is a near-synonym for "nuclear") is associated with "reactor." If a number of authors were writing on the subject and each were consistent in using either the terms "nuclear reactor" or "atomic reactor" but not both, then the terms "nuclear" and "atomic" would not co-occur, and hence, they would have a small association factor. Two associative expansions would solve this problem. The Arthur D. Little study found that based on their document collection data base, synonyms did occur in the initial term expansion but agreed with Stiles that a second generation expansion of terms can contribute to a more complete formulation of association profiles.

Once the original request terms are expanded twice through the use of the association factor, each term in the new list (whose association factor is greater than some threshold) is assigned a term weight based on the following formula:

$$W_i = \frac{\sum_{j=1}^{N} A_{ij}}{N}$$

W_i Weight of term i

A_{ij} Association factor between terms i and j.

N Total number of terms in the final expanded list.

This weight is a measure of the probabilistic relevance of each term to the original request.

The expanded list of request terms is then compared with the manually generated index terms which describe each document. When there is a match, the weight of that term is assigned to that document. The sum of the matched term weights is the document's relevance number.

$$R_k = \sum_{i=1}^{N} W_i$$

for all i contained in the index of document k

R_k relevance number of document k

This document number is a probabilistic measure of the document's relevance to the user's original request. Thus, the documents in the collection may be ranked in the order of relevance, and the user might request to examine those documents whose relevance number is above some threshold.

The entire process is illustrated schematically in Figure 12. The full text of the document collection is used to generate a vocabulary file which contains all of the different words (with perhaps the exception of the 100 or so

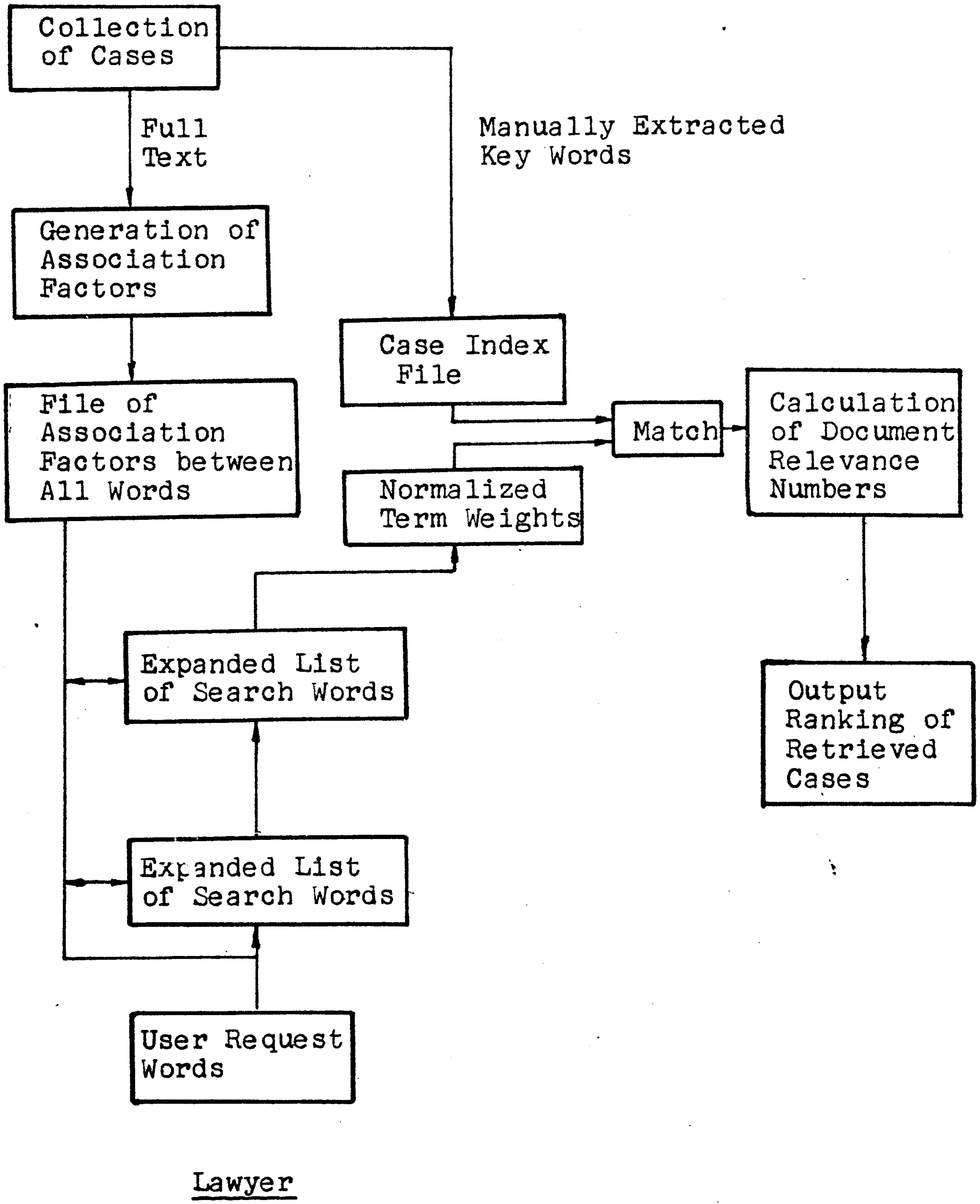

Collection of Cases
Full Text
Manually Extracted Key Words
Generation of Association Factors
Case Index File
File of Association Factors between All Words
Match
Calculation of Document Relevance Numbers
Normalized Term Weights
Expanded List of Search Words
Output Ranking of Retrieved Cases
Expanded List of Search Words
User Request Words
Lawyer

Fig. 12 Association Factor Retrieval System

most common words) used in the documents and the association factor between all possible pairs of words. Key words describing each document are extracted to form a word file indexing each of the documents. The attorney's key word requests are augmented with associated terms found in the vocabulary file. This process is repeated a second time on the expanded list. Each search term in this list is assigned a term weight based on its normalized association with the other terms. The terms are then matched against the document index words, and document relevance numbers are computed.

3.3.2 Automatic Indexing

Perhaps the largest single source of error in the early association factor systems stemmed from the manual extraction of key words which would describe each document. Stiles and Lyons were able to trace many of the output errors back to erroneous manual indexing. The quality of retrieval systems dependent on manual indexing is limited by the skill and quality of the analyst performing the indexing. Two qualified analysts will often disagree on the indices for the same case. "Manual indexing requires unchallenged acceptance of another person's classification thereby limiting every lawyer using the index to the ability of the indexor."[42]

[42] William A. Fenwick, "Automation and the Law: Challenge to the Attorney," Vanderbilt Law Review, XXI (March, 1968), p. 261.

It is therefore desirable that the proposed improved system under design include some means by which the key descriptive words of a document could be automatically extracted without the errors and cost attributable to human intervention. The ABF-IBM research project "Legal Research Methods and Materials," developed such an algorithm. Essentially, this algorithm quantifies word frequency distribution such that informing words may be separated from non-informing words. The histograms of word frequency distributions in Figure 13 illustrate how this might be done. Non-informing words (such as "again," "before," "to," "another,") occur fairly often in many documents. On the other hand, informing words (such as "nuclear," "poison," "wiretapping," "contract,") occur zero times or only a few times in most documents and occur many times in a relatively small number of documents. By measuring the degree of "skewness" of the word distribution, one could task a computer to isolate the informing words in a case, thereby realizing automatic case indexing. The ABF-IBM study revealed that the best measure of word distribution which could be used to discriminate between the two kinds of words was the "ratio of raw occurrences (for a given word) to the reciprocal of the coefficient of variation of its within document frequency, normalized for document length."[43]

[43] Sally F. Dennis, "The Design and Testing of a Fully Automatic Indexing-Searching System for Documents Consisting of Expository Text," *Information Retrieval*, ed. George Schecter (Washington: Thompson Book Company, 1967), p. 75.

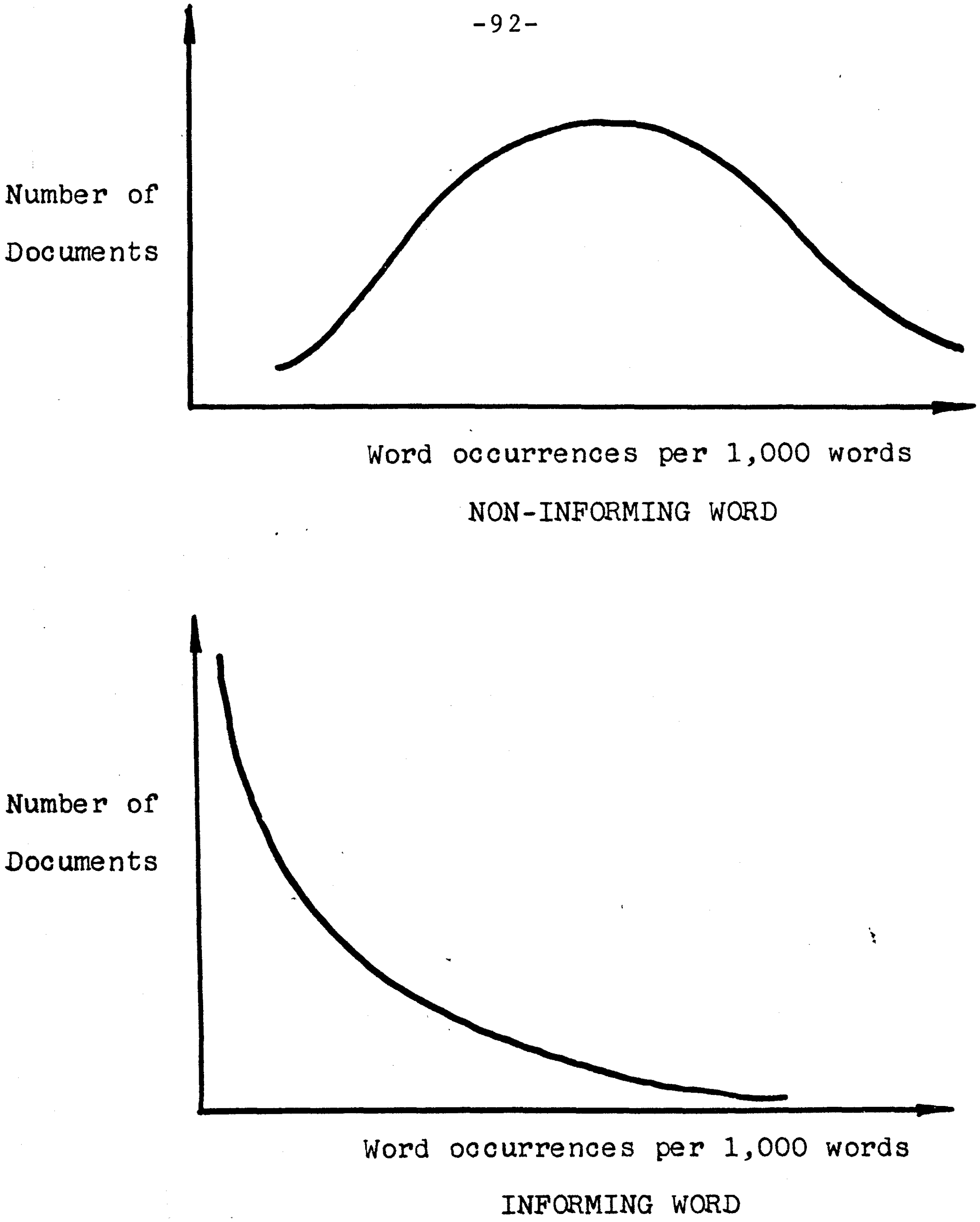


Fig. 13 Word Frequency Curves

The mathematical formulation for this measure is presented in Appendix I.

3.3.3 Relevance Ranking

In analyzing the association factor technique, a number of improvements to the basic algorithm can be made to optimize the system's performance with regard to the retrieval of case law. Aside from the automatic indexing mentioned above, most of these modifications fall into the general category of relevance ranking.

The first such modification was suggested by R. P. Anderson of Lehigh University. This modification is in the form of a correction to the document relevance formula to take into account the fact that the number of terms which index documents may vary substantially. Take the absurd example in which some encyclopedic document such as Webster's Dictionary is entered into a data base. The terms automatically extracted from the dictionary in order to index it would probably include all of the informing words on the system. Thus, any set of request terms would match the dictionary's index terms which would result in the dictionary's always being retrieved as the most relevant document almost independent of the request. Anderson's correction, which was found experimentally, normalized the document relevance number based on the number of indexed terms.[44]

$$\text{Document Relevance Number} = \frac{(R)\ (t)}{(T)}$$

R = Relevance number as measured by Stiles

t = Number of terms indexing that document which match with the search terms

T = Total number of terms indexing that document

The remainder of the algorithm modifications involve the interaction of the lawyer with the system in order to "steer" the search onto the right path. Perhaps the matching of expanded search terms and the summing of terms weights to arrive at a document relevance number is sufficient for obtaining the most knowledge on a given subject, but the lawyer's needs are somewhat different. He wants to know what is the current status of the law as applied to a given problem. Thus, a fifty-year-old case which contains all of the search terms is actually less relevant than a case decided last year on the same subject, but containing only a few of the search terms. Similarly, a U.S. Supreme Court ruling may be more relevant than a case from a state appeals court. However, the system cannot have rigid rules with regard to date and court in computing case relevancy. Legal scholars,

[44] Ronald R. Anderson, "An Associativity Technique For Automatically Optimizing Retrieval Results," (Sponsored by the National Science Foundation under grant No. GE-2569 and by the Office of Naval Research, contract Nonr- 710 08), Lehigh University Center for the Information Sciences, 1968, p. 12.

judges, or attorneys, may wish to analyze the sequence of cases which have led to a particular legal concept. Thus, the system should retain flexibility with regard to relevance ranking. This calls for interaction between the system and the user.

The first such proposed interaction of the lawyer and the system concerns the computation of search term weights. In the basic association factor system, the weights of the search terms are derived from a computation of the normalized association factor of each term with all others on the expanded list. This weight is interpreted as the probabilistic relevance of each term to the user's request. This is satisfactory only if the original key words in the request were at an optimal level of specificity. If the original key words were too specific or not specific enough, the computed term weights would not actually reflect their true relevance to the lawyer's request. Since the attorney is structuring the request, he is in an excellent position to assess term relevance and, therefore, should be allowed to intervene and modify the term weights based on his own perception of relevancy. An example might serve to illustrate this point. Suppose the original key words included the term "surveillance." An associative expansion might yield the terms, "spy," "follow," "photograph," "electronics," "wiretapping," and "eavesdropping" with their corresponding term weights.

If the attorney was not interested in any of the possible forms of electronic surveillance, he should be able to eliminate words such as "wiretapping" from the list. Otherwise, irrelevant cases on wiretapping would be presented in the output. Similarly, if photographs played a key role in the problem at hand, the attorney should be able to increase the weight of that term. If the expanded list of terms sparked the lawyer's imagination such that he thought of additional key words not on the list, he should be able to either add these words with an estimate as to their weights or initiate another associative expansion on a new set of key words.

In this proposed system, the attorney will be presented with the expanded list of terms and their weights after the associative expansions. At this point, the attorney can modify the weights of any of the terms. This could easily be done based on a scale of 0 to 10 in which the weight zero would remove the term from the search list while a weight of ten would indicate that that term is very relevant to the search request. New terms with estimated relevance wieghts could be added to the list by the attorney. The attorney is also given the option of performing another expansion either on the existing list or a set of new terms before proceeding with the actual data base search. It is only after the attorney is completely satisfied with the expanded list of search terms andtheir relevance weights that the system may continue the document retrieval algorithm.

Goldblum and others have suggested that the attorney should be able to impose a set of constraints or bounds on the system which would exclude the searching of some parts of the data base.[45] For example, such a constraint could take the form of requesting only Massachusetts cases since 1950. It is in the opinion of this author that these suggestions do not go far enough in providing the interactive flexibility needed by the lawyer to custom tailor the search to meet his exact needs. In this proposed system, the attorney will have the option of reordering relevance ranking of cases through a linear combination of factors which quantify the "importance" of cases as measured by the number of times that case has been cited, and the court and date of the opinion, all within some jurisdictional bound.

In some sense, the number of times that a given case is cited or referenced in subsequent opinions, is a measure of the importance of that case. As an example, the case of Brown V. Board of Education (1954), which was the landmark decision involving public school desegregation has been cited again and again in subsequent disputes. The document relevance number of a given case could be modified to take this into account by multiplying the number by the ratio of

[45] Goldblum, op. cit., p. 59.

the number of times that case has been cited to the maximum number of citations of any case in the output list of cases.

The number of times a particular case was cited in later disputes is not an accurate measure of case importance all by itself. This measure must be augmented with an expression for the timeliness of the decision. For example, the doctrine expressed in *Betts v. Brady* (1942) was the law for many years and the case was cited numerous times in similar disputes involving the right of a defendent to a state-appointed attorney. However, when the *Betts v. Brady* decision was overturned by *Gideon v. Wainwright* (1963), that became the law of the land and was subsequently cited. In 1966, *Miranda v. Arizona* drastically modified and expanded the doctrine of *Gideon v. Wainwright*. This later decision is now the law. Based strictly on the number of times that a case was cited, *Betts v. Brady* would be judged more important than *Miranda v. Arizona*. However, because this later case served to nullify or modify the law expressed in previous cases and because it is the current state of the law, it should be more relevant to the attorney who is trying to predict how the law will be applied to a dispute at hand.

In the proposed system, the attorney will have the ability to assign weights to different time periods. On a scale of 0 to 10, the attorney can specify which time frames are important to him in his search for case law. A zero would mean that cases in that time frame would not be considered while numbers from 1 to 10 would modify the document relevance numbers of the cases to provide a higher relevance ranking to those cases decided within the time frame of interest. The following is an example of how an attorney could specify time period weights if the most recent cases were most important to him:

Dates	Weight	Meaning
1960-1973	10	of highest importance (relevance)
1950-1960	8	important
1930-1950	3	of marginal value
before 1930	0	cases before 1930 are irrelevant

Someone interested in the development of a legal concept might give the same weight to all time periods.

The reasoning used in the above discussion is also applicable in analyzing the impact of the type of court which issued the opinion, on the relevance of a case. The United States Supreme Court is the highest court of the land, and as such, all of its decisions are binding on all other federal and state courts. In this sense, a Supreme Court ruling on a given issue might be more relevant than a state appeals court on the same issue. However, a relevant appeals court ruling might contain a more timely expression of a legal concept which has not yet reached the Supreme Court. Furthermore, only a very small fraction of legal disputes are decided in the Supreme Court. In this system, the attorney will have the option of assigning weight numbers to the different kinds of courts whose decisions are contained in the data base. If the attorney believes that the relevant case law may be found in the Supreme Court decisions, he will assign a high number (such as 10 on a scale of 0 to 10) to the Supreme Court and lower numbers to other courts. In a general case law search, all courts might be given the same weight.

The reordering of case relevance ranking through measures of subsequent case citations, date of decision, and type of court, must somehow be mathematically combined and formulated into the equation for document relevance number. This may be accomplished in the following manner. The original document relevance number can be multiplied by a "lawyer correction" factor which is a linear combination of the three weighted measures of citations, date, and court. The coefficients of the terms in the linear combination may be specified by the lawyer. To put it in another way, the relevance measure of subsequent citations is automatically calculated while the measures of date of decision and type of court are based on lawyer-assigned weights. These three measures are now each assigned a weight coefficient by the lawyer to indicate the relative importance of each measure as judged by the lawyer. For example, the lawyer might decide that the date of a case is a more important measure than the type of court--he would, therefore, specify a higher coefficient for date than for court. These measures and their coefficients are added together to form a lawyer-correction factor. The calculated document relevance number is multiplied by the lawyer-correction factor to obtain the final ranking. This process is not nearly as complicated as it sounds from the above explanation. A description of how the attorney would perform this operation is contained in the next chapter. The choice of a linear combination of lawyer-assigned weights is by no means the only method for improving the document-relevance formula. One could think of a number of ways to modify the document-relevance equations to take into account important variables specified by the lawyer. Just as with the problem of selecting a measure of associativity, different formulas for document relevance can be experimentally tested once a system has been implemented and a data base established. From the test results, one could determine which document-relevance formula is most responsive to the lawyer's interaction and needs in controlling case law search.

The final lawyer interaction with the search system algorithm is the constraining of the search with respect to jurisdiction. This option has been used in other automated case law systems. Essentially, this allows the lawyer to specify, for example, that only cases in Massachusetts be searched, or only cases in New York, Massachusetts, and

New Hampshire. This final constraint results in the document-relevance formula illustrated in Figure 14.

3.3.4 Summary of Algorithm

The algorithm for the proposed computerized case law retrieval system is shown graphically in Figure 15. As before, the full text of the collection of cases is used to generate the association factors between all of the words in the vocabulary file. The full text is also automatically analyzed in order to extract informing words which will index each case. Other descriptive information such as the date and court of the case, is also added to the index file. The association factors are used to expand the initial request terms two times. The lawyer may add or delete words from the final list and modify the term weights of any of the words. These words are matched against index words, and then document relevance numbers based on word matches and lawyer-controlled variables are calculated. The cases may now be ranked in order of probabilistic relevancy to the lawyer's request.

$$R_k \begin{cases} = 0, & \text{for k outside jurisdiction or time frame of lawyer's request.} \\ = \left[\underset{\text{for i contained in index of k}}{\overbrace{\sum_{i\ 1}^{N} \left(\frac{\sum_{j\ 1}^{N} A_{ij}}{N} \right)}^{\text{Stiles' Number}}} \overset{\text{Lawyer Term Weight Correction}}{M_i} \right] \underset{\text{Anderson's correction for number of index terms}}{\left(\frac{t_k}{T_k}\right)} \overbrace{\left[C_1 \left(\frac{S_k}{\max S}\right) + C_2 D_k + C_3 P_k \right]}^{\text{Lawyer Correction Factor}} \end{cases}$$

Where:

R_k = document relevance number of case k.
A_{ij} = Association factor between search terms i and j.
N = total number of search terms.
M_i = Lawyer's correction to weight of search term i.
t_k = number of terms indexing case k which match the search terms.
T_k = total number of terms indexing case k.
C_1, C2, C_3 = Lawyer assigned coefficients (weights) to the measures of citations, date of opinion, and type of court, respectively.
S_k = number of times case k was cited in subsequent cases.
max S = largest value of S in those cases relevant to the lawyer's request.
D_k = lawyer assigned weight to the time period in which case k was decided.
P_k = lawyer assigned weight to the court in which case k was decided.

Fig. 14 Relevance Number Equation

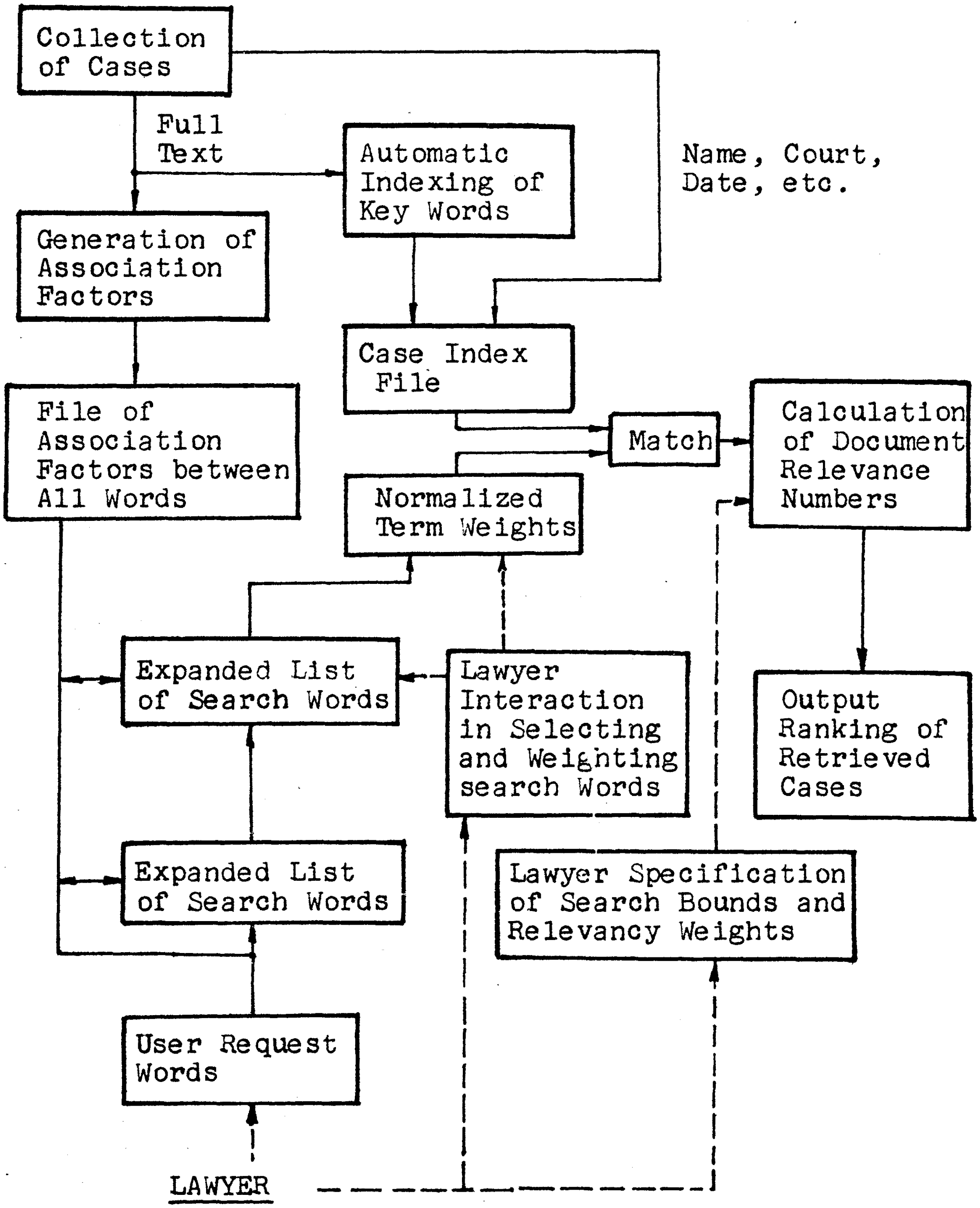

Collection of Cases
Full Text
Automatic Indexing of Key Words
Name, Court, Date, etc.
Generation of Association Factors
Case Index File
File of Association Factors between All Words
Match
Calculation of Document Relevance Numbers
Normalized Term Weights
Expanded List of Search Words
Lawyer Interaction in Selecting and Weighting search Words
Output Ranking of Retrieved Cases
Expanded List of Search Words
Lawyer Specification of Search Bounds and Relevancy Weights
User Request Words
LAWYER

Fig. 15 Proposed Retrieval System

Chapter 4 IMPLEMENTATION OF THE ALGORITHM

Before proceeding to a systems hardware design for the implementation of the algorithm presented in the last chapter, it is important to set forth some guidelines with regard to the implementation. Since this system will be designed for use by practicing attorneys, certain implementation criteria must be satisfied for the system to be successful. This set of criteria will be useful in making various tradeoffs in the systems design.

(1) The system should be fast and direct. A lawyer cannot afford to wait days or weeks while his request is batch processed along with many other requests. Furthermore, in order for the lawyer to interact with the system as specified in the algorithm, the system must be able to respond in "real time." The system should also be direct; that is, the attorney should be able to use the system himself. Use of the system through an intervening analyst would be costly, time-consuming and inefficient.

(2) The system should be easy to use. Lawyers for the most part are "non-technical" people. Since they will be the ones actually using the system, the operator interface should be simple and easy to use. The lawyer must be able to readily operate the system to its fullest advantage without any kind of major orientation/training program.

(3) Although an economic analysis of the system will be made in Chapter 5, the design should reflect the fact that the system is aimed for use by private practicing attorneys. As such, any realistic design must yield a system which is affordable in the legal profession.

(4) The system must retain maximum flexibility. A case law search by subject is only one way in which a lawyer performs legal research. If the lawyer knows the citation of a particular decision, the system should be able to retrieve all other cases which cite that decision. Also, the system should be flexible with regard to different kinds of output. Depending on the needs of the attorney, the system should allow the option of outputing only case titles and citation, or title, citation and headnote, or the full text of the decision.

4.1 SYSTEM ORGANIZATION

Even before beginning a systems design, it is clear that a computer system powerful enough to index and search through two and a half million cases will be large and expensive. Practical considerations of cost will prevent each lawyer from having his own system. This implies that some central computing system must service the requests of many lawyers. In this way, the cost of the system will be distributed among many users to the point where the system is affordable. The requirement that the system be fast and direct implies that each subscribing law firm has to have some kind of remote terminal through which the attorneys may task the system to perform searches, and through which the system informs the attorney of the results of a search. This is all within the state of the art of computer technology. Present day time-sharing systems (computer systems which are designed to handle many users simultaneously) are capable of servicing up to several hundred remote terminals. In most time-sharing systems, the remote terminals are connected to the central computing system through telephone lines. Devices known as modems transform computer information into signals which can be transmitted over telephone lines and transform those signals back into computer form at the other end of the line. Except for the requirement that there be a good quality telephone line connecting them, there is no limit as to how far apart remote terminals may be from the central computing system. Furthermore, multiplexors allow several terminals in the same general vicinity to time share a single telephone line, thereby minimizing long distance charges.

Thus it seems that the most reasonable system organization is to have a large central computing system which actually performs the search algorithm. Each subscribing law firm would have some kind of remote terminal which may be connected to the system via telephone lines. Since the cost of the central system is distributed among many users, one of the design goals is to perform as much processing as possible in the central system in order to minimize overall cost.

4.2 CENTRAL COMPUTING SYSTEM

Although the equations for the association factor, the automatic indexing, and the document-relevance number appear to be complicated, they are actually very simple operations for the computer to perform. The most burdensome task in the computing system is the management and manipulation of the extremely large data files which are required to store all of the necessary information for the retrieval algorithm. It is the choice of the storage media for those files which will dictate the architecture of the central computing system.

Six different files of varying length are needed for the efficient processing of search requests. These six files and their use are briefly summarized as follows:[46]

(1) Dictionary file: As the name implies, this file contains a list of all of the different words used in the collection of cases. Beside each word in the file is a unique numeric code. To promote computational efficiency, the

[46] The structure of some of these files was inspired by a similar effort by Goldblum (op. cit., p. 64-68). Basic improvements include the addition of term association and dictionary files and differences in the content and size of the files.

numeric code will be used in place of the word during most of the program's operation. When a request is entered into the system by a user, the word is "looked up" in this dictionary file and the numeric code is saved for further computation. Similarly, when the system outputs a list of words to the user, the dictionary file translates the numeric codes back into alphabetic words.

(2) Word-association file: This file contains the numeric codes representing the words in the dictionary file and the value of the association factor between all of the words in the file. This file will be used to expand the lawyer's key word request with associated words.

(3) Word-document file: This is the file which indexes all of the cases in the collection. The file contains the numeric codes of all of the words in the dictionary file. Beside each code are document numbers which refer to each case which is indexed by that corresponding word. The expanded list of search terms is compared to the numeric codes.

(4) Case summary file: A summary of each case is contained in this file. Included is information such as the title of the case, cited and citing cases, court, date, etc., plus a very brief capsulized description of the case itself. Parts of this file are presented to the user to enable him to decide whether or not to examine the full text of the case opinion.

(5) Case citation file: The citation of each case in the collection and its corresponding document number is stored in this file. This allows for rapid transition from a case citation directly to the case summary file.

(6) Case opinion file: The full test of the opinion of each case is contained in this file. This is the case law which the attorney ultimately wishes to examine.

The fastest and computationally most efficient manner by which these files can be stored is in the form of random access memory (core or solid-state) directly inside the main computer itself. However, this would be prohibitively expensive. In order for the system to be practical, there must be some trade off between speed of the system and the cost. Whenever there is an enormous amount of data to be stored in a computer system, a common approach is to store large portions of the data in peripheral memory devices which are controlled and accessed by the computer. Figure 16 illustrates the hierarchy of common memory systems. Notice the inverse relationship between speed and cost. In selecting the storage media for the various files in the system, one must fall back on "engineering judgment" to decide which devices represent an effective compromise between speed and cost. In general, files which are short and tend to be accessed fairly often by the computer are stored in high speed memory systems. On

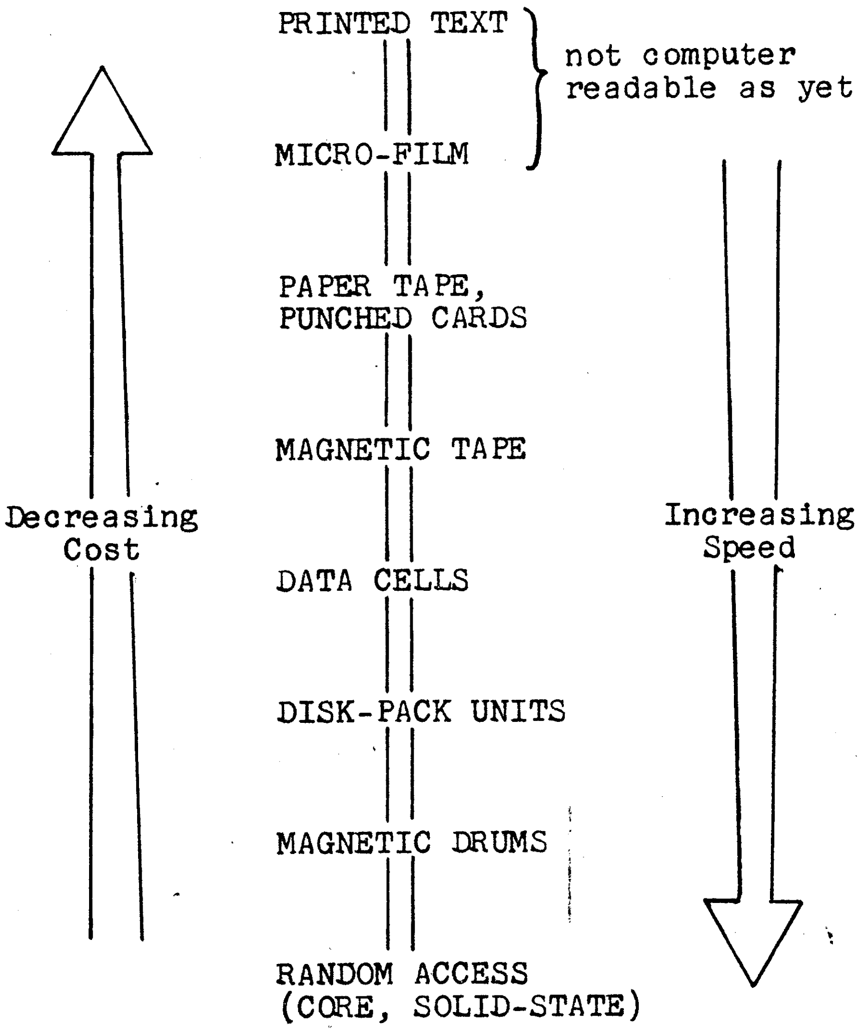


Fig. 16 Hierarchy of Memory Systems

the other hand, long, infrequently accessed files are stored in slower, more economical devices. Long files which are accessed often tend to be stored in medium speed, medium cost devices.

The contents, size and nature of each file will now be examined in order to make some judgments as to the appropriate memory system in which to store the file.

Dictionary file: Each record in the dictionary file contains an alphabetic word with its corresponding numeric code. There are as many records in the file as there are different words in the case collection (with the exception of 100 or so of the most common words). The first problem is deciding how much memory storage to allocate in each record for the alphabetic word. By truncating each word to six letters, the software and storage is simplified, and many grammatical variants are eliminated. However, truncation creates homographs. Based on a thesaurus of 2,400 words, Dennis found that 16% had natural homographs. The truncation of words to six letters created an additional 1.6% homographs.[47] It seems that much more is gained than is lost by truncation. Thus, each record requires six bytes[48] of

[47] Dennis, op. cit., p. 77.

[48] A byte is eight binary (0 or 1) bits of memory. This unit is commonly used for expressing the size of memory systems. (One alphanumeric character is typically stored in one byte of memory).

storage for the alphabetic word. The size of the numeric descriptor code and the total number of records in the file depends on the total number of words. An analysis of almost three thousand cases taken directly from West's Northeastern Reporter yielded about 25,000 different words.[49] Adding a safety margin, the dictionary file will be designed to store 30,000 records. Two bytes of memory storage contains 16 binary bits. These bits can represent numbers from 0 to 2^{15}-1 (roughly 0 to over 64,000). Thus, two bytes seem to be more than enough storage for each numeric code such that each of the different words can be assigned a unique numeric code. The total size of the file is computed as follows:

Record	Alphabetic word	6 Bytes
	Numeric code	2 "
	Total	8 Bytes/Record

Since there will be 30,000 such records in the file, that total amount of memory required is 240,000 bytes. This is a relatively small file compared to other files in the system. However, it only needs to be accessed whenever request terms are entered into the system and whenever the system outputs search terms to the lawyer for analysis. As such, a magnetic disk might be a suitable storage media for the dictionary file. Since the capacity of a disk pack such as the IBM 3330 is approximately one hundred million bytes, the dictionary file would occupy only a very small fraction of the available storage.

[49] Dennis, op. cit., p. 73.

Word association file: One way to look at this file is to envision a square matrix. All of the words in the dictionary file would be listed on each of two right angle axis of the matrix. If one wanted to find the association factor between two words, one finds the row in the matrix headed by one of the words and a column headed by the other word. The intersection of that row and column would contain the association factor between the two words. The Arthur D. Little study performed association experiments using 1000 x 1000 and 3000 x 3000 matrices.[50] If all of the different words were to be used in forming a matrix, this matrix would be 30,000 x 30,000 with a total of 900,000,000 entries. However, not all of these entries are useful. The association factor between a word and itself does not need to be stored. Similarly, a square matrix would contain two entries for each pair of words. This, of course, would be redundant. The association factor between "search" and "seizure" is exactly the same as the association factor between "seizure" and "search." Taking this into account, the 30,000 x 30,000 matrix can be reduced to 449,970,000 entries. This is still too large for a file which must be accessed fairly often. Since most words in general are not related to most other words, the storage requirements for the word association file

[50] NASA Contractor Report, op. cit., p. 7.

can be reduced by only storing those entries whose association factor is larger than some threshold. The danger involved in storing only a partial matrix is that the lost information might produce distortions in the final output ranking. Anderson experimented with the effects of using a threshold cut-off in storing association coefficients and came to the following conclusions:

> 1. The highest document relevance numbers are primarily the result of a few terms with high associativity coefficients rather than several terms with low associativity coefficients and,
>
> 2. The weights of the terms deleted by the cut-off are not significant enough to cause any appreciable fluctuation in the document relevance numbers.[51]

If only entries above a threshold are stored, it would be inefficient to use a matrix format in the file. Instead, the file will consist of a record for each word in the dictionary file. Each record will contain the numeric word code, the numeric code for all associated words (whose association is greater than some threshold) and the corresponding association factor. Assuming that on the average, each word will have twenty other words associated with it, the size of the average record is computed as follows:

Numeric word code	2	Bytes	
Associated numeric codes	40	"	(20 @ 2)
Association factor	20	"	(20 @ 2)
Total	62	Bytes/record	

[51] Anderson, *op. cit.*, p. 10.

Thirty thousand different words implies a total file size of about 1,860,000 bytes. Considering the size of the file, a magnetic drum might prove to be a suitable storage device. Since this file must be accessed fairly often, an appropriate procedure might be to transfer large sections of this file into the computer random access memory during the actual processing.

Word document file: This is the file in which the numeric word codes index corresponding cases in the collection. Thus, each record will contain a numeric code and document numbers of cases which are indexed by that word. Three bytes of storage for each document number is more than sufficient to give each of the two and a half million cases a unique number. Through the manipulation of this number and the extra bits in the three allocated bytes, the address of the corresponding record in the case summary file may be obtained. The size of this file depends on how many cases on the average are indexed by each term. An estimate for this value is obtained in the following way: Assuming that each case is indexed by twenty different words, the total of two and a half million cases would result in fifty million index words. However, since there are only 30,000 different words, each word would index an average of 1,700 cases.

Therefore, the average record size would be

Numeric word code	2	Bytes	
Number of case indexed by that term	3	"	
Document numbers	5100	"	(1700 @ 3)
Total	5105	Bytes/record	

(The number of cases indexed by each term is stored for "housekeeping" purposes and is also used in the initial computation of association factors.) Thirty thousand records yields a file size of little larger than 1.5 x 10^8 bytes. Two IBM 3330 disk packs would have more than sufficient capacity to store this file.

Case summary file: Each record in the case summary file contains pertinent information about each case. This file is used primarily by the attorney in deciding whether or not to examine the full text of the decision. The content and storage requirements of each record is estimated as follows:

Case citation	10	Bytes	
Date, court, state, etc. plus capsule summary	850	"	
Cited cases	200	"	(20 @ 10)
Citing cases	200	"	(20 @ 10)
Total	1260	Bytes	

Since there are two and a half million cases, the total storage required for the case summary file is somewhat less than 3.2 x 10^9 bytes. This is too large to be practically stored in random access memory or magnetic drum systems.

Fortunately, this file is accessed infrequently such that a slower storage device is acceptable. The case summary file may be stored either on magnetic tapes, data cells, or disk systems. As an example, since the IBM 3330 disk pack has a storage capacity of 10^8 bytes, about thirty-two such devices would be needed.

Case citation file: This file is merely a list of the coded citations along with their corresponding document numbers. This allows direct entry into the case summary file when only the citation is known.

Case citation	10 Bytes
Document number	3 "
Total	13 Bytes/record

Thirteen bytes per record times two and a half million cases yields a file size of thirty-two and a half million bytes. This file could easily be stored on a part of one of the system's disk drives.

Case opinion file: Based on the analysis of cases in the Northeastern Reporter from 1959 to 1962, Dennis found that the average length of a case is 1385 words.[52] Assuming that the average word contains seven letters, it would require almost ten thousand bytes to store the text of one case. This implies about 2.5 x 10^{10} (twenty-five thousand million) bytes are needed to store the full text of all of the reported

[52] Dennis, op. cit., p. 72.

cases. Aside from the initial indexing, the central computer system does not "use" the full text--only the lawyer needs it. This fact coupled with the extremely large amount of necessary storage has prompted the decision to store the text locally at the remote terminals in the form of microfilm. How this is done will be explained in the next section.

In addition to the memory devices needed to store the above described files, the only other peripheral equipment needed in the central computing system are multiplexors and modems through which the system may communicate with remote terminals. The choice of the actual computer is not particularly critical. It should, however, be large enough to:
(1) control all of the peripheral storage devices, and,
(2) execute large time-sharing programs involving many users. One of the larger IBM system 370 mainframes would be a suitable choice.

4.3 REMOTE TERMINAL

At the remote terminal, a lawyer has to be able to task the central computing system to perform a search, interact with the system during the search, and receive the search results. He also has to read the opinion of the cases retrieved to obtain the actual case law.

One way he could do this would be to use the computer retrieved citation to directly find the opinion in a volume of West's Reporters. This is undesirable for two reasons. First, most law firms do not have a complete West system. Thus, the research would be delayed until the lawyer went to a law library. Second, it seems almost unreasonable that a computerized system powerful enough to take a few key words, search through two and a half million cases and come up with the citations of the most relevant decisions, still requires that a lawyer wade through a massive jungle of thick, leather-bound volumes to read case opinions. In the last section, it was shown that it would be impractical to store the full text of the decisions in the central computing system. As an alternative, microfilm cartridges have been chosen as the storage medium for full text in this system. The Eastman Kodak Company manufactures a product line which seems to be particularly suited to this application. The storage medium is a microfilm reel packaged in a plastic cartridge. Using 16 mm film, a standard 215 foot film reel with a 50 to 1 photographic reduction and dense packing of frames could store the equivalent of about 20,000 8½ x 11 pages of text. Assuming case opinions average five pages of printed text, only 625 microfilm cartridges are required to store the decisions of all two and a half million reported cases. Since each cartridge measures 4" x 4" x 1", the volume occupied by the 625 cartridges would be less than that of a standard four-drawer file cabinet. As a side note, microfilm available today is archival in quality; i.e., it has a lifetime of hundreds of years.

The microfilm cartridges are plugged into a microfilm reader illustrated in Figure 17. This reader has two very important features. If the reader is provided a digital signal corresponding to a frame number on the reel, the reader will automatically advance the film such that the desired frame is projected onto the screen. This takes less than six tenths of a second. In this system, the microfilm reader will be automatically controlled so that all the lawyer has to do is snap in the correct cartridge. The other important feature of the microfilm reader is its ability to make paper copies of the information displayed on the screen. The electrostatic copier located on the bottom of the reader generates copies much in the same way that standard office copiers operate. This allows the lawyer to make a hard copy of the decision without transcribing the text from the screen or finding the same opinion in a printed form.

Lawyer communication and interaction with the central computing system is accomplished through a TV-like alphanumeric display and keyboard such as the one shown in Figure 18. The lawyer uses the typewriter-style keyboard to enter his search words, thresholds, and citations. As the keys are depressed, the letters will appear on the display

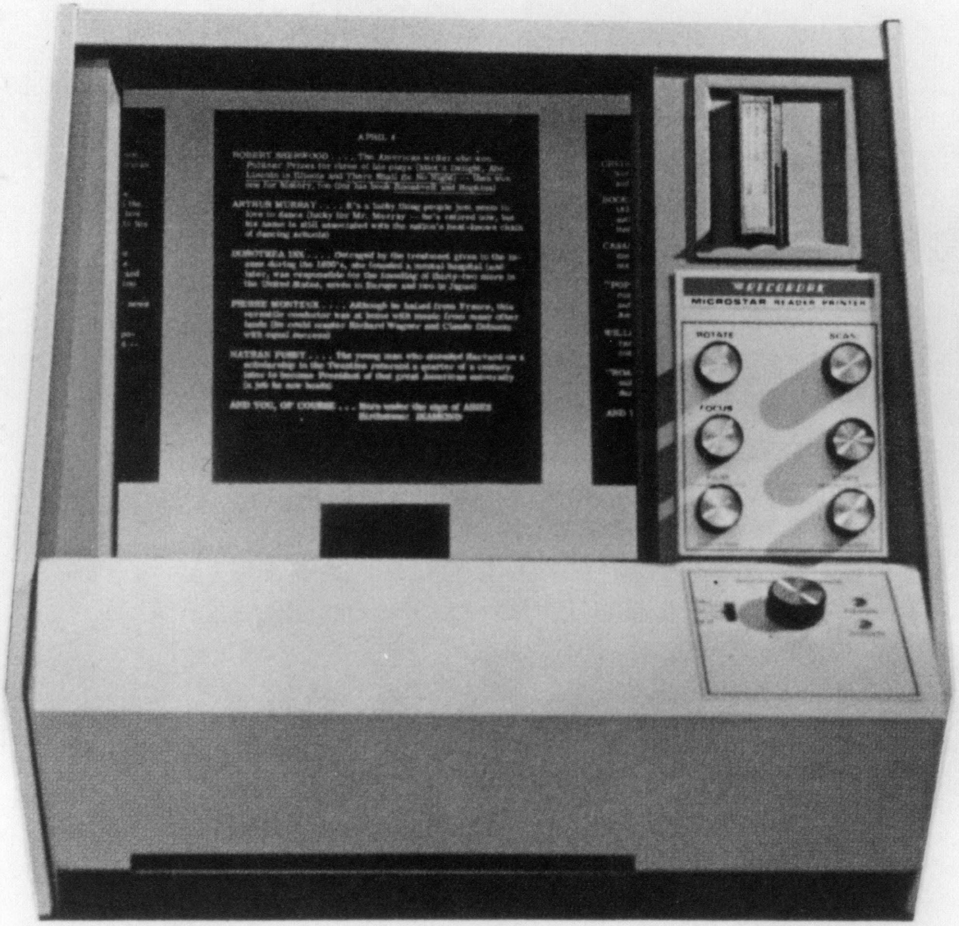


Fig. 17 Micro-film Reader/Electro-static Copier

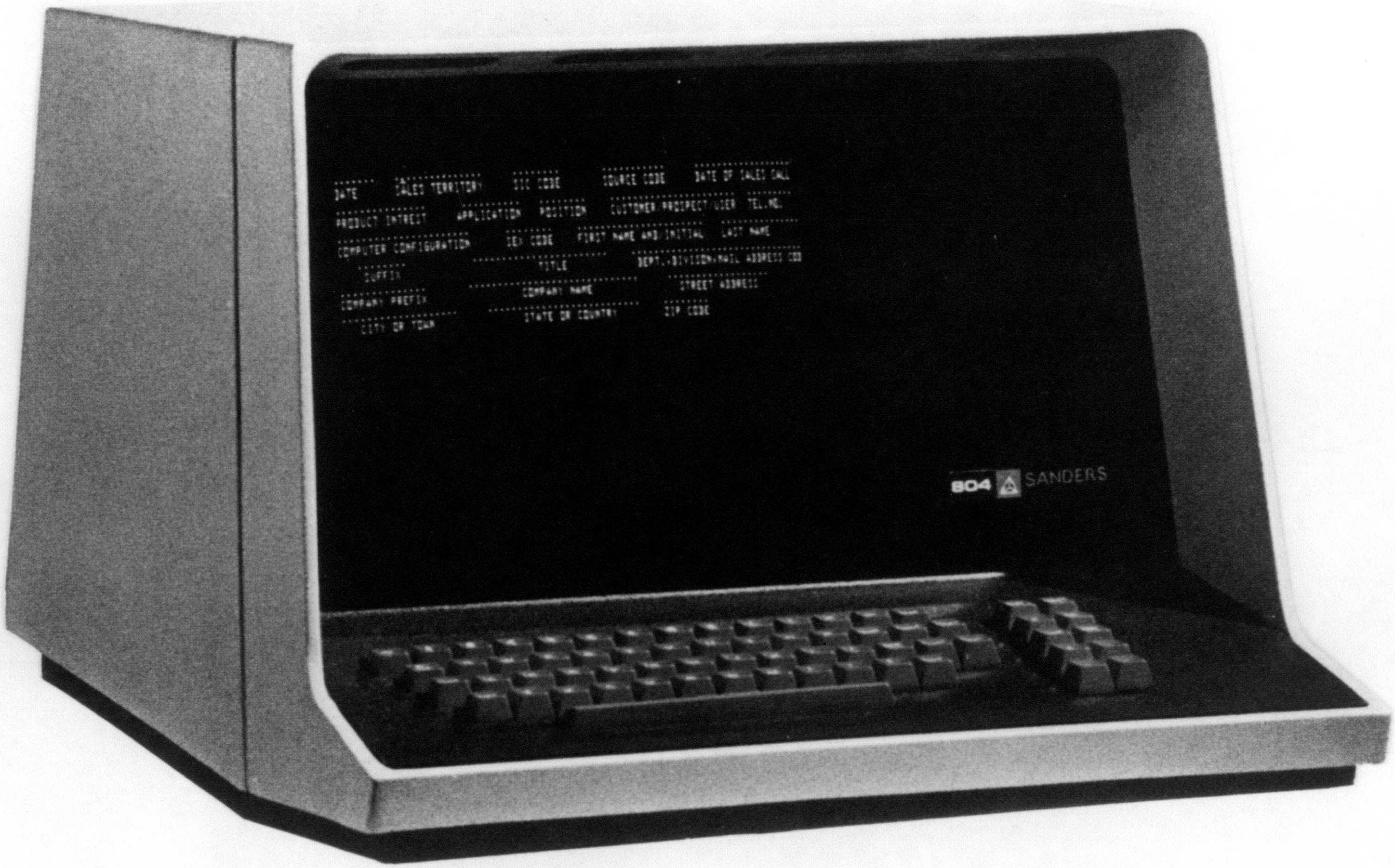


Fig. 18 Remote Terminal Display/Keyboard

screen. Once the request has been properly composed, the lawyer will depress one of the control keys which will cause the data on the screen to be transmitted to the central computing system. Search results from the central computer are also presented on the lawyer's display screen.

The display/keyboard terminal chosen for this system contains an internal "mini-computer" or microprocessor. This microprocessor acts like a small computer. It is used for formatting the communication from the lawyer to the central computer and vice versa, and also for controlling the display, keyboard, and microfilm reader.

The final piece of equipment needed to complete the attorney's remote terminal is a modem. As mentioned before, a modem is a device connecting the lawyer's terminal to a standard telephone. This device translates computer information into signals which can be transmitted across telephone lines, and also translates incoming signals into computer compatible form. Since the amount of data being transmitted between the remote terminal and the central computing system is minimal, any standard low cost modem would be acceptable.

A block diagram of the lawyer's remote terminal is illustrated in Figure 19.

4.4 SYSTEM OPERATION

The operation of this computerized case law retrieval system is now described from the point of view of the attorney

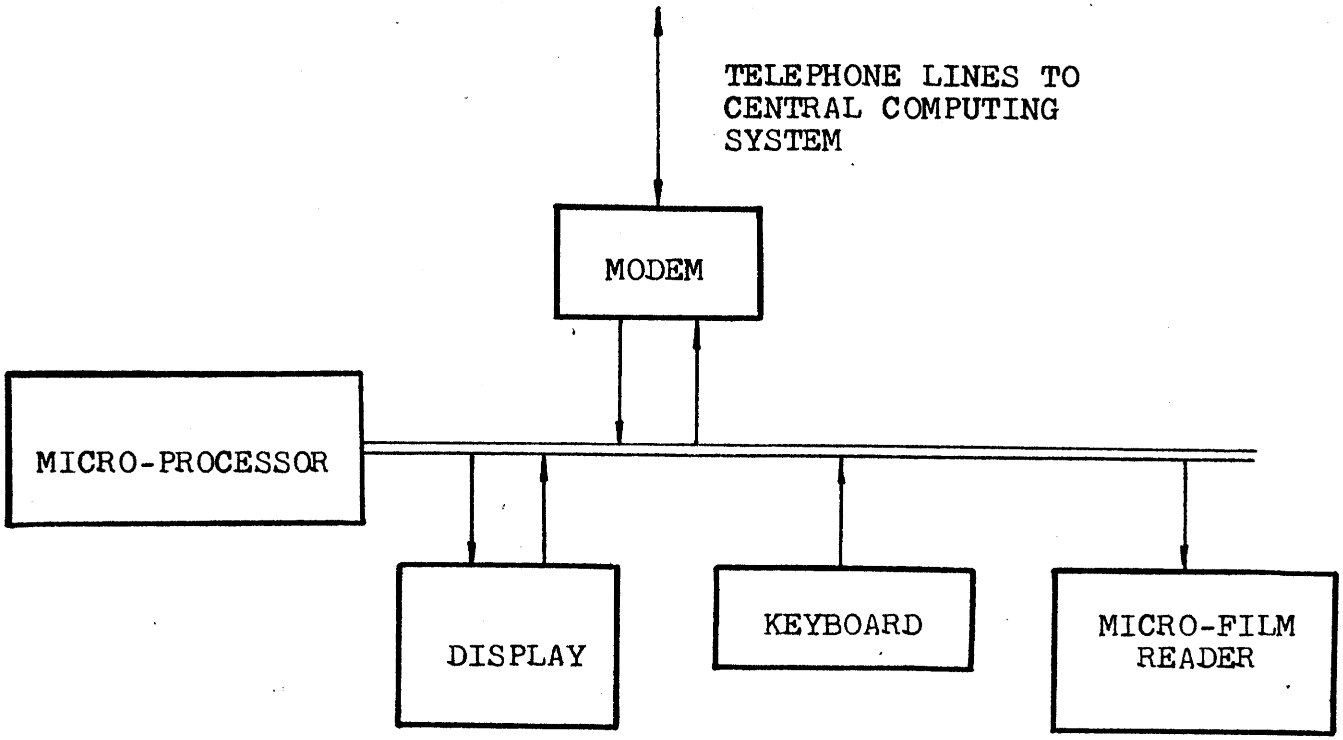


Fig. 19 Block Diagram of Remote Terminal

using the system. After the attorney turns the terminal on, he will dial into the central computing system much in the same manner as one dials a long distance telephone call. Once the connection between the terminal and the main computer has been established, the computer will present a series of questions on the display screen in order for the attorney to "log-in." These questions include the lawyer's name, name of the firm, password, and billing account number. The purpose of this log-in procedure is to make sure that an authorized subscriber is using the terminal and that charges for computer time are correctly billed. After the lawyer has successfully logged into the system, he is asked what kind of search he wishes to perform (i.e., a search by subject, citation, case title, etc.). The lawyer will respond by typing the appropriate type of search and depressing the RETURN key. Since it is the most complicated, a subject search will be examined here. The lawyer, therefore, types "SUBJECT" and hits the RETURN key. The system responds by presenting a format on the display screen which the lawyer can use to specify system bounds and thresholds. Using the keyboard, the lawyer will enter onto the format the jurisdictions of interest (such as MASS, or MASS + NY + NH, or ALL), the dates of interest with a weight (0 to 10) indicating the relative importance of that time period (1960-1973,10; 1930-1960,8; BEFORE 1930,0), and the courts of interest also with a 0 to 10 weight (SUPREME COURT,10; etc.). Next, the lawyer must make a judgment as to the relative importance of the number of times a case was cited, the date of the decision, and type of court in the computation of case relevancy. For example, the lawyer might feel that the date of a case is much more important to his needs than the type of court. He would, therefore, give date a higher number than the other two measures. Alternatively, he might chose to give all three measures the same weight. Once the form has been completed, he will again depress the RETURN key. The main computer will transfer all of the information on the screen into temporary storage for future use. The lawyer is now asked to enter his key word request. The format of the search request is identical to that of a KWIC system--a series of key words connected by logical operators. For example, if the lawyer is interested in laws regarding protective equipment for motorcyclists, he might type MOTORCYCLE AND (HELMET OR WINDSHIELD OR GOGGLES). When he is finished typing, he will hit the RETURN key. The main computer will read each term from the screen and "look it up" in the dictionary file to obtain the corresponding numeric code. These codes are used to access the word association file and find the codes of associated words. The codes of these words are used to find other associated words (i.e., two associative expansions). This expanded list of numeric codes is ordered based on normalized associative term weights, translated back into alphabetic words via the dictionary file, and presented on the attorney's display screen with corresponding weights. The lawyer must now analyze this list of words and weights and decide if any are to be added, deleted, or modified. After he makes the corrections, he has three options. Hitting the INITIALIZE key will recycle the system to the point where he is asked to specify key words. Hitting the RETURN key will cause the system to perform another word expansion on the list of corrected words on the display. It is only when the SEARCH key is depressed that the system reads the search words and weights on the screen and begins the retrieval process. Using the dictionary file, the central computing system translates the words back into numeric codes and accesses the word-document file. Using the document relevance number formula, the system computes the relevance of all the cases indexed by the search words, ranks the cases in order of relevancy, and stores them in a temporary file. The lawyer is then informed of how many cases there are in the temporary file and is asked how many cases he wants outputed and in what form. The lawyer may answer this by typing (CIT + SUM,5, CIT,15) which would mean the citations and case summaries of the five most relevant decisions and just the citations of the next fifteen most relevant cases. Or, he could type any other combination which he desires. If he wishes to examine the actual decision of a case, the lawyer will position the cursor on the citation of interest on the screen and depress the TEXT key. The system will respond by presenting on the screen the number of the microfilm cartridge on which that decision is stored. The lawyer snaps that cartridge into the reader, and the system will automatically advance the film until that particular case is projected on the screen. The lawyer can now read the decision and manually advance the microfilm. If he wants a record of the decision the lawyer can hit one button and the microfilm reader will produce an electrostatic copy of the information on its screen. At this point, the lawyer can obtain additional cases retrieved by the system, enter a new key word request, or go back and start the entire process all over again. An operational overview of the remote terminal is illustrated in Figure 20.

4.5 INITIAL FILE GENERATION

The most severe implementation problem with this system is the initial generation of the system files. Actually, this is really a cost problem which may be reduced by the development of improved technology.

The decisions of all the reported cases are presently stored in printed form. Therefore, before any automatic indexing and searching can be performed, the text must be

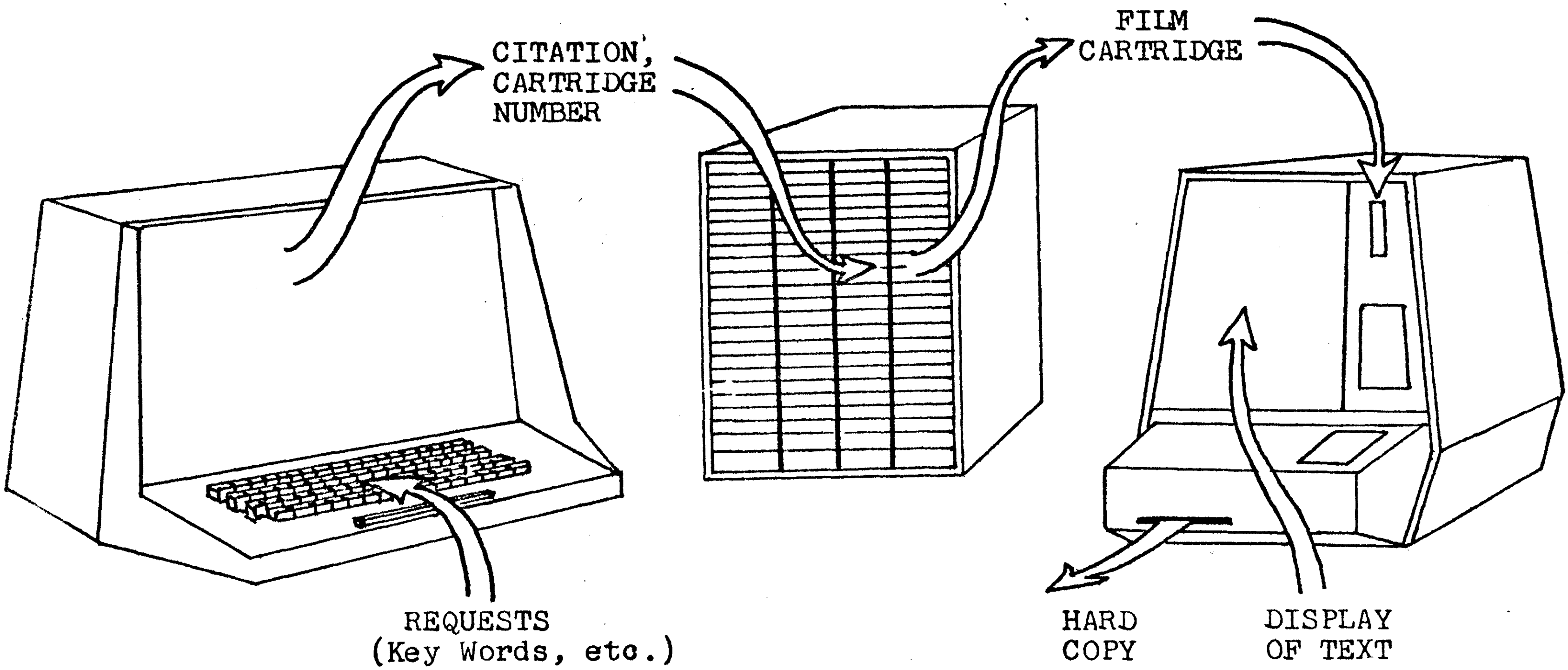


Fig. 20 Overview of Remote Terminal Operation

translated into a form which the computer can accept. Up until a few years ago, the only reliable methods for transforming printed data into computer readable form involved human intervention. This usually meant an operator would read the text and type it on the keyboard of a data entry system. The data entry system would produce punched cards, paper tape, magnetic tape, or some other media which a computer could read. The high labor costs associated with the data entry of large files is one of the primary reasons why many computerized case law retrieval systems never advanced beyond small laboratory experiments. Assuming that each case could be keypunched, checked and corrected in the average time of one hour, then an army of 1200 keypunch operators working forty hours a week would require over one year just to transcribe the printed text of all of the cases into computer readable form. At four dollars an hour, this would amount to a total cost of over ten million dollars just for the direct keypunching labor. This does not even include items such as the data entry system, storage and office space.

In order to counter the errors and high costs attributable to human data entry, a number of companies (including IBM, ECRM and Compu-Scan) have recently developed optical scanners. An optical scanner is a device which uses a light source, optical fibers, and light sensors to transform light and dark sections of a printed page into electrical impulses which are entered into a computer. By focusing the scanner on one printed character, the computer can analyze the light and dark areas via the elecurical pulses and determine what kind of character it is. By noting the separation between characters, the computer can isolate words and sentences as it "reads" the text without any human intervention. Since optical scanning is a relatively new technology which is still undergoing intensive development, the optical scanning equipment available today is somewhat limited with regard to speed, type of printed material which can be read, and error rates.

The typical scanning speed of the optical scanners surveyed is about 150 characters per second. Since the entire collection of case law contains an estimated 2.5×10^{10} characters, one optical scanner operating 24 hours per day would require approximately 5.3 years to read all of the cases. Using ten scanners operating in parallel would reduce this to a little over six months. Since the cost of these units ranges from sixty to a hundred and fifty thousand dollard each, the cost of ten scanners could exceed one million dollars.

The second major limitation of these devices is that at present, only certain forms of printed material can be read automatically. Specifically, high speed scanning can only be realized with standard, evenly spaced characters.

Examples of this kind of text includes material typed by most standard typewriters, computer printouts, and offset printing of typed material. Irregularly spaced characters such as one would find in most printed books, newspapers, material typed by IBM "Executive" model typewriter, etc., cannot be optically scanned reliably at the present time. Most scanners operate by focusing a viewer on an area the size of one character. When characters are irregularly spaced, only part of a character or perhaps more than one character will occur in the field of vision, thereby "confusing" the computer.

This limitation would prevent direct scanning of cases in their present form. The characters in the text of cases are irregularly spaced so that the margins in the printed volumes of the Reporters will be straight. Thus, in order to use optical character recognition to transform printed cases into computer acceptable form, two possible courses of action are available. The first is to wait the projected two to five years until scanners are perfected which can read irregular text such as that found in most books. The second alternative is to transcribe the cases into media which can be optically scanned. One such approach would be to type the cases on a standard typewriter. Typing at 60 words per minute, over 900 typists working 40 hours per week would be required to keep up with the ten optical scanners. The cost of this direct typing labor would amount to about three million dollars.

Error rates of optical character recognizers is the last problem. Using the typical performance figure of one error in 250,000 characters, optical scanning of all the cases would produce an expected one hundred thousand character errors. These errors could either be corrected by a proofreader after the cases have been scanned, or they can be left until the cases are automatically indexed. Since character errors in words will tend to yield unique "words" (for example, an error in reading "action" might produce the pseudo-word "actlon"), words containing character errors would be extracted by the automatic indexing as "informing" words. This would greatly speed up the proofreading process in that the proofreader needs only to analyze the list of informing words and their corresponding cases. If manual transcription of the cases is necessary prior to optical scanning, then the human errors in typing would overshadow the errors in automatic reading.

Once all of the text is in computer compatible form, there is the time and cost of entering all of the data into the system and performing the necessary computations and manipulations in order to generate the required files. Perhaps an example will serve to illustrate the magnitude of this problem. After the word-document file is created, the number of documents indexed by each term is used to calculate association factors. Although only those association factors above some threshold will be preserved, all possible factors have to be computed to determine which ones in fact are above the threshold. As was mentioned in a preceding section, a vocabulary of 30,000 different words results in about 450 million different possible pairs of words, and hence, 450 million association factors. If a computer required 100 milliseconds (one tenth of a second) to fetch the necessary data, perform the computation, and store one association factor, then the computer would have to operate continuously, twenty-four hours a day for almost a year and a half in order to compute all of the association factors needed for the word association file. This time, of course, could be reduced through the use of several computers and parallel processing.

All of these problems are really "one-time" problems in the sense that they are non-recurring. Once the system is established and all of the files have been generated, the use of the system and the updating of the files pose no major technical difficulties. Periodically (perhaps once a month), the files could be updated with all of the recent decisions. This could be done at night so as to minimize any possible disruption of service. Unlike previous case decisions, much of the printed material today is composed by computerized systems. Computerized photocomposition is a process by which a text is stored on magnetic tapes and an output device creates a copy of the text. The copy and the tapes are edited until the text is in final form. A final output copy is made which is then used in offset printing. By this method, not only is the document printed, but a magnetic tape of the text is produced as a byproduct. This tape can be directly read by a computer, thereby eliminating the need for either keypunching or optical scanning. S. J. Skelly has proposed the system illustrated in Figure 21 as an efficient method by which legislative bills may be published.[53] A similar process could be instituted for publishing case decisions.

One of the advantages of the computerized case law retrieval system is that it is so easy to update. As new words are created (such as the word "computer" a few decades ago), they are entered into the files along with their association factors. Since the system is not dependent on any kind of subject index, the development of new legal issues and concepts does not necessitate complete reorganization. As the body of case law grows, so does the data base.

The cost of system initialization and set up cannot be readily determined without a more detailed design of the system structure. However, based on the above analysis, this

[53] S. J. Skelly, "Computers and the Law," Saskatchewan Law Review, XXXIII (Fall, 1968), p. 170.

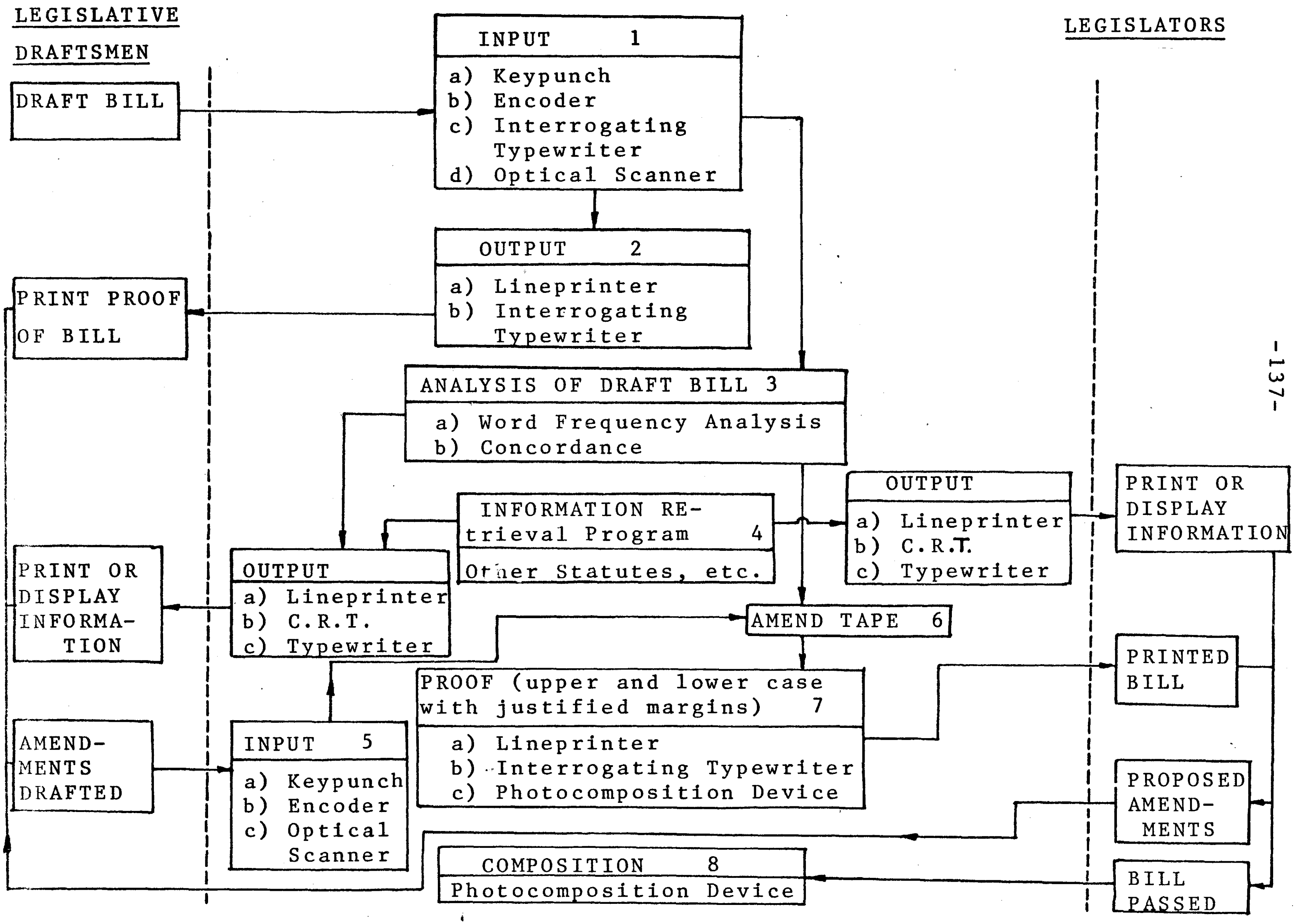


Fig. 21 Computerized Bill Processing

cost could easily run into the tens of millions of dollars. It would not be unreasonable to expect that funds for such an effort originate from the public sector. Society as a whole would benefit by the development of this kind of technology. Aside from the direct benefits of having "better" law, the document retrieval algorithm will have applications in many other fields of endeavor.

4.6 TECHNICAL PROBLEMS

Aside from the problems of creating the initial files, the only other technical problems foreseen at this time involve word indexing. Specifically, there are two problems: (1) phrases, and (2) numbers. Taken by themselves, the words "cause," "of," and "action" have little informational value and would, therefore, not be used to describe or index a document. However, the phrase "cause of action" does have informational content which would be lost in the proposed system. The inclusion of phrases would create many difficulties. First of all, the dictionary file would have to be expanded to include all "informing" phrases. It is not known just how many informing phrases there are in the body of case law, but there is sure to be a large number of them. This, of course, increases the storage requirements and the complexity of the system (remember that the number of association factors which need to be calculated is proportional to the square of the number of words or phrases). Also, present computing systems cannot automatically extract phrases from expository text. Thus, human analysts, with all of the associated errors, costs, and inefficiencies, would have to be used to isolate phrases in the case opinions. The other technical problem is that of numbers. Should they be treated the same as words? Are numbers "informing words"? If so, the total number of different words in the files would also have to be greatly increased.

Before any computer system is developed, it is recommended that a thorough study be made of the potential difficulties arising from phrases and numbers in an association factor system.

Chapter 5 ECONOMIC ANALYSIS

The costs of the computerized case law retrieval system are broken down into three areas in order to perform economic analysis: (1) one-time costs of system set up, (2) microfilm library costs, and (3) law firm operating costs.

The costs of the development of such a system and the creation of the necessary data files were discussed briefly in the previous chapter. A reasonable cost estimate of the total effort would be in the tens of millions of dollars. It is hoped that the social importance of document retrieval technology will prompt the government to underwrite a major portion of the initial costs. The Federal Government should have a particular interest in these kinds of systems since the U.S. Government is probably the world's largest producer of documents and information. Although twenty or thirty million dollars seems like a lot of money, it is an exceedingly small fraction of the current government expenditures. It is often said that the high value of human life is a justification for the expenditure of hundreds of millions of dollars in medical research. This author concurs with Fredrick E. Smith when he said, "I submit that the benefits of protecting the rights of people, and providing equal access to the succor of the laws is...[an ample] justification for providing optimum legal information services."[54]

The cost of the microfilm library is analyzed separately because the size of the library is variable depending on the particular needs of the firm. The cost of each 215 foot microfilm cartridge is about twelve dollars (including processing). Thus, an entire collection of cases (2.5 million) consisting of 625 cartridges would cost about seven and a half thousand dollars. Presumably, if such a computerized system were to be made operational, the economies of scale in producing all of the microfilmed cases would reduce this cost. The cost of a complete microfilm library can be justified based on the savings which would accrue from (1) using microfilm instead of purchasing printed volumes, and (2) the reduction in office space needed for the law library. One microfilm cartridge contains 20,000 "pages." Since 20,000 pages of bound, printed text would certainly cost more than twelve dollars, the marginal cost of using microfilm is less than that of printed text (the cost of the microfilm reader is included under operating costs). Storing an existing law firm library on microfilm is cost effective based on the reduction of space necessary to store the library. For any given law firm, a quick calculation of the cost of floor

[54] Fredrick E. Smith, "Computer Applications to Legal Documentation: What is Not Being Done," Law Library Journal, LXIV (May, 1971), p. 114.

space and size of library could produce the answer as to how long it would take for the microfilm to pay for itself.

It is difficult to perform an accurate cost-benefit analysis on the law firm operating costs of such a system. First of all, the benefits of the system will accrue to many different people--lawyers, clients, the public at large, courts, etc. Secondly, since the system is not yet implemented, it is hard to estimate the potential improvement in law. Lastly, even if the improvement were known, it is difficult to place a dollar figure on the value of "better law." Because of these problems, a very narrow (conservative) view will be taken in performing a cost-benefit analysis. It will be assumed that with a computerized system, a lawyer will maintain the same quality of case law research as was performed manually, only he will be able to do it faster. Thus, the only benefit that will be considered is the saving of lawyers' time.

The cost side of the picture has several components. It is assumed that the nonrecurring system design and development costs will be funded by the government and that the cost of the microfilm library is justified on the basis of savings in book purchases and reduced office space. Thus, the benefits of saved time must be balanced against the operating costs of the central computing system and the cost of each law firm's remote terminal. The operating cost of the central computing system is a variable cost which is distributed among many law firms. As such, it seems reasonable that each law firm's contribution to these costs be a function of how much that firm uses the system. As is the case with most time-sharing systems, the law firms will be billed at some hourly rate for the amount of time that the terminal was connected to the central system. The remote terminal itself is a fixed cost for the law firm. Either the terminal is rented, in which case there is a monthly rental charge, or it is purchased, in which case there is an annual contribution to the purchase cast of the terminal.

The value of the lawyers' saved time is now compared to the fixed and variable operating costs in order to perform a cost-benefit analysis. Since the basic premise is that the lawyer will perform case research faster with a computerized system, the fraction r ($0 \leq r < 1$) will be used to denote the fraction of time needed to perform automated case research. For example, if the lawyer spent H hours per week performing case research manually, he would only need to spend rH hours with an automated system. The mathematical formulation for the benefits (value of saved time) and the costs (both fixed and variable) for a law firm operating an automated system for one year is shown in Figure 22. If the costs are set equal to the benefits, a "break-even" analysis can be performed to determine at what point the system just

$(52)\,N H S\,(1-r)$	BENEFITS - yearly value of saved time.
F	FIXED COST
$(52)\,N H C r$	VARIABLE COST

where: N = Number of lawyers in the firm.

S = Average hourly wage of the lawyer.

H = Average number of hours per week spent by each lawyer in performing case law research manually.

C = Cost per hour of using the central computing system via remote terminal.

r = Fraction of time needed to perform case law research with an automated system.

F = Yearly contribution to fixed costs.

BREAK-EVEN ANALYSIS

Benefits = Variable costs + Fixed Cost

$$(52)\,N H S\,(1-r) = (52)\,N H C r + F$$

$$H = \frac{F}{N\left[S\,(1-r) - r C\right] 52}$$

Fig. 22 Break-even Equation

"pays for itself." By solving this equation for H, one can determine the average number of hours per week each lawyer in the firm must spend in manual case research before the system has a benefit cost ratio greater than one.

The estimated fixed costs of the remote terminal are itemized in Figure 23. If this total cost of $25,000 is amortized over six years, the law firm's yearly contribution of fixed costs (F) would be about $4,200. For the purposes of this analysis, the value of a lawyer's time is estimated to be $15 per hour (which corresponds to about $31,000 per year). Although this is considerably less than the typical "billing rate" of law firms, it is used to reflect the fact that most research is performed by lower paid, junior members of law firms. Since the break-even equation in Figure 22 is being solved for H, the only other variables which must be specified are C, r, and N. The hourly cost of remote terminal connection to the central computing system (C), cannot be accurately calculated without knowing the exact price of the main system and the number of law firms which will subscribe to the service. However, since the cost of present day commercial time-sharing systems averages about $25/hour, the break-even analysis will be performed for different values of C in the range from $15 to $35 per hour. Similarly, the value of r, which indicates how much faster an automated system is compared to a manual system, cannot be

ESTIMATED FIXED COSTS

Micro-film reader and electro-static copier	$ 9,000
Alpha-numeric display and keyboard console	3,000
Micro-processor	6,500
Modem unit	500
Subtotal	$ 19,000
Installation, cabling, enclosures, etc.	1,500
Manuals, training, etc.	500
Maintenance	4,000
TOTAL COST	$ 25,000

Fig. 23 Remote Terminal Fixed Costs

determined until after the system is built and tested. Therefore, the computation will be performed across a range of values for r. Lastly, N, the number of lawyers in the law firms will vary from firm to firm. A computer printout of the bread-even value for H for various values of C, r, and N is contained in Appendix II. Some of the data from Appendix II has been used to generate the curves in Figure 24 which show the break-even point as a function of H and r for different values of C in a four man law firm. A Missouri Bar study indicates that 16.7% of a lawyer's time is normally spent performing legal research.[55] This amounts to about 6.7 hours a week. From the curves in Figure 24, an H of 6.7, a C of $25/hour requires an r of about .3 in order for the system to break even in a four man law firm. An r of .3 implies that a lawyer with an automated system must be able to perform legal research 3.3 times faster than he could manually. An r of .3 seems to be well within the capability of the proposed system. If the system is faster or if more time is spent in case research, or both, the benefits will exceed the costs. This kind of analysis can be performed for different values of C and for different size law firms.

[55] Morris Cohen, "Research Habits of Lawyers," Jurimetrics Journal, IX (June, 1969), p. 191. [This number compares favorably with a similar study conducted in Canada (which is also a "Common Law" county) which estimated the percentage to be 21% (reported in The Ottawa Citizen, December 9, 1972, p. 9)].

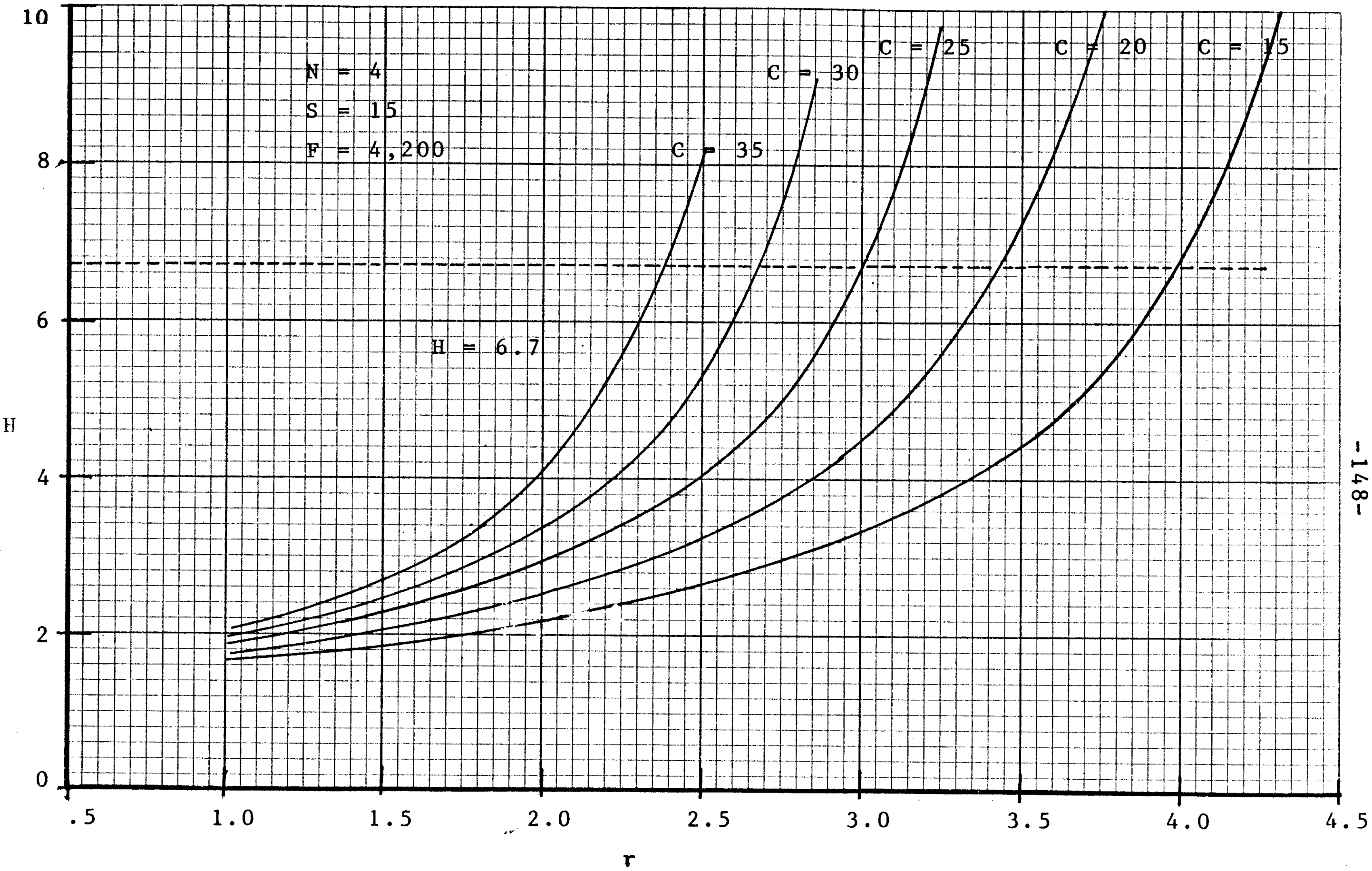


Fig. 24 Break-even Curves

Although the cost-benefit analysis of law firm operating costs for the automated system already appears favorable, it should be noted that the above analysis was conservative in the sense that it assumed that the automated system was no better than manual operation--only faster. The system's potential for improving the quality of legal research would tend to tilt the balance on the benefit side. Another possible benefit of the system's remote terminal is its potential use in the data management of the law firm's normal operations. By adding some additional memory and a line printer to the microprocessor, display and keyboard configuration, the remote terminal could provide services such as client billing keeping accounts of attorney time, recording cash flows, drafting standard legal forms such as wills, and payroll processing.[56] The variety of law firm management functions which a computer terminal could perform once it is installed tends to improve any cost-benefit analysis made on the automated case law retrieval system.

In summary, the economic analysis argued for government funding to initially develop the computerized case law

[56] See Boris Ellison, "The Computer as an Economic Solution to Law-Office Record-Keeping Problems," Jurimetrics Journal, X (June, 1970) or F. Patmon, "Total Systems Approach to the Practice of Law," Law Office Economics and Management, XI (February, 1971), p. 501.

retrieval system. Once established, the above cost-benefit analysis indicates the operating costs of the system are more than offset just by the savings in lawyers' time. The additional benefits of improved research and law office management are further justifications for the computerized system.

Chapter 6 LEGAL AND BEHAVIORAL PROBLEMS

6.1 LEGAL PROBLEMS

Literature on the general subject of computerized case law retrieval systems suggests that there are two potential legal problems associated with such systems. These are (1) copyright laws, and (2) the unauthorized practice of law.

Although the decisions contained in the cases themselves are not copyrighted, most of the supplementary materials such as headnotes, synopses, and lists of citations found in legal reference systems such as West's are subject to copyright protection. The way in which the automated system is designed, supplementary case material such as headnotes and lists of citations are included in the case summary file. Certainly, one could start from the actual cases themselves and extract all of the citations and other material needed for the file, but this would be an enormous waste of time, money, and effort, given that this has already been done by commercial publishers. It would seem that the most reasonable approach would be to negotiate with the concerned publishers for the use of needed copyright protected material.

The question of whether the use of an automated case law retrieval system by laymen constitutes unauthorized practice of law has been raised many times. Attempts at just defining what "practice of law" actually is has led to controversies, ambiguities, and inconsistencies. With regard to unauthorized practice and computer systems, the American Bar Association Committee on Unauthorized Practice has made the following statement:

> a retrieval system is unobjectionable so long as it is merely a means of storing textual information for later retrieval. In that respect, it is similar to a library. So long as it is a library, there would appear to be no unauthorized practice of law problems present. When, however, the system becomes so sophisticated that facts are fed into it from which the system draws legal conclusions based on specific legal analysis, it would involve the practice of law.[57]

The ABA further states that if the system does "practice law," then its use must be restricted to lawyers (members of the Bar) and the use of the system by laymen would constitute unauthorized practice. In light of the above statements, how does the proposed system stand with regard to unauthorized practice? The system accepts key words (facts?) and outputs cases believed to be relevant. Is the output of relevant cases in some sense a "legal conclusion"? This is a debatable question. In the opinion of this author, since the computer does not use "specific legal analysis" but, rather, lawyer-supplied words, weights and the statistical relationships

[57] "Computer Retrieval of the Law: Challenge to the Concept of Unauthorized Practice?" University of Pennsylvania Law Review, CXVI (May, 1968), p. 1273.

between words, it cannot draw any kind of "legal conclusion." Thus, the availability of a computerized case law retrieval system would not constitute any more of a threat to unauthorized practice than the existence of the West system in a public library. There is also the following side issue. If computer case retrieval systems are implemented and if they demonstrate a significant improvement over manual search systems, then the popularity of these systems could result in a decline in the availability of manual index/search systems. "If such a situation comes to pass, exclusion of laymen from automated retrieval systems may be a denial of their right to have access to 'the law.'"[58] This might present a problem if, in fact, laymen were excluded from the use of such systems via unauthorized practice. Barring such a grave error, it is envisioned that if manual index systems in the legal area were to be discontinued, libraries would contain microfilmed decisions available for all to read and would also have a remote terminal which would be available for searches at some reasonable fee.

6.2 BEHAVIORAL ISSUES

Perhaps the greatest obstacle to the implementation of computerized case law retrieval systems is gaining the acceptance of practicing attorneys. With the exception of

[58] *Ibid.*, p. 1283.

the typewriter and the Xerox machine, the legal profession has remained immune to virtually all technological advances since the invention of the printing press. In the medical profession, large hospitals, research clinics, and universities engage in medical research. There are no such counterparts in the legal profession investigating advanced legal methods. The daily work of physicians involves new medicines, new equipments, and techniques. Lawyers, on the other hand, use the same tools of trade that they have always used. Certainly the nature of the law is dynamic and changes constantly, but in interpreting the law and applying it to problems at hand, the lawyer has no more resources available to him than he had fifty years ago. In fact, because of the increased number of laws and cases, his job is more difficult now than it was fifty years ago. The need for stability in law has led to the reliance on traditions and precedents as expressed in concepts such as "judicial restraint" and stare decisis. Perhaps this kind of philosophy accounts for some of the perceived reluctance to change legal research methods.

Lawyers, for the most part, are not technically oriented. Except for its associated legal problems, technology is not a part of a lawyer's daily work as it is, for example, with physicians. Lawyers don't understand (and, therefore, don't trust?) computers. The greatest interaction between lawyers and computers is probably in the form of monthly credit card bills--this already puts the digital computer in a bad light.

When asked to react to a computerized case law retrieval system, the following remarks are typical of those solicited from practicing attorneys:

> "I don't want a computer doing my job for me."
>
> "I cannot program computers."
>
> "Computers are too expensive."
>
> "I don't need a computer, I get paid for performing legal research."
>
> "Current case law research is a good way to break in new lawyers."
>
> "I don't trust computers."[59]

The first remark, "I don't want a computer doing my job for me" almost sounds as if the attorney is afraid that the legal profession will be dissolved and replaced by an electronic wizard. Certainly, the computer will do some of the work which is currently being done by the lawyer, but what kind of work is it? The computer will perform the clerical, mechanical tasks of searching through millions of cases and retrieving some appropriate documents. The attorney is still absolutely essential in the characterization of the

[59] These statements are paraphrases of the actual reactions of lawyers interviewed by Kinwood Harris, MIT, Class of 1973.

problem at hand, reading the decisions of previous cases, interpretting the law, and predicting how the law will be applied to the current dispute. He then must advise his client and/or argue principles of law before the bench. The computer only acts as a "middleman" in relieving some of the drudgery associated with case law research. By no means is the system intended to replace the legal skills and expertise of the lawyer.

The response of "I cannot program computers" rings of the fear that such systems would necessitate that the lawyer go back to school to become a computer expert. This, of course, is not the case. Not knowing how to program a computer does not prevent one from operating a computer. The analogy can be made that one doesn't have to know how the internal combustion engine works in order to take advantage of using and driving an automobile. The computerized case law retrieval system already will be programmed and all the lawyer will have to do is to steer it onto the right path.

The notion that computers are expensive is a common one. In fact, the proposed centralized computing system is expensive. However, lawyers' time is also expensive, particularly when this expense is summed across all of the practicing lawyers in the U. S. The economic analysis in the previous chapter justified the operating costs of a computerized system based on the savings in lawyer time.

The response, "I don't need a computer, I get paid for performing legal research" is just the problem which the system is trying to solve. Lawyers should be compensated for the use of their professional skills and talents. But spending countless hours in a law library wading through outdated indices and searching through thick volumes of Reporters does not require a high caliber of skill and talent. By reducing the time necessary to perform legal research, the lawyer can devote more of his time to using his irreplaceable professional know-how in interpreting the law and applying the law to his client's problems.

The time and drudgery of case law research has resulted in the delegation of this task to junior members of law firms. This practice is cloaked by the premise that it is a good way to "break in" new lawyers, and is perpetuated by the line of reasoning which goes "I had to do it when I joined the firm; therefore, others have to do it when they join the firm." If case law research is actually a good way to "break in" lawyers, then the use of computers in case law research is probably even a better way to "break in" new lawyers since it would allow more research to be performed in the same amount of time.

The response "I don't trust computers" is probably the most serious behavioral problem and one which cannot be readily answered. In order for the attorney to have faith in the system, he would have to work with both the automated system and with manual operations for some time. "Ultimately, total credibility of any automated researching system probably must be predicated upon lengthy and continuing comparisons of computer results and results obtainable through traditional research methods."[60] This, of course, would be very expensive and time-consuming. R. W. Robins has suggested that a possible means of circumventing this problem is to have an organization such as the ABA investigate and certify automated case law retrieval systems.[61] Aside from the technical problems of how such an agency would test a system for certification, it is not at all clear that a favorable edict issued by a regulatory organization would cause lawyers to accept the computer and "believe" in its results.

Perhaps these behavioral problems will be solved by the impact of students currently graduating from law school. More and more law schools are offering introductory courses in computers as part of their curriculum. Basic understanding of computers will serve to eliminate the fear. Also, present law students are not tied to the security of the

[60] Robins, op. cit., p. 714

[61] Ibid.

traditional legal research system; they have a "let's see if it works" attitude toward the innovative use of computers in the legal profession.[62]

[62] Jeffrey A. Meldman, "Law Student Attitudes Toward Computers and Legal Research," Jurimetrics Journal, IX (June, 1969), p. 210.

Chapter 7 SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Traditional methods of performing case law research are no longer adequate to cope with the increasing volume of reported decisions. Since the decisions of litigated disputes constitutes a body of law in the U.S., the attorney must be able to accurately search through the collection of case decisions to determine what is the law and predict how it will be applied to the controversy at hand. This search for relevant cases through the use of cumbersome, outdated indices, digest, and key topics, is probably the most burdensome and least satisfying aspect of a lawyer's job. Besides being extremely time-consuming, manual indices do not guarantee that the lawyer will find those cases relevant to his problem. Rigid hierarchial indices tend to be most accurate in the dormant areas of law and least accurate in the dynamic, changing areas. In performing a search, the lawyer is limited by the quality and skill of the editors who index the cases. Pigeonholing cases into a limited set of legal issues causes many kinds of errors and distortions. All in all, the lack of an efficient, accurate means for obtaining relevant case decisions can lead to expensive delays, erroneous decisions based on inaccurate or incomplete information, and injustices resulting from the inaccessability of the law.

Early attempts at applying computers to case law retrieval were little more than mechanizations of existing manual systems. Although they retained all of the deficiencies inherent to manual indexing and abstraction, these "Point-of-Law" systems offered the advantage of increased speed and the capability to search on more than one topic simultaneously. The next major advance in this technology was the development of systems which were not dependent on indices. By storing full text, KWIC systems are able to circumvent all of the problems of manual classification. The major drawback of KWIC is the requirement that user search request words must exactly match words in the retrieved text. Of the various attempts at computerizing case law retrieval, KWIC-type systems have enjoyed the greatest success and a number of such systems are in operation today. Although they did not result in operational systems, experiments applying the association factor to case law research have demonstrated that the exact match requirements of KWIC can be overcome through the use of probabilistic techniques. In addition, the association factor technique is capable of ranking retrieved cases in order of probable relevancy.

The potential of the association factor algorithm for overcoming the index and language problems inherent in other manual and automated systems has led to its selection as the basis for the proposed computerized case law retrieval system. Suggested improvements to the basic technique include the use of automatic indexing and extensive lawyer interaction in the calculation of document relevance numbers. A preliminary hardware systems investigation indicates that a time-shared system consisting of a large central computer and many remote terminals located in the offices of law firms would be a technically feasible implementation. A cursory economic analysis of this implementation produced the result that the operating costs of the computerized case law retrieval system are more than offset just by the savings in lawyers' time. Other benefits such as "better" law and general law office data management would serve to improve the cost-benefit analysis.

It is recommended that a thorough study by made of the proposed computerized case law retrieval system. The technical, legal, and behavioral problems itemized in the body of this document must be investigated. Once these problems have been addressed, an effort should be made to secure government funds for a complete system implementation and test program. The development of such a system could have a profound affect, not only on the administration of justice, but also on the dissemination of information and knowledge in all other fields of endeavor.

BIBLIOGRAPHY

Allen, L.E. "Beyond Document Retrieval Toward Information Retrieval," *Minnesota Law Review*, XLVII (April, 1963) 713.

Anderson, Ronald R. "An Associativity Technique for Automatically Optimizing Retrieval Results," Lehigh University Center for the Information Sciences. Naval Research Contract Nonr-(710)08.

Baxendale, J.C.L. "Indexing and Third Selected Bibliography on Computers and the Law," *Rutgers Journal of Computers and the Law*, II (1971), 88.

Beard, Joseph J. "Information Systems Application in Law," *Annual Review of Information Science and Technology*, VI. Edited by Carlos Cuadra. Chicago: Encyclopedia Britannica, Inc., 1971.

Cohen, Morris. "Research Habits of Lawyers," *Jurimetrics Journal*, IX, No. 4 (June, 1969).

"Computer Retrieval of the Law: Challenge to the Concept of Unauthorized Practice?" *University of Pennsylvania Law Review*, CXVI (May, 1968), 1261.

"Computerized Legal Research in Countries Outside North America," *Jurimetrics Journal*, XII (March, 1972), 119.

Davis, Richard P. "Let There be LITE -- Legal Information Through Electronics," *Jurimetrics Journal*, VII, No. 2 (December, 1966), 118.

________. "The LITE System," *JAG Law Review*, VII, No. 6 (November-December, 1966).

Dennis, Sally F. "The Design and Testing of a Fully Automatic Indexing-Searching System for Documents Consisting of Expository Text," *Information Retrieval*. Edited by George Schecter. Thompson Book Co., 1967.

Dickerson, F.R. "Electronic Searching of Law," *American Bar Association Journal*, XVII (May, 1971), 167.

Eldridge, William B., Dennis, Sally F. "The Computer as a Tool for Legal Research," *Law and Contemporary Problems*, XXVIII, No. 1 (1963), 76.

Fay, Robert J. "Full-text Information Retrieval," Law Library Journal, LXIV (May, 1971), 167.

Fels, E.M., Jacobs, J. "Linguistic Statistics of Legal Indexing," University of Pittsburgh Law Review, XXIV (June, 1963), 771.

Fenwick, William A. "Automation and the Law: Challenge to the Attorney," Vanderbilt Law Review, XXI, No. 2 (March, 1968), 228.

Goldblum, Edward J. "Application of Computers to Retrieval of Case Law." Unpublished Master's thesis, Sloan School of Management, Massachusetts Institute of Technology, 1968.

Harrington, William G. "Computers and Legal Research," American Bar Association Journal, LVI (December, 1970), 1145.

Harrington, William G., Wilson, H. Donald, Bennet, Robert L. "The Mead Data Central System of Computerized Legal Research," Law Library Journal, LXIV (May, 1971), 184.

Hoffman, Paul S. "Lawtomation in Legal Research: Some Indexing Problems," Modern Uses of Logic in Law, (March, 1963), 16.

Hoppenfield, Ellias C. "Law Research Service, Inc." Modern Uses of Logic in Law, (March, 1966), 46.

Hudson, Clayton A. "Some Reflections on Legal Information Retrieval," Osgoode Hall Law Journal, VI (December, 1968), 259.

Jones, Paul E., Curtice, Robert M., Giuliano, Vincent E., Sherry, Murry E. "Application of Statistical Association Techniques for NASA Document Collection," NASA Contractor Report CR-1020. Cambridge, Mass.: Arthur D. Little, Inc., 1968.

Kayton, Irving. "Retrieving Case Law by Computers: Fact, Fiction, and Future," George Washington Law Review, XXXV (October, 1966), 1.

Kehl, William B., Horty, John F., Bacon, Charles R.T., Mitchell, David. "An Information Retrieval Language for Legal Studies," *Communications of the Association for Computing Machinery*, IV (September, 1961), 9.

Lancaster, F. Wilfred, Gillespie, Constantine J. "Design and Evaluation of Information Systems," *Annual Review of Information Science and Technology*, Vol. V, Edited by Carlos Cuadra. Chicago: Encyclopedia Britannica, Inc., 1970.

Lancaster, F. Wilfred. *Information Retrieval Systems*. New York: John Wiley & Sons, Inc., 1968.

"Legal Information Retrieval Systems and the Revised Copyright Law," *Valparaiso University Law Review*, I, No.2, (Spring, 1967), 359.

Lyons, John C. "New Frontiers of the Legal Technique," *Modern Uses of Logic in Law*, (December, 1962), 256.

Maron, M.E., Kuhns, J.L. "On Relevance, Probabilistic Indexing and Information Retrieval," *Journal of the Association for Computing Machinery*, VII (1960), 216.

Meldman, Jeffery A. "Law Student Attitudes Toward Computers and Legal Research," *Jurimetrics Journal*, IX, No. 4 (June, 1969), 207.

Melton, Jessica F., Bensing, Robert C. "Searching Legal Literature Electronically: Results of a Test Program" *Minnesota Law Review*, XLV (1960), 229.

Melton, Jessica F. "English Language Problems: Mechanized Search Techniques," *Law Library Journal*, LVI (November, 1963), 432.

Mermin, S. "Computers, Law, and Justice: An Introductory Lecture," *Wisconsin Law Review*, (Winter, 1967), 43.

"Proceedings of the Special Committee on Electronic Data Retrieval," *Modern Uses of Logic in Law*, (March, 1962), 44.

Robertson, E. "The Parametric Description of Retrieval Tests," *Journal of Documentation*, XXV, No. 1 (March, 1969), 1.

Robins, W. Ronald. "Automated Legal Information Retrieval," *Houston Law Review*, V, No. 4 (March, 1968), 691.

Salton, Gerard. *Automatic Information Organization and Retrieval.* New York: McGraw-Hill, Inc., 1968.

Skelly, S.J. "Computers and Legal Information Retrieval," *Ottawa Law Review*, III (Spring, 1969), 433.

________. "Computers and the Law," *Saskatchewan Law Review*, XXXIII, No. 3 (Fall, 1968), 167.

Smith, Fredrick E. "Computer Applications to Legal Documentation: What is Not Being Done," *Law Library Journal*, LXIV (May, 1971), 113.

Stiles, H. Edmund. "The Association Factor in Information Retrieval," *Journal of the Association for Computing Machinery*, VIII (1961), 271.

________. "Automatic Indexing and the Association Factor," *Information Systems Compatibility*. Edited by Simon M. Newman. Washington, D.C.: Spartan Books, Inc., 1965.

Tapper, Conlin F.H. "Research and Legal Information by Computer," *Chicago Bar Record*, XLVIII, No. 8 (June-July, 1967), 226.

Troy, Frank J "Ohio-Bar Automated Research -- A Practical System of Computerized Legal Research," *Jurimetrics Journal*, X, No. 2 (December, 1969), 62.

Wilson, Robert A. "Computer Retrieval of Case Law," *Southwestern Law Journal*, XVI (September, 1962), 409.

APPENDIX I

MEASURE USED FOR AUTOMATIC INDEXING[63]

$$\frac{R}{K} \geq 10{,}500 \Longrightarrow \text{Informing Word}$$

$$\frac{R}{K} < 10{,}500 \Longrightarrow \text{Non-informing Word}$$

R= Number of raw occurrences

K is computed as follows:

N = Number of documents in sample

L_d = Number of words of running text in document d

$f_{c,d}$ = Number of occurrences of word c in document d

$$g_{c,d} = \frac{f_{c,d}}{L_d}$$

$$\bar{g}_c = \frac{\sum_{d=1}^{N} g_{c,d}}{N}$$

$$s_c^2 = \frac{\sum_{d=1}^{N} (g_{c,d} - \bar{g}_c)^2}{N - 1}$$

$$K_c = \frac{\bar{g}_c^{\,2}}{s_c^{\,2}}$$

[63] Dennis, op. cit. p. 93.

APPENDIX II

BREAK-EVEN ANALYSIS

LAW FIRM OF 4 LAWYERS
(N= 4)

C	R	H
15	0.1	1.68
15	0.15	1.92
15	0.2	2.24
15	0.25	2.69
15	0.3	3.37
15	0.35	4.49
15	0.4	6.73
15	0.45	13.46
20	0.1	1.76
20	0.15	2.07
20	0.2	2.52
20	0.25	3.23
20	0.3	4.49
20	0.35	7.34
20	0.4	20.19
20	0.45	UNFEASIBLE
25	0.1	1.84
25	0.15	2.24
25	0.2	2.88
25	0.25	4.04
25	0.3	6.73
25	0.35	20.19
25	0.4	UNFEASIBLE
25	0.45	UNFEASIBLE
30	0.1	1.92
30	0.15	2.45
30	0.2	3.37
30	0.25	5.38
30	0.3	13.46
30	0.35	UNFEASIBLE
30	0.4	UNFEASIBLE
30	0.45	UNFEASIBLE
35	0.1	2.02
35	0.15	2.69
35	0.2	4.04
35	0.25	8.08
35	0.3	UNFEASIBLE
35	0.35	UNFEASIBLE
35	0.4	UNFEASIBLE
35	0.45	UNFEASIBLE

```
LAW FIRM OF 6 LAWYERS
    ( N= 6 )

    C                  R                  H

    15                 Ø.1                 1.12
    15                 Ø.15                1.28
    15                 Ø.2                 1.5
    15                 Ø.25                1.79
    15                 Ø.3                 2.24
    15                 Ø.35                2.99
    15                 Ø.4                 4.49
    15                 Ø.45                8.97

    2Ø                 Ø.1                 1.17
    2Ø                 Ø.15                1.38
    2Ø                 Ø.2                 1.68
    2Ø                 Ø.25                2.15
    2Ø                 Ø.3                 2.99
    2Ø                 Ø.35                4.9
    2Ø                 Ø.4                 13.46
    2Ø                 Ø.45               UNFEASIBLE

    25                 Ø.1                 1.22
    25                 Ø.15                1.5
    25                 Ø.2                 1.92
    25                 Ø.25                2.69
    25                 Ø.3                 4.49
    25                 Ø.35                13.46
    25                 Ø.4                UNFEASIBLE
    25                 Ø.45               UNFEASIBLE

    3Ø                 Ø.1                 1.28
    3Ø                 Ø.15                1.63
    3Ø                 Ø.2                 2.24
    3Ø                 Ø.25                3.59
    3Ø                 Ø.3                 8.97
    3Ø                 Ø.35               UNFEASIBLE
    3Ø                 Ø.4                UNFEASIBLE
    3Ø                 Ø.45               UNFEASIBLE

    35                 Ø.1                 1.35
    35                 Ø.15                1.79
    35                 Ø.2                 2.69
    35     @           Ø.25                5.38
    35                 Ø.3                UNFEASIBLE
    35                 Ø.35               UNFEASIBLE
    35                 Ø.4                UNFEASIBLE
    35                 Ø.45               UNFEASIBLE
```

```
LAW FIRM OF 8 LAWYERS
    ( N= 8 )

    C                  R                  H

    15                 Ø.1                Ø.84
    15                 Ø.15               Ø.96
    15                 Ø.2                1.12
    15                 Ø.25               1.35
    15                 Ø.3                1.68
    15                 Ø.35               2.24
    15                 Ø.4                3.37
    15                 Ø.45               6.73

    2Ø                 Ø.1                Ø.88
    2Ø                 Ø.15               1.Ø4
    2Ø                 Ø.2                1.26
    2Ø                 Ø.25               1.62
    2Ø                 Ø.3                2.24
    2Ø                 Ø.35               3.67
    2Ø                 Ø.4                1Ø.1
    2Ø                 Ø.45             UNFEASIBLE

    25                 Ø.1                Ø.92
    25                 Ø.15               1.12
    25                 Ø.2                1.44
    25                 Ø.25               2.Ø2
    25                 Ø.3                3.37
    25                 Ø.35               1Ø.1
    25                 Ø.4              UNFEASIBLE
    25                 Ø.45             UNFEASIBLE

    3Ø                 Ø.1                Ø.96
    3Ø                 Ø.15               1.22
    3Ø                 Ø.2                1.68
    3Ø                 Ø.25               2.69
    3Ø                 Ø.3                6.73
    3Ø                 Ø.35             UNFEASIBLE
    3Ø                 Ø.4              UNFEASIBLE
    3Ø                 Ø.45             UNFEASIBLE

    35                 Ø.1                1.Ø1
    35                 Ø.15               1.35
    35                 Ø.2                2.Ø2
    35                 Ø.25               4.Ø4
    35                 Ø.3              UNFEASIBLE
    35                 Ø.35             UNFEASIBLE
    35                 Ø.4              UNFEASIBLE
    35                 Ø.45             UNFEASIBLE
```