# Beyond Boolean Search:
# FLEXICON, A Legal Text-Based Intelligent System

DAPHNE GELBART J. C. SMITH

University of British Columbia
Faculty of Law Artificial Intelligence Research Project
Vancouver, Canada V6T 1Y1

ABSTRACT

FLEXICON is a new litigation support system being developed for the effective retrieval of legal documents by legal and paralegal professionals. It includes a text analysis and processing component, processing raw text and intelligently extracting key information in the form of electronic "headnotes". It also includes an innovative non-boolean search and retrieval mechanism. As well, it provides many features that improve legal research such as a menu-driven interface, thesauri, relevance feedback and retrieval by topic.

## 1. INTRODUCTION.

The essence of legal research in common law jurisdictions is the retrieval of relevant decided cases and related legal information. An effective information retrieval system is thus an essential litigation support tool. Legal professionals access more information than any other group of professionals. On a daily basis, lawyers access electronic databases that contain tens of millions of documents. The large volume of legal information and the enormous effort required to manually abstract and index cases for every domain of law call for a system which automates the process of generating a database of document profiles and case summaries and effectively searches the database to retrieves cases and other legal documents relevant to a user's request. The improved access to legal cases in electronic form, either transferred directly from the courts or scanned from hardcopy documents, as well as the improvement in personal computers and storage technology make automation feasible where it was earlier not possible.

While a number of computerized legal information retrieval systems are available to respond to this need, including QuickLaw and CAN/LAW in Canada and WESTLAW and LEXIS in the US, it appears that no existing system fully addresses the special needs of judges, lawyers and legal researchers. Existing systems use generic search mechanisms that fail to address many specific needs of legal professionals. Most existing systems allow the user to conduct keyword searches of mostly unstructured case databases via boolean queries. Boolean searches can be confusing to legal or paralegal users and often result in matches that are under-inclusive or over-inclusive, retrieving large numbers of cases, many not relevant, or missing relevant cases which do not exactly match the user's search request. Most systems also use awkward command languages that have to be memorized and provide little or no assistance to the user in formulating a query.

The cases retrieved by existing systems are generally recovered in full text, which the user must examine to ascertain whether or not the case is relevant. To avoid laboriously paging through large quantities of retrieved information, some form of document summary is needed to allow users to determine the relevance of documents retrieved by the system. Publishers of legal cases often provide case headnotes, abstracts, indexes and digests. These tools require extensive human time and expertise to prepare. When legal information is stored and retrieved electronically, this process can be automated. The recent trend of courts to produce electronic decisions in standard format, which may be instantly transferred to a database upon release, suggests the usefulness of automatically generated case summaries that would be available for immediate use and would be fully integrated with a search and retrieval system.

The University of British Columbia FLAIR (Faculty of Law Artificial Intelligence Research) project was established to develop techniques for the processing and retrieval of legal information. FLEXICON* (Fast Legal EXpert Information CONsultant) is FLAIR's new intelligent text-based system [Gelbart and Smith, 1990]. FLEXICON offers the three components of a third generation legal information retrieval systems as defined by Bing [Bing, 1987]: a search function in the form of a non-boolean effective retrieval mechanism; a relevance function in the form of computer-generated profiles, case summaries and HYPERTEXT links to rapidly determine relevance of retrieved cases; a source function, in the form of electronic and hardcopy summaries and full text of retrieved cases. In addition, FLEXICON provides a menu-driven user interface, assistance in formulating meaningful search requests via thesauri and relevance feedback as well as many other features of interest to legal professionals.

A typical case contains a factual story, a description of the set of legal issues which the story gives rise to, a statement of the applicable law, and a resolution to the issues when the law has been applied to the facts. The statements of law within the case contain legal concepts and references to cases and statutes. We can build up a profile of a case in terms of the relationships between four parameters: concepts, cases, legislation and facts. By measuring the frequency and proximity of these parameters in relationship to each other, we can produce a database of profiles of cases which can serve as



* FLEXICON: (c) Copyright 1989, Daphne Gelbart and J.C. Smith.

summaries of their counterparts in the document database. We can then retrieve cases from the database by comparing them against a query composed of the four types of keyword terms and selecting those cases whose profiles are most similar to the query.

While most commercial legal information retrieval systems are based on boolean search and manually indexed case databases, several interesting approaches to legal information retrieval were suggested including the conceptor-based retrieval and norm structures suggested by Bing [Bing, 1987, a; Bing '87, b], the semantic network model based on conceptual structures suggested by Hafner [Hafner, 1981], the connectionist approach suggested by Belew and Rose [Belew 1987; Rose and Belew 1989], the retrieval of arguments from legal cases [Dick, 1987], as well as rule-based and case-based expert systems linked to databases of legal cases developed by our group [Smith and Deedman, 1987; MacCrimmon, 1989, Kowalski, 1991]. FLEXICON attempts to provide cost-effective quality retrieval of cases and documents from any domain of law beyond that provided by conventional boolean search and without spending an inordinate amount of time, cost and effort on the construction and maintenance of semantic networks, conceptor links or domain rules.

## 2. TEXT ANALYSIS AND PROCESSING.

Automatic text analysis and processing serve two important purposes in FLEXICON: (1) to generate document summaries or "headnotes" that will serve as a means for the user to rapidly familiarize him- or herself with the document's content, establishing its relevance to his or her needs; (2) to structure the text database and generate document representatives containing index information necessary and sufficient to match the document with a user's query. The text analysis mechanism employed by FLEXICON goes beyond simple keyword extraction and uses legal knowledge to intelligently recognize terms that are significant representatives of legal documents. FLEXICON recognizes complex legal phrases based on approximate matching of words and word ordering. While true natural language understanding and processing through the use of robust parsers is not feasible at present, we are currently able to use the computer in the intelligent processing of text, its classification and the extraction of significant information, not through true text understanding but by the combined application of legal knowledge and computational linguistic methods.

### 2.1 Document Profiles.

Document analysis involves scanning the raw text of the document and automatically extracting key information that can serve as a document representative or profile. We can build up a profile of a case in terms of the relationships between four parameters meaningful to legal professionals: concepts, cases, legislation and facts. The fact terms represent the factual story on which the case is based. Legal concepts represent a statement of the applicable law and the resolution to the issues when the law has been applied to the facts. Case citations stand for related issues previously decided on point or by analogy. Statute citations are references to applicable legislation. FLEXICON employs an improved mechanism of case and statute citation extraction, recognizing case citations referred to by a shortened version of the style of cause and statute section numbers referred to without mention of the statute's title and, if necessary, converting citations to a standard format.

While Tapper [Tapper 1979; Tapper 1984] suggests that citations are superior to keywords as document representatives since they have neither synonyms nor homographs and they serve as short coded expressions standing for complete issues, we maintain that profiles based on the four types of outlined keywords provide the most complete representation of legal cases. The four types of keywords provide added flexibility in document retrieval by allowing the user to base a search on specific dimensions of law. For example, the user may retrieve relevant cases based on factual information or else search for cases that share common legal issues. The document profiles include all the information necessary and sufficient to match documents with a user's query, providing a compact and structured representation of the text, excluding only noise words, thus making it unnecessary to revert to the text at search time.

The four types of profile keywords are weighted by factors reflecting their relative significance. The weight factors of concept and fact terms are proportional to the document term frequency and inversely proportional to the number of documents in which the term resides, favouring terms that occur frequently in the particular document, but giving less weight to terms that tend to occur in many cases in the collections. The weight factors of citations require additional analysis. Citation weight factors are proportional to the document term frequency. Unlike concept and fact terms, frequently cited cases or statutes tend to be the more important citations and therefore the assignment of higher weight for frequently cited cases and statutes should be tested. Additional weight factors and their relationships will be tested for case citations including the age of the case, the court level and the remoteness of the jurisdiction [Tapper 1979; Tapper 1984]. As well, the relevance of a cited case can be deduced by comparing its profile to that of the citing case using the matching procedure outlined in section 3.4 below and assigning higher weight factors to cases that most resemble the citing case.

### 2.2 Electronic "Headnotes".

The document profiles produced by the automatic text analysis component of FLEXICON serve as a basis for the automatic construction of electronic "headnotes" which we call *flexnotes*. Like the manually constructed headnotes of printed law reports, flexnotes provide easily scanned summaries of judgments. The flexnotes generated by FLEXICON are designed to provide legal professionals with the essence of the case and a means to rapidly decide relevance of retrieved cases.

The flexnote, which is automatically generated from the text of the case, differs from the headnote presented in legal publications. Figure (1) shows a portion of a flexnote. First, header information is displayed, such as the style of cause of the case, date, jurisdiction and the judges that heard the case. This is followed by an automatically generated classification of the case to a subject of law (such as criminal, constitutional, private and public law), based on analysis of the number and type of statutes, the number and type of cases and the style of cause of the case. Next, a summary of the concepts, facts, case citations

GADUTSIS ET AL. v. MILNE ET AL.
ONTARIO HIGH COURT OF JUSTICE
BRITISH COLUMBIA REGISTRY
BEFORE: PARKER
DECEMBER 20, 1972

Private Law

| CONCEPTS | CASES |
|---|---|
| [ 110] zoning | [ 3] Hedley Byrne & Co. Ltd. v. Heller & Partners Ltd |
| [ 47] lease | [ 2] Windsor Motors Ltd. v. District of Powell River |
| [ 39] regulations | [ 2] Rutter v. Palmer [1922] 2 K.B.) the words employ |
| [ 36] rent | [ 1] Mutual Life & Citizens' Ass'ce Co. Ltd. v. Evatt |
| [ 33] servants | [ 1] Mersey Docks Trustees v. Gibbs (1866), L.R. 1 H. |
| [ 27] municipal corporation | [ 1] Marschler v. G. Massers Garage [1956] O.R. 328 |
| [ 27] waived | [ 1] Glengoil SS. Co. v. Pilkington (1897), 28 S.C.R. |
| [ 27] special damages | [ 1] Dixon v. City of Edmonton [1924] S.C.R. 640 |
| [ 24] sentence | [ 1] Canada Steamship Lines Ltd. v. The King [195[illegible] 2 |
| [ 21] stipulation | |
| **STATUTES** | **FACTS** |
| [ 2] Ontario Planning Act | [ 15] permit |
| -[ 1] | [ 15] Kerenvi |
| -[ 1] s. 61.2 | [ 11] restaurant |
| [ 1] Civil Code | [ 11] building permit |
| -[ 1] s. 1019 | [ 10] Shimski |
| | [ 10] Petropoulos |
| | [ 10] Toronto |
| | [ 8] Milne's |
| | [ 8] [illegible]ployees |
| | [ 8] Markham St |

KEY PARAGRAPHS

\001\ ROBINS, J. (orally):-- On September 10, 1973, the plaintiff John Bosworth Limited ("Bosworth") entered into a written agreement of purchase and sale with one Louis Train by which Train agreed to purchase and Bosworth to sell some 86 acres of land in the Township of Whitchurch for the price of $430,000. The transaction was closed on November 19, 1973, when a deed was registered in the name of the defendant Professional Syndicated Developments Limited ("Syndicated"), a company incorporated by Train and others for the purposes of the transaction. On closing, Syndicated paid the moneys then due and gave back a mortgage of $330,000 to secure the balance of the purchase p:ice. Because that mortgage went into default, Bosworth brought this action for foreclosure and other usual relief.

\002\ Syndicated acknowledges that the mortgage is in default but defends the action and counterclaims for rescission or damages on the basis of allegations to the effect that Bosworth's real estate agent induced it to purchase the lands by representing that they were zoned "Industrial M2"; that Syndicated closed the transaction and gave back the $330,000 mortgage relying on that representation; that the representation was untrue and constituted a material misdescription; that, by reason of the misrepresentation, Syndicated "received property which was different in nature, quality, and substance from that which it was represented to be"; and that the representation was made by or on behalf of the plaintiff fraudulently, well knowing the same to be false, or recklessly and not caring whether it was true or false. Alternatively, Syndicated alleges that the agreement was entered into on the belief of both parties that the property was zoned M2 Industrial, that it was not so zoned and there had, therefore, been a total failure of consideration.

\006\ The minutes of the Council of the municipality reflect that the lands were considered industrial and treated as such. Council was obviously anxious to have these lands developed industrially and at one stage went so far as to indicate that it would redesignate them rural if Bosworth did not proceed with the industrial park it proposed for the site. In the summer of 1973, Bosworth put the property on the market for sale as industrial lands through W. E. Fockler Real Estate and another broker brought in by Fockler, Michael Jay Real Estate Limited.

figure 1. A portion of a sample FlexNote

and statute citations is presented in the FLEXICON quadrant structure, ordered in decreasing order of the term's weight factors. Finally, key paragraphs that appear to express the essence of the judgment are displayed. In the future, we also plan to include Authorities Considered citations in the flexnotes, referencing legal text or academic journals cited by cases in the database.

The "important paragraph extraction" module of FLEXICON uses legal analysis in the intelligent evaluation and selection for display of significant fact, issue and law paragraphs that best represent the case. The paragraph extraction is, both, inductive: to provide the means to quickly determine relevance of retrieved cases, as well as informative: to supply information about the facts and law discussed in a case. Each paragraph is analyzed by considering factors including key phrases, significant concept and fact terms, citation patterns, paragraph position, continuity and length. Dictionaries of phrases that tend to occur in "issue", "fact" and "law" paragraphs, weighted according to the phrase's significance, have been compiled by the FLEXICON team's legal analysts. The program eliminates unimportant "quotation" paragraphs or very short paragraphs. It then computes scores for important paragraph classes based on weighted relevant factors and selects high scoring paragraphs for display. While the extracted key paragraphs can occasionally miss important issues which were not emphasized in the original case but came to be significant in retrospect, they are produced instantly, consistently and at minimal cost.

The various methodologies used to create computer-generated case extracts are well summarized by Paice [Paice, 1989]. Unlike most computer-generated abstract research reported, which concentrates on extracting sentences and is, therefore, faced with serious problems of textual continuity and anaphoric references, FLEXICON extracts whole paragraphs, giving priority to groups of contiguous paragraphs. In order to ensure adequate balance and coverage of the essence of a case, it attempts to distinguish between paragraphs discussing the facts of the case, the law applied and issues discussed and to represent high-scoring paragraphs from each class in the abstract. Since existing legal cases are usually unstructured, FLEXICON can not rely on textual superstructures in the form of headings and sections and must distinguish between various components of a case according to rules supported by corresponding lexicons. The idea of abstract frames or schemas, as suggested by Paice, was considered for specific subdomains of law, such as damages for soft-tissue injury (Whiplash), which tend to follow a script. In order to apply to a wide domain such as law, case schemas will require automatic classification of cases to narrow subdomains, for which corresponding scripts will be created, yielding improved case abstracts. FLEXICON currently provides limited automatic case classification. We are currently investigating the feasibility of a finer classification based on automatic clustering of case profiles.

We plan to test and tune our paragraph extraction program on a large database of representative cases. Our existing test results reveal that well prepared case extracts can be used successfully as case summaries, especially when coupled with document profiles indicating the significant legal concepts, facts, case citations and cited legislation. Other research results confirm this finding [Paice, 1989]. In fact, recent research in a different domain reveals that no marked differences in comprehension are reported when information was presented with full text, by abstract or extract [Hunter, 1988].

### 2.3 Full Text Browsing

The flexnote is followed by the text of the case with numbered paragraphs, allowing unique references to portions of the text which are independent of the pagination method employed when printing the case. While the flexnote is intended primarily for on-line use, presented to the user as a brief summary screen with function keys allowing selective viewing of various headnote components and a HYPERTEXT feature from keywords and citations to their occurrences in the text, a hardcopy of the flexnote is also generated and inserted in front of the case text. The user can print selected cases in the FLEXICON format with the FLEXICON headnote, thus creating a private library of legal documents.

## 3. SEARCH AND RETRIEVAL.

Unlike existing legal information retrieval systems that use generic search and retrieval mechanisms, the FLEXICON search is specific to the legal domain. Blair and Maron [Blair and Maron, 1985] have demonstrated the low effectiveness of boolean search in large databases, of which the users are often unaware. In an experimental retrieval of legal documents from a large database, the recall of retrieval by lawyers, in their domain of expertise, was only 20 percent, while the lawyers believed that they were retrieving over 75 percent of all relevant documents. Blair and Maron explain the low recall of large information retrieval systems in terms of the necessity to formulate queries that will reduce the output overload characteristic of searching large databases. Typically, a user enters one or more query terms, connected by AND, OR or NOT operators. The user then examines the number of documents retrieved and adds terms connected by AND operators to reduce the output overload, until a sufficient and manageable volume of documents is retrieved. Boolean queries in large databases tend, therefore, to reflect the extent of the output requested rather than the information necessary to describe the domain of retrieved documents in detail, since the inclusion of too many ORed terms causes output overload while including too many ANDed terms quickly reduces the probability of retrieval to zero.

The FLEXICON document retrieval methodology is based on ranking relevant documents according to the similarity of their profiles to the user's query and not on the basis of the presence or absence of a single term. The FLEXICON user is encouraged and assisted in formulating informative search requests specifying in detail the characteristics of relevant documents and improving the quality of the search without causing output overload or reducing the retrieval recall when searching large databases. Many of the problems characteristic of boolean search are avoided with the FLEXICON search methodology. For example, the inclusion of homographs in a boolean query could lead to the retrieval of irrelevant documents [Choueka, 1985] due to terms ambiguity. The FLEXICON search, however, will place non-relevant documents at the bottom of the ranked list since their overall similarity to the query is expected to be very low.

In analogy to document profiles, the FLEXICON search requests are composed of four keyword types meaningful to legal users: legal phrases, facts, case and statute citations. The statement of the applicable law is represented by legal concepts. Relevant issues, on point or by analogy, are represented by case citations. Applicable legislation is represented by statutes or specific sections and paragraphs of a code or act. Factual information can be specified by fact terms. As mentioned, citation query terms have the advantage of neither missing relevant documents that include synonyms of query terms nor retrieving irrelevant documents that include homographs of query terms (e.g.: "will"). However, since FLEXICON can match synonymous concept and fact terms by employing a synonym thesaurus, we maintain that queries containing the four types of outlined keywords best represent the cases to be retrieved and produce the best search results.

While the technique employed by FLEXICON to automatically analyze and process documents can be easily applied to process queries written in natural language and convert them to the four lists of weighted terms in analogy to document profiles, we believe that search profiles derived from natural language queries will be inferior to those entered directly by the user, given an easy to use interface, assistance in the selection of useful terms and a means to indicate the significance of specific terms. Queries written in natural language are expected to be brief and are unlikely to include repeated terms to be used for frequency analysis for the purpose of term weighting. Spelling or typing errors and the use of non-standard citations or terms not in the system dictionaries can be difficult to detect. As well, the user may spend unnecessary time and effort in producing queries accounting for all the information necessary for an effective search while still using a meaningful natural language statement. For these reasons we have designed a simple, elegant and fully menu driven user interface for FLEXICON search.

An option that we might make available in forming queries with FLEXICON is "retrieval by example", i.e. retrieving cases similar to some sample database cases entered by the user. FLEXICON will automatically formulate a query consisting of the supersets of terms occurring in the document profiles of those cases. The user can then add and delete terms to customize the query to his or her needs.

3.1 The Search User Interface.

In the first search specification screen the user has the option of defining the scope of the search by specifying parameters such as the subject of law, date range, court, jurisdiction or judges.

In the following and primary FLEXICON search specification screen, the user is instructed to enter into four quadrants legal concepts, facts, case and statute citations which he or she thinks are relevant to the search. The FLEXICON user may base a search on a specific dimension of law by entering, for example, only factual terms, in which case FLEXICON will compare the query to the fact quadrants of document profiles only. The most relevant documents, however, tend to match the user's query in several dimensions of law, sharing similar legal concepts, a similar fact pattern, relying on common legislation and/or displaying relevance by analogy via the citation of common cases.

The user can type the entered terms; however, in order to facilitate data entry, avoid spelling and typing errors and establish a standard citation referral, FLEXICON provides dictionaries of legal concepts, fact terms and case and statute citations. The dictionary terms displayed during a given consultation reflect any prior selections of definition parameters. For example, upon the selection of a given subject of law, the concept dictionary will display only those legal phrases relevant to that domain. The dictionary provides two data entry modes: the user can select letters from the alphabet and then select from lists of terms starting with that letter. Alternatively, the user can enter the beginning of a word or name and the system will display all the terms or citations that start with that prefix. In order to improve the search quality, the user has the option of indicating the significance of each term entered by qualifying each term as high, medium or low (the default is medium). The user can also specify the maximum number of cases he or she wishes to retrieve. By default, FLEXICON returns those cases whose matching score with the user's query exceeds a predefined threshold.

Unlike most existing systems which use awkward command languages that require memorization of a specific syntax and familiarity with the structure of the document segmentation, the FLEXICON search specification mechanism is easy to use and designed to provide maximal assistance to the user and minimize the error probability in data entry.

3.2 Query Refinement: the Related Terms Thesaurus.

FLEXICON provides assistance in formulating an effective search profile via a network of associations between related terms. After entering terms on the search specification screen, the user can view, related terms for each concept, case and/or statute citations which, at the user's discretion, can be added to the original query. This is in contrast to most existing systems that provide little or no assistance to the user in forming a search request and lack any domain expertise that could improve the search outcome.

The related terms thesaurus is automatically constructed from the text database. The basic thesaurus algorithm links terms that tend to statistically co-occur in many documents, assigning a measure of association between terms that is inversely proportional to their lexical distance in the text. While Attar and Fraenkel rightly argue that the thesaurus functionality works best in a local context due to the extent of processing and the size of data structures which may prohibit its construction in conjunction with large databases [Attar and Fraenkel, 1981], we were able to substantially reduce the processing effort and space requirements by limiting the thesaurus to associations among legal concepts, case citations and statute citations. This limitation of the thesaurus to terms representing legal issues has the added advantage of producing associations that are generally of global scope. While the thesaurus terms are not always applicable in a given context, the FLEXICON thesaurus does not perform automatic terms substitution. Rather, the user adds relevant thesaurus terms to the query at his or her own discretion.

In addition to recording statistical co-occurrence of terms in the text, the FLEXICON thesaurus links will reflect the citation mechanism of cases and statutes, forming associations

between appealed and original cases and between similar statutes in different jurisdictions. We are also considering providing the FLEXICON user with information regarding changing statute sections and subsections as well as periods in which statutes are in effect.

3.3 Relevant Document Selection.

The simple boolean search used by existing systems is restrictive in that the search outcome may depend on the presence or absence of a single term. Boolean queries ANDing many terms tend to retrieve very few or no relevant documents, missing many relevant ones. Boolean queries ORing many terms tend to retrieve large volumes of documents, often due to the out-of-context occurrence of search terms in text. In order to get manageable volumes of retrieved information, users tend to formulate short and uninformative queries. The search approach in FLEXICON uses the model suggested by Salton and represents both document profiles and queries as ordered lists of weighted keywords [Salton and McGill, 1983; Salton, 1989; Salton and Buckley, 1988]. Relevant documents are retrieved by comparing a query composed of the four types of keyword terms to document profiles stored in the FLEXICON database. The FLEXICON query definition screen is demonstrated in figure (2). A portion of a case profile is displayed in Figure (1) (The hardcopy lists all the citations but only the most important legal concepts and fact terms). Both query and document terms are weighted by several weight factors, to improve the accuracy of the match. As indicated in the discussion of document profiles, document terms are weighted according to the term's frequency of occurrence in the document, its overall frequency in the data collection and other information qualifying the case. Query terms are also weighted by the overall frequency in the data collection, as well as by a user-entered significance factor.

The user's query is compared to document profiles generated by the text analysis component of FLEXICON, producing a matching score which assesses the degree of similarity between query and profile. Variations of the Cosine formula suggested by Salton [Salton 1989] are used to compute the extent of match between the four document and query corresponding quadrants and to produce a matching score based on a weighted average of the individual scores. The documents whose profiles score highest are returned to the user in decreasing order of matching score. In order to improve the efficiency of the search, an inverted index of terms is constructed and used to produce the initial list of cases which could be relevant. Only those cases are matched with the query to produce the final list of ranked, best-matching cases.

Specialized matching functions have been developed to best match case and statute citations. In order to enhance matching on the basis of case citations, FLEXICON will use citation cross reference information to compile, for each database case, the list of cases citing the case, as well as corresponding appeal cases (which may not explicitly cite the original case but can be identified by the parties involved in the legal suit). The search function will compare the cases entered in the query to cases citing or appealing a given case, as well as to the cases cited by the case. This feature provides the capability to retrieve appeal cases as well as earlier cases than those entered in the query on the basis of case citations. For example, a relatively recent case citation entered in a user's query can not match cited cases appearing in the profiles of earlier cases but can match the list of citing cases compiled for that case. Specialized statute matching functions were also developed. For example, FLEXICON will fully match a statute citation occurring in the query with no section number with any occurrences of specific section numbers of that statute in document text. It will provide, however, only a partial match between a specific section number in the query and citations of the statute, with no section numbers, in the text. As well, FLEXICON can match case citations that occur in various formats with those stored in the FLEXICON format.

3.4 Beyond Textual Term Matching: The Synonym Thesaurus.

The synonym thesaurus allows FLEXICON to match queries and documents that share similar terms semantically but not textually. The similarity matching routine recognizes complete or partial matches between legal concept and fact terms belonging to the same thesaurus class, thus retrieving cases that might be missed by search mechanisms based on simple keyword match. Unlike the related terms thesaurus whose function is to improve the search profile, this thesaurus function is performed automatically by the search program, without user intervention.

3.5 Viewing the Search Results.

The result of the FLEXICON search is a list of documents ranked in descending order of similarity to the query and a histogram plotting the matching scores of retrieved cases, providing a visual means to determine the subset of best-matching documents. Figure (3) shows the list of cases retrieved by a FLEXICON search and their matching scores. The user can view the flexnotes generated by the text analysis component of FLEXICON in order to determine the relevance of retrieved cases. A flexnote summary screen is first displayed. The user can then view other sections of the flexnote, browse the full text of the document by following HYPERTEXT links, copy, edit and print selected information. This is in contrast to existing systems which do not always provide headnotes, in which case the user must laboriously page through large quantities of text to determine relevance.

3.6 Interactive Search: Relevance Feedback.

Relevance feedback describes a process whereby terms found in profiles of documents retrieved by an initial query can be used to refine that query and allow the search to be repeated until the user is satisfied. While inspecting the profiles of retrieved relevant documents, the user may be reminded of terms that could improve the original query. FLEXICON allows the user to easily and selectively add terms of his or her choice to the original search profile, optionally reassign the term significance and repeat the search. This process can take place by scanning the flexnotes of high-ranking retrieved cases. Alternatively, FLEXICON can produce a "collective retrieved cases' profile", consisting of four quadrants which represent the union of profiles of high-ranking retrieved cases, sorted according to term weights. Rather than examine the profiles of individual retrieved cases, the user can perform relevance feedback more efficiently by examining the collective profile, which lists terms according to their overall weight in the collection of retrieved cases.

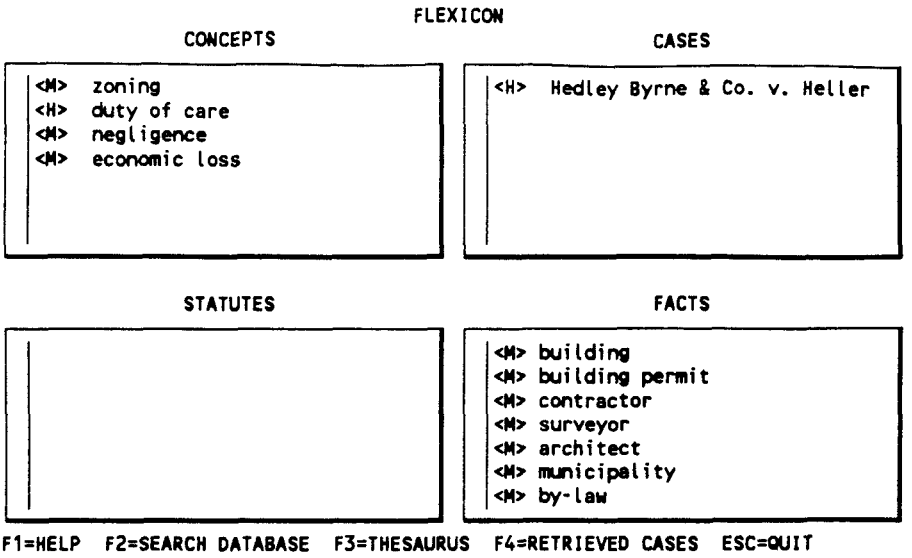


figure 2. A sample FLEXICON Search Profile Screen

| Style of Cause | Score | Histogram |
|---|---|---|
| Gadutsis et al. v. Milne et al. | 22.1 | |
| 392980 Ontario Ltd. v. City of Welland e | 21.2 | |
| The Town of The Pas v. Porky Packers Ltd | 16.1 | |
| John Bosworth Ltd. v. Professional Syndi | 15.9 | |
| Dominion Paving Ltd. v. Vaughan (Town) | 13.4 | |
| Farm Credit Corp. v. Cherniwchan | 11.9 | |
| Nielsen v. Watson et al. | 11.6 | |
| Grand Restaurants of Canada Ltd. v. City | 10.3 | |
| Dominion Chain Co. Ltd. v. Eastern Const | 9.5 | |
| Hofstrand Farms v. The Queen | 8.0 | |
| Hunt et al. v. T.W. Johnstone Co. Ltd. e | 7.5 | |
| Sulzinger v. C.K. Alexander Ltd. et al. | 6.9 | |
| Foster Advertising Ltd. v. Keenberg (Man | 6.8 | |
| Cumby v. Snow (Nfld. C.A.) | 6.5 | |
| Attorney-General For Ontario v. Fatehi e | 6.4 | |
| Central Investments & Development Corp. | 6.2 | |
| Royal Bank of Canada v. Aleman | 6.1 | |
| Sodd Corporation Inc. v. Tessis | 5.9 | |
| Nunes Diamonds Ltd. v. Dominion Electric | 5.7 | |
| University of Regina v. Pettick | 5.2 | |
| Morrison v. Mccoy Bros. Group (Alta. C. | 4.8 | |
| Silva et al. v. Atkins et al. | 4.5 | |
| Queen v. Cognos Inc. (H.C.J.) | 4.5 | |
| Blair v. Canada Trust Co. | 4.3 | |
| St. Lawrence Cement Inc. v. Farry Gradin | 4.3 | |
| East Toronto Presbytery Centennial Corp. | 3.9 | |
| Sedco v. William Kelly Holdings Ltd. | 3.4 | |
| Olsen v. Poirier et al. | 3.1 | |
| Abacus Cities Ltd. (Trustee of) v. Bank | 3.0 | |
| Hayward v. Mellick | 2.8 | |
| Moojelsky v. Rexnord Canada Ltd. | 2.5 | |
| A-1 Products Corp. v. Metro Waste Paper | 2.5 | |
| Kamloops (City) v. Nielsen et al. | 2.3 | |
| Derco Industries Ltd. v. A.R. Grimwood L | 1.9 | |
| Hendrick v. De Marsh | 1.9 | |
| Fuller v. Ford Motor Company of Canada L | 1.8 | |
| Gordonna Ltd. v. St. John'S (City) | 1.2 | |
| Heeney v. Best et al. | 0.5 | |
| Haig v. Bamford | 0.1 | |

figure 3. A sample List of Retrieved Cases

3.7 Query Library: Topic Search.

FLEXICON allows a user to formulate effective search profiles by enhancing the initial query with the related terms thesaurus and by further refining it with relevance feedback. Queries thus created can be saved and reused by the user or by others. Experienced users can create and save queries of interest, catalogued by topics. Less experienced users will then be able to select a query of interest from the library and submit it as is or modified to search the database with minimal effort. Sample queries will be provided with the FLEXICON library that demonstrate to the user the structure of well-formulated queries in key domains.

Saved queries can also be used to filter relevant cases of interest to legal professionals specializing in specific domains. A set of well-constructed queries can retrieve relevant cases and allow users to view all existing or new cases in their domain of expertise, using the flexnotes to rapidly determine relevance. Cases of interest can be printed with flexnotes, providing a private hardcopy library of selected cases.

3.8 Discussion of the FLEXICON Search.

The FLEXICON search, as demonstrated by figures (1) through (3), represents only preliminary results since the current prototype handles only a small database of 50 cases in the domain of pure economic loss. We are currently expanding the system to handle any number of cases. We will report measurements of the effectiveness of the FLEXICON search as well as comparisons to existing systems in future publications. Results reported by other researchers indicate that the effectiveness of the vector space model search used by FLEXICON compares favorable with both boolean search [Herman and Candela, 1989] and domain-specific expert systems for full text retrieval [Gey and Chan,1989]. The feasibility of using the vector space model to search very large databases (including a legal database of 40,000 cases) was demonstrated by Herman and Candela [Herman and Candela, 1989] who achieved excellent performance of an optimized system using statistical ranking similar to that used by FLEXICON.

## 4. A FLEXICON SEARCH SCENARIO.

The FLEXICON user has the option of selecting an existing query, as is or modified, from a query library catalogued by topic. Alternatively, the user can define a new search request as follows. First, the user can narrow down the search, by selecting choices from menus, according to criteria including the subject of law, date range, province or country. Next, the user enters query terms, via term dictionaries, into the four quadrants of the primary query definition screen, optionally indicating the significance of query terms. The user may then examine the related terms thesaurus to add concepts, cases and statutes related to user-entered terms. The user then performs the search and views the list of retrieved documents ranked according to their match with the query. Finally, the user can view the headnotes of retrieved documents to determine relevance and browse or print the full text of selected documents. While viewing document profiles, the user may choose to select terms that will then be added to the original search request and repeat the search. Queries which are determined effective can be saved in the query library for future reference.

## 5. PROJECT STATUS AND FUTURE WORK.

The first version of the FLEXICON system has been completed and is running on IBM PC's or compatibles. It includes text processing and case summary generation, search and retrieval and relevance feedback. The related terms thesaurus and the synonym thesaurus are being implemented. A robust memory management system is developed to allow FLEXICON to efficiently search large document databases on conventional user machines. The system will also be optimized to run on CD-ROM optical disks. An electronic database of British Columbia and other Canadian judgments is being constructed and is used to test the effectiveness of the search.

In order to make FLEXICON a true litigation support tool, we wish to explore to what degree case-based reasoning can also be automated and incorporated into the existing system to produce case retrieval as well as expert predictive capability without the tremendous manual effort required to construct traditional advisory systems. Preliminary work towards the development of an automated case-based component has begun, using the methodology developed by FLAIR to predict the outcome of legal cases, based on the disposition of similar cases. We plan to compare the relevant case retrieval and prediction capacity of the automatic case-based reasoner component suggested above to the Nervous Shock Advisor and the Whiplash Knowledge System [Smith and Deedman, 1987; Gelbart and Smith, 1990] developed by FLAIR. We will also attempt to expand the case analysis component of FLEXICON to note the history of the case, following actions by the courts that apply, distinguish, overturn, overrule or appeal the case. As well, the system will point out all the cases that cite and appeal a given case and attempt to report the case holding.

## 6. PUBLICATION PLANS

Following the development of a complete and robust system, we will attempt to make it available to the public as a commercial product. We are planning to publish FLEXICON on CD-ROM optical disks which will provide users with complete private legal libraries at low cost. As well, users will be able to print their own copies of selected cases with attached flexnotes, thereby alleviating the need for manual handling of text and reducing the cost of subscribing to a large number of printed law reports and summaries of recent cases.

FLEXICON will allow users to search large databases for relevant cases as well as to get advise about the expected outcome of legal issues, at their leisure and at a fixed price. Searching off-line, the FLEXICON user can take advantage of the related terms thesaurus and relevance feedback features to repeatedly improve the search results. The complete system on CD-ROM optical disks can be periodically updated at minimal cost. The advisory system requires no additional maintenance to reflect changes in the law, as the system's expertise is automatically updated by newly added legal cases.

FLEXICON will be published on CD-ROM optical disks, providing users with a complete library of legal documents at

low cost. As well, users will be able to print their own copies of selected cases with attached flexnotes, thereby alleviating the need for manual handling of text and reducing the cost of subscribing to a large number of printed law reports and summaries of recent cases. A CD-ROM system will also alleviate the cost of subscribing to the more costly and less effective on-line search systems.

## 7. CONCLUSION.

FLEXICON is a new, intelligent, text-based system designed for legal professionals that combines ease of use, automatic generation of electronic "headnotes" and effective case retrieval. It has been designed to serve as a bridge between the outdated technology based on manual indexing and boolean search used by existing information processing systems and the ideal system based on true natural language understanding and expert domain knowledge.

While FLEXICON provides full text retrieval, the text is processed producing document profiles which serve as representatives of legal cases. Case profiles, composed of four types of parameters meaningful to legal professionals, are used in FLEXICON not only for term indexing but as a basis for case summaries and as document representatives used for similarity matching and ranking of retrieved documents as well as for effective relevance feedback. The non-boolean FLEXICON document retrieval methodology encourages and assists users in formulating informative search requests, resulting in effective document retrieval while avoiding many of the problems characteristic of boolean search.

The FLAIR group intends to provide continuing support and enhancement to FLEXICON. The combination of a menu-driven interface, automatically generated case summaries, thesauri, an effective search and retrieval mechanism, an advisory capacity, and low cost retrieval at fixed price using CD-ROM optical disks, represents a step forward towards a complete litigation support system.

## ACKNOWLEDGMENTS

This research was supported by the Law Foundation of British Columbia and the Social Sciences and Humanities Research Council of Canada.

The FLAIR team members and student researchers who have contributed to the FLEXICON project include: Professor J.C. Smith, Daphne Gelbart, Keith MacCrimmon, Don Johnson, Bruce Atherton, Deborah Graham, Max Krause, Tanya Goldenshtein, Derek May, Chris Tennant, Martino Cavaleri, Lenneah Theroux and Ean Matthews.

## REFERENCES

Attar, R. and A. Fraenkel, "Experiments in Local Metrical Feedback in full-Text Retrieval Systems", Information Processing & management, Vol 17, No 3, 1981.

Belew, Richard, K. "A Connectionist Approach to Conceptual Information Retrieval", in Proc. 1st International Conference on Artificial Intelligence and Law, Boston, 1987.

Bing, Jon, "Performance of Legal Text Retrieval Systems" The Curse of Bool", in Law Library Journal, Volume 79, No 2, 1987.

BIng, Jon, "Designing Text Retrieval Systems for Conceptual Searching", in Proc. 1st International Conference on Artificial Intelligence and Law, Boston, 1987.

Blair, D. C. and M.E. Maron, "An evaluation of retrieval effectiveness for a full-text document retrieval system", Communications of the ACM, Vol 28 Number 3, March 1985.

Choueka, Y., "Responsa: An operational full-text retrieval system with linguistic components for large corpora", in Proc of the IALL Meeting, Jerusalem, 1985.

Dick, J., "Conceptual Retrieval and Case Law", in Proc. 1st International Conference on Artificial Intelligence and Law, Boston, 1987.

Gelbart, D.Z. and J.C. Smith, "Towards a Comprehensive Legal Information Retrieval System", in Proc International Conference on Database and Expert Systems Applications, Vienna, 1990.

Gey, Fredric and Chan Wingkei, "Comparing Vector Space Retrieval with the RUBRIC Expert System", SIGIR FORUM, Volume 23, No 1, 1989.

Hafner, Carole, "Conceptual Organization of Case Law Knowledge Bases", in Proc. 1st International Conference on Artificial Intelligence and Law, Boston, 1987.

Herman, Donna and Gerald Candela, "A Very Fast Prototype Retrieval System Using Statistical Ranking", SIGIR FORUM, Vol 23, No 4, 1989.

Hunter Morris, Andrew, "The Use of Computer Generated Abstracts", Ph.D. thesis, Business Administration, General Information Science, Texas Tech University, 1988.

Kowalski Andrzej, a paper describing the Malicious Prosecution Advisor case-based reasoner, in Proc. 3rd International Conference on Artificial Intelligence and Law, Oxford, 1991.

MacCrimmon, Marilyn T. "Expert Systems in Case-Based Law: The Hearsay Rule Advisor", in Proc. 2nd International Conference on Artificial Intelligence and Law, Vancouver, 1989.

Paice, Chris "Constructing Literature Abstracts By Computer Techniques and Prospects", Information Processing and Management, Vol 26, No 1, 1990

Rose, Daniel E. and Richard K. Belew, "Legal Information Retrieval: A Hybrid Approach", in Proc. 2nd International Conference on Artificial Intelligence and Law, Vancouver, 1989.

Salton, G. and M.J. McGill. Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

Salton, G. and C. Buckley. "Term Weighting Approaches in Automatic Text Retrieval". Information Processing and Management, Vol 24 (1988), pp 513.

Salton, G., Automatic Text Processing, Addison-Wesley, 1989.

Smith, J.C. and C. Deedman, "The Application of Expert Systems Technology to Case-Based Reasoning", in Proc. 1st International Conference on Artificial Intelligence and Law, Boston, 1987.

Tapper, Collin, "An experiment with citation vectors", Data Processing and The Law, 1984.

Tapper, Collin, "Citations as a tool for search law by computer",Computer Science and Law, 1979.