A connectionist approach to conceptual information retrieval

Richard K. Belew
Computer Science & Engineering Dept.
University of California - San Diego
rik@sdcsvax.ucsd.edu

"...(a) word is not a crystal, transparent and unchanged, it is the skin of a living thought."

Chief Justice Holmes, in Towne v. Eisner, 1918.

Abstract

This report proposes that recent advances using low-level connectionist representations offer new possibilities to those interested in free text information retrieval (IR). The AIR system demonstrates that this representation suits the IR domain well, particularly the special problems attending the more sophisticated forms of *conceptual* retrieval required in legal applications. Also, the natural way in which connectionist representations allow *learning* means that AIR can avoid the high costs associated with manual indexing while providing comparable results. The paper begins by motivating the importance of legal information retrieval, from the perspectives of both the Law and artificial intelligence (AI). Our approach is then compared to traditional methods for IR, and to more recent work using higher-level *symbolic* representations from AI. After a brief introduction to connectionist representations in general, the AIR system is presented. The paper closes with evidence that this system does, in fact, begin to support the use of those "open textured" concepts that make the Law both a very difficult and a very illuminating domain for AI research.

1 Introduction

It has been well documented that conventional approaches to information retrieval (IR) are inadequate to the task. A key problem is that these systems rely too heavily on the presence of *words* rather than the *concepts* standing behind these tokens. This is particularly true of legal IR systems. Recently, a number of researchers have investigated the use of knowledge representation techniques from AI in the hope of providing more conceptually based retrievals.

These *symbolic* representations also suffer from major problems, however. For example, they often require criterial definitions of the concepts which are not possible with the *open textured* concepts typical of the Law. Even more worrisome is the fact that the Law must be manually encoded for these systems to work. This threatens to limit the impact of this technology to a small fraction of the legal text, and prohibits adaptation of the indices with the natural evolution of legal language.

From the perspective of AI, the Law is a particularly attractive domain in which to study natural language because it at once embodies all of the centrally important questions of understanding natural language, but works with text that is crafted with more precision than most text.

Connectionism is emerging as a new, significantly different and promising new *sub-symbolic* knowledge representation technique in AI. This paper reports on experiments with AIR, a connectionist approach to conceptual information retrieval. It will be shown that this representation is very natural to the task, particularly at capturing the open texture required by legal concepts. Also, connectionist representations support some very powerful forms of adaptation quite naturally and AIR demonstrates this capability as well.

However, connectionist representations also have certain problems. In fact, the relative advantages and disadvantages of symbolic and sub-symbolic representations are quite complementary. Trade-offs between these knowledge representation techniques take a very concrete form in the legal IR task. This suggests legal IR is an interesting domain in which to explore the relationship between symbolic and sub-symbolic representations.

This paper begins by motivating conceptual information retrieval (CIR) systems as an important research area, for the Law, for information retrieval, and for AI. Next, we review some of the most important approaches to this problem. Section 4 characterizes the work on connectionist representations from which this research draws. Section 5 then describes the AIR system, a connectionist approach to the problem of free-text information retrieval. Several properties of this associative approach to IR that seem particularly useful in the legal domain are demonstrated. AIR also adapts its representation of keywords and documents on the basis of the browsing patterns of its users. It will be argued that such adaptation is an important property of CIR systems. However, a description of AIR's learning algorithm is beyond the scope of this paper. The details of this algorithm, its relation to other connectionist learning schemes, and the performance of the algorithm in simulation and in a human subjects experiment is reported elsewhere [3]. The paper concludes with a description of our current work aimed at integrating this connectionist representation with taxonomic and thesaurus information, represented using more standard, symbolic knowledge representation techniques.

2 Motivations for conceptual information retrieval research

2.1 Working without a legal calculus

Broadly speaking, work in computer applications to the Law fall in one of two categories. It is either based on work in analytic jurisprudence or it automates some manual function lawyers must perform frequently. The former approach is based on the belief that their exists a *legal calculus*, perhaps based on deontic logic, that will someday provide a solid *mathematical* foundation for legal reasoning.[1] This work has and will certainly continue to extend our understanding of both the Law [1] and automated reasoning. But as [13] and others have noted, "The methodology for ascertaining the Law cannot be based on a legal calculus for (at present) none exists." (p.78)

The second, more modest approach — automating tasks that are already part of standard legal procedure — has allowed important advances in legal applications of computing in the absense of such a legal calculus. Many legal tasks have been analyzed and partially automated, as demonstrated by the wide range of applications reported at this conference.

Legal information retrieval seems particularly promising as a mechanism for dramatically altering the way law is practiced, according to a recent workshop on legal expert systems [11], and it is no wonder. Text is the stuff of the Law, in a very fundamental sense. Lawyers live in a world filled with documents, legislation, opinions and regulations, and all of these sources of text are increasing at a staggering rate. To date, most legal IR applications have been aimed at nationally vended information sources, shared by many lawyers. Recent advances in computer hardware, particularly the optical disk, promise to make similar technology available to individual lawyers and firms for in-house litigation support. In short, the need for legal IR systems is real.

According to a recent ABA survey, almost half of all lawyers are in a one-person firm and two-thirds are in firm of less than three. It seems, then, that the majority of all lawyers are *generalists*, having a small set of clients for whom they provide a broad range of legal services, rather than *specialists* in one particular area of the Law serving a large number of clients in only one way. For the CIR system builder, this is good news. It is difficult to imagine that a CIR system could provide many relevant citations unknown to a lawyer who is expert in that field. But it seems not at all unlikely that even rudimentary aids could be very useful to lawyers who are novices, relative to a particular legal area. The ABA statistics indicate that the majority of legal IR system users are in this category.

2.2 The value of the Law to AI

Lawyers are trained to use more precise and consistent language than the average writer, and it can be argued that legal prose is therefore more amenable to computer analysis (than newspaper articles, for example). The fact that the Law attempts to describe situations in the most complete, consistent, and unambiguous manner possible gives hope to the prospect of successful analysis of legal natural language. As Stamper [31] notes,

> ...legal prose in the hands of a skilled draughtsman is purged of those logical complexities and ambiguities which make the rigorous interpretation of ordinary language difficult, if not impossible. The logical and semantic structures employed in legal prose being relative simple ones, it is hoped to explicate them fully.... This is much less ambitious than attempting to model in a computer the understanding of ordinary language. (p. 103)

This is certainly not to say that legal writing is ideal. One of the most interesting facts emerging from Laymen Allen's analysis of legislation is that a great deal of the Law contains ambiguity [1]. In fact, at least some of this ambiguity is intentional: legislators unable to resolve their differences of opinion instead hide it behind ambiguous phrasing.

Nevertheless, lawyers are professionally trained to use natural language as precisely as possible. It is important to contrast the lawyer's precise use of *natural language* with the precision achieved by computer professionals using *artificial* programming languages. The Law and computer programming are alike in that they both form large, coherent systems for describing their respective worlds in consistent and rational terms, but there the similarities stop. The legal system has been constructed in full knowledge that the world is in fact a very *noisy*, uncertain and inconsistent place, whereas the world of computers is only beginning to confront these issues, particularly in the context of AI. In fact, some some of the Law's special mechanisms designed to manage ambiguity (e.g., voting, jury trials, dissenting opinions) may provide important ideas to computer programmers for how their systems might also manage the noisy world.

All this again suggests CIRs as an interesting domain in which to explore natural language understanding. It at once embodies all of the centrally important questions of understanding natural language, but works with text that is crafted with more precision than most text samples. Also, this precision takes a radically different form than typically used with computers. For all these reasons, the Law can teach AI a great deal.

2.3 The need for adaptation in IR systems

After a long period of disinterest, AI has again returned to the problems of getting computers to *learn* or *adapt* as a result of their experience. These questions are now recognized as critical to AI, in part because of the tremendous task of manually programming all of the knowledge intelligent systems need, but also because any reasonable definition of intelligence requires the ability to improve in the face of experience.

There are at least two very good reasons why adaptation must be a *sine qua non* for any realistic IR system. The first is that the additional structure required by conceptual information retrieval (as described in Section 3) is extremely expensive to develop manually. When this structure takes the form of editorial enhancements (as in the WESTLAW system), it means that lawyers (already an expensive resource) must be trained still further in the stylistic and indexing conventions being used. Given the workloads and time-pressures facing these editors, it is difficult for them to build consistent, accurate structures. If even more structure is being added (as proposed by Hafner and Krovetz; see Section 3.3 below), the lawyers must be trained still further, as knowledge engineers.

The cost of manually constructing the additional structure required by conceptual information retrieval may well be justified for certain, particularly important document bases (e.g., Supreme Court decisions). But for the vast majority of legal text it is difficult to find adequately trained editors/knowledge engineers and justify their expense. Adaptive mechanisms offer the benefits of conceptual information retrieval without the costs of manual indexing. A less subtle but equally important form of growth every IR system must be able to gracefully accomodate is the constant incorporation of new documents.

[1] Or at least that *logical* quarter of the Law defined by [7].

However, there is a second reason for requiring adaptive mechanisms in an IR system: The indexing vocabulary of IR systems, like other forms of natural language, is fundamentally *dynamic*. This problem is shared by automatic indexing techniques, as well as those for which the cost of manually adding structure is justified. However the indexing structures are generated, the meaning of keywords in an IR

In summary, any IR system that does not have adaptive capabilities has fundamental limitations. If it is to benefit from the additional structure used in conceptual information retrieval, this structure must be added manually, at very high cost. And even if these manual enhancements are made, a non-adaptive IR system immediately begins to obsolesce as its vocabulary ages.

It is worth noting that the converse is also true: the IR problem also offers some advantages as an area in which to study adaptation. Because IR systems retrieve probabilistically and because there will never be a single, correct answer to a user's query, the system is allowed to make "mistakes." This is a critical feature for any learning application, because learning systems *must* be allowed to make mistakes. Users of a database system would not tolerate mistakes (such as retrieving an incorrect salary). But because IR systems are fundamentally probabilistic, a learning algorithm which promises to improve the probability of correct retrievals will be allowed to respond incorrectly, since even traditional, non-adaptive IR systems do so.

3 Approaches to conceptual information retrieval

It has been well documented that conventional approaches to information retrieval (IR), while potentially very useful, are inadequate to the task. For example, in one post-hoc analysis, Maron and Blair showed that a state-of-the-art IRS was able to deliver only 15% of documents (subsequently) judged "relevant" and 30% of documents judged "critically relevant" [4]. There is therefore great room for improvement. A key problem is that these systems rely too heavily on the presence of *words* rather than the *concepts* standing behind these tokens. This is particularly true of legal IR systems. Here we highlight the methods behind two of the most important LIRS now in commercial use (LEXIS and WESTLAW), and also describe some recent attempts to use knowledge representation techniques borrowed from AI to codify the semantic information contained in documents.

3.1 Full-text indexing

The most common approach in IR systems is to automatically index documents. Almost universally, this means generating an inverted index for all keywords of the text, so called full-text retrieval. The advantages of such a thorough index are first that retrieval can obviously be based on any word or combination of words, and second that the retrieval procedure is simply explained to a user.[2] The simple but devastating problem with this approach is its reliance on a simple occurrence of a word. One of the most basic results of modern linguistic analysis is that the presence of a particular word *token* is neither sufficient nor necessary evidence of a reference to the concept standing behind that word [35]. For example, the concept of CAR can be used without actually using that word (by using the words AUTO or VEHICLE instead). Conversely, the presence of the word CAR does not necessarily mean that one is discussing things with four wheels (perhaps the first element of a list in the programming language LISP).

LEXIS is the least structured system to be considered. Each legal document is represented by rudimentary citation indices (court, date, judge) as well as the full text. More recent embellishments to LEXIS include AUTOCITE, which provides WAS-CITED-IN links to all those cases referring to the current document. Since these indices and links can all be generated automatically, the addition of new law to the database is a relatively simple task. It will become even simpler when: a) character recognition technologies are perfected; or b) LEXIS obtains access to the machine-readable version of the Law generated as a by-product by legal publishing houses.

Much of modern IR research has concerned itself with refining full-text retrieval by analyzing word occurence frequencies within a particular document and comparing them to the word frequencies occuring in the entire document.[3] There are certainly very sound arguments for the information contained in such word frequency analyses; in fact, the system to be described in Section 5 *begins* with just this type of data. However, any system which depends completely on the occurence of word tokens, or statistics derived from these, is inherently limited. Such statistics provide a piece of the IR solution but cannot do it all.

3.2 Editorial enhancement

One way to enhance full-text retrieval is to *manually* add indexing information to each document in the text-base. The manual procedure has a person who is expert in the subject of discourse (for example, the Law) read each document and provide *editorial enhancements*, i.e., text and indices ancillary to the actual text of the document.

In the WESTLAW system, the editorial enhancements take the form of appropriate keywords, taxonomic classification in West's *key number* system, and summarizing head notes. West's key numbering system is particularly interesting because it was developed to facilitate searching of West's printed volumes, long before computerized LIRSs were available, and has been used by lawyers and refined over a period of many years. It can be viewed, therefore, as a legal "artifact" reflecting important relationships of the Law it represents.

Sprowl highlights the trade-offs offered by WESTLAW [30]:

> ... Naturally, some information is lost in the process of condensing judicial decisions into headnote summaries, but something else is gained if the headnotes are clear, concise, and well-indexed.

The "something else" is the additional structure provided by the system. Its structure models to some extent the structure of the Law, and this additional legal knowledge is one of the features WESTLAW has to offer.

As Sprowl notes, there are problems with editorial enhancement as well. If the editors characterization of a document is accurate, the indexing will help users, but if it is inaccurate the document is effectively hidden. And considering the time pressure under which most documents must be classified by editors, as well as the fact that each document's classification is subjected to the idiosyncratic biases of each individual editor, such misclassifications are almost guaranteed to oc-

[2] This second advantage becomes significant when full-text retrieval is compared with more sophisticated approaches which may be theoretically more desirable, but are also more difficult to explain and justify to a user.

[3] It is interesting to note that while the use of these statistics, particularly in the form of *weights* on the indices and query terms, has been discussed in the literature for 15 years, almost no commercially vended IR systems make use of this feature. One reason may be the fact that retrieval using weights is more difficult for the user to comprehend.

cur. Finally, the entire process of manual indexing — having each document read by an expert in the field — is extremely expensive.

3.3 Representing semantics using knowledge representation

Recently, a number of researchers have begun to investigate the use of knowledge representation techniques from AI in the hope of providing more conceptually based retrievals.

Hafner's Legal Information Retrieval System (LIRS) is as different from WESTLAW as that system was from LEXIS [12]. Based on an elaborate model of legal knowledge, each document is given a much more subtle description than is possible with either of the previous examples. LIRS uses a *document description language* (DDL) to formally describe the contents of each case. Basic legal concepts such as FORGERY, and DOCUMENT are combined to describe situations. These concepts and situations are then used to describe particular aspects of the case such as:

- What was the legal basis of the lawsuit?
- What legal situation does the case exemplify?
- What other cases were cited as supporting, or as not controlling?
- What was the court's decision?
- Were formal, defining criteria of a legal concept enumerated as part of the decision?

Clearly, this is a much richer description of a document than a list of key-numbers. Just as clearly, a substantial effort is required to code any case into DDL. First, an enormous vocabulary of basic legal concepts must be defined and related; Hafner's experimental system dealt only with a small corpus of documents on negotiable instruments. Then individuals trained both in the Law *and* in the peculiarities of DDL must painstakingly translate each case into the LIRS formalism. Only then it the system ready to assist a user.

Krovetz proposes the use of KRL, another important frame-based representation [18]. He cites KRL's ability to organize knowledge along a number of dimensions and subsequently retrieve it via comparison with related concepts as particularly important in legal applications.

These projects both propose to support CIR by augmenting the raw text of documents with a characterization of its content, written using a sophisticated knowledge representation language. Both Hafner's and Krovetz's use frame-based representations, but semantic networks [22], logic programming [19], heuristic state space [6], even the relational data model [31] have been used for legal applications. Each of these representations offers a different set of epistemological primitives and supports different inference procedures. A careful consideration of exactly which of these is most suited to the CIR task is therefore worthwhile.

There is a fundamental flaw inherent in all such *symbolic* representation schemes, however. These languages make use of the rigor and deductive power of *logic* and hence are also severly limited by logic's inability to deal with the imprecision and inconsistency of the world. For us, working on the problem of LIR, this lack means that logic-based KRLs are unable to represent *open textured* concepts. This is an extremely important deficit, and one that the AIR system addresses directly; this argument is presented in Section 6.

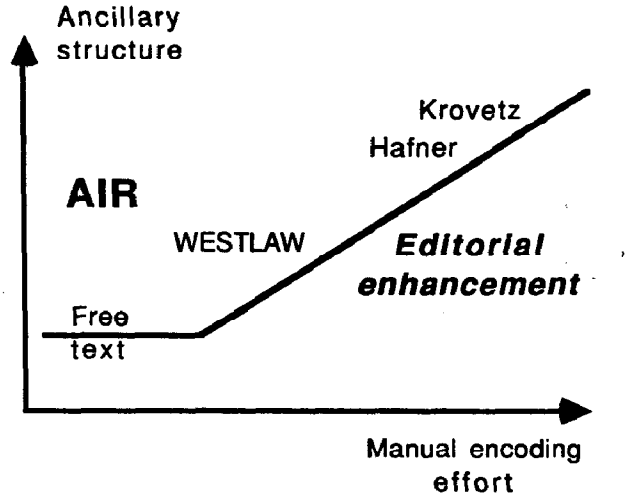

Figure 1: Approaches to CIR

3.4 Summary

AIR's relation to the preceding three approaches is shown in Figure 1. All of these systems vary along two important dimensions: the amount of *additional structure* they provide to support CIR, and the *manual effort* expended in providing this structure. Full-text indexing systems make as much use as possible of the purely *syntactic* information derived from the occurence of word tokens. There appear to be fundamental limits to this source of information, but it also makes minimal demands in terms of manual indexing effort. Traditional editorial enhancements (like WESTLAW's headnotes and key numbers), as well as more recent attempts using knowledge representation languages (like Hafner's and Krovetz's), both rely on highly-trained people to encode the *semantic* content of each document. These systems currently offer the only means by which this critical form of information can be used for sophisticated CIR. The AIR system, to be described in Section 5, promises to generate this additional structure automatically, using machine learning techniques.

4 Connectionist representations

Within the last few years there has emerged a distinctly different approach to knowledge representation. These representations are a response to perceived limitations of the *symbolic* representations used by most current AI systems. Each of the traditional knowledge representation schemes mentioned in Section 3.3 are rooted firmly in the *symbol processing paradigm* in which intelligent behavior is assumed possible only by systems which manipulate *abstract* symbols. A critical assumption in this paradigm is that symbols are, in fact, abstractions; i.e., their internal structure is irrelevant; Newell has called this the "physical symbol system hypothesis" [25].

The essence of the connectionist approach is that it finds this internal structure not only relevant but of central importance. In connectionist representations, symbols correspond to *ensembles* of *sub-symbolic* elements rather than to any one element. This fine-grained approach to representation gives connectionist networks both their power and their limitations. In fact, connectionist networks are better viewed as a *complement* to symbolic representations than as a replacement for them. Connectionism approaches the task of representing knowledge

at a different, lower level than that used by AI's more traditional representations. Semantically meaningful concepts correspond to large, distributed sets of units, rather than to one particular unit. This work therefore can be viewed as rejecting Newell's physical symbol system hypothesis.

While the interest in connectionist representations is fairly recent, many of its basic tenets hark back to much earlier work. In many ways, the connectionist approach could be called *neo-cybernetic*. The nets being used are not very different from the nerve-net models investigated as far back as 1943 [21]. Going back even farther, William James emphasized how fundamental simple associations are to human thought [17,16]. One feature of the recent connectionist work is that it is less concerned with replicating psychological behaviors or neurophysiological data than with building mechanisms that satisfy certain information processing goals. This moves the research much more squarely into computer science. Recent interest in models for massively parallel computation provides a second important motivation for interest in connectionist models on the part of computer scientists.

The basic structure in connectionist representations is a weighted graph. It is important to note that the "knowledge" in these systems is captured exclusively in the weights on the links. "Programming" a network implies putting a particular set of weights on links, and "learning" implies changing the weights. Processing is effected by propagating a transient quantity, called *activity*, throughout the network. Activity is a real-valued quantity, typically between zero and one, although some systems (including AIR) allow negative activities (between 0 and -1) to propagate. It should be noted that the *state* information in connectionist systems is therefore extremely distributed: each node and each link contains approximately one computer word. The complexity of these systems results from the fact that the *global* state depends on the many, subtle interactions among simple, local elements. The recent two-volume text by Rumelhart, McClelland et al. is a good primer on the various design parameters of connectionist systems [28].

One key property that makes connectionist systems attractive is that they are *reconstructive*:

> This means that the system yields the entire output vector (or a close approximation to it) even if the input is noisy or only partially present, or if there is noise in the memory matrix. [2], p. 19.

In the context of CIR, this property means that queries need not be exact; queries that are "close" to the descriptions of documents will cause them to be retrieved. This ability is at the heart of AIR's ability to capture the nuances of open textured legal concepts (see Section 6).

A second critical feature of connectionist representations is that there is a wealth of ideas on how to allow *learning* in these structures. One class of results date back to Rosenblatt's work on the Perceptron [26]. However, the Perceptron was shown to have serious shortcomings [23], and theoretically well-understood learning algorithms for more elaborate connectionist structures have only recently been discovered [15,27].

Connectionist systems have now been successfully applied to tasks in vision, linguistics, and speech recognition and generation [28,20]. The AIR system applies a connectionist representation to the problem of free-text information retrieval. Our project has demonstrated that this is an appropriate representation for the task, and also suggests ways in which the sub-symbolic connectionist representations might be merged with more symbolic forms of knowledge.

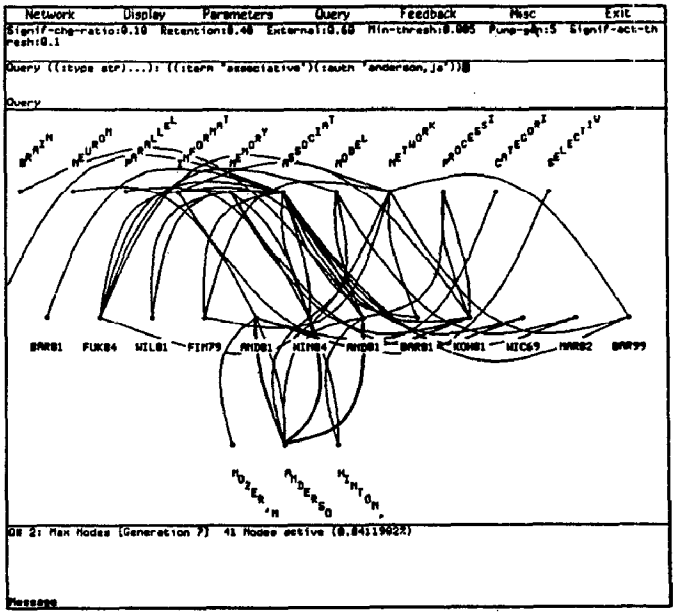

Figure 2: AIR's user interface

5 The AIR system

The AIR[4] system represents a connectionist approach to conceptual information retrieval. This section will first give an overview of AIR's retrieval process, and then discuss some details of the basic mapping from the IR problem into a connectionist representation.

5.1 Overview of the retrieval process

In AIR, the user's query causes some activity to be placed on each of the nodes corresponding to features mentioned in the query. This activity is allowed to propagate, first to the immediate neighbors of the query nodes, then on to their neighbors, and so forth. AIR's "answer" are those nodes whose maximum activity level reaches significant levels before the propagation terminates.

Figure 2 shows AIR's interface during a typical query. The window is divided into five panes, but only two are of real interest. The very top line is a command bar; these commands were used by me but not, in general, by "real" users. Moving down, the second window and the very last window contain statistics about the net, parameter settings, and details about the retrieval process. The third and fourth windows are the interesting ones.

The third window is where the user types the query (in this case asking for information about the word ASSOCIATIVE, as used by J.A. ANDERSON), and AIR responds by drawing the network shown in the fourth pane. The nodes of this network correspond to the documents, keywords, and authors AIR thinks are relevant to the user's query. AIR draws this network response as a tri-partite graph: nodes on the top level correspond to keywords, the middle level to documents, and the bottom layer to authors. The arcs connecting nodes represent the associative links between keywords, documents and authors. Heavier lines imply stronger weights. AIR uses directed links whose directionality is represented by the concavity of the arcs; a clockwise convention is

[4] AIR stands for Adaptive Information Retrieval.

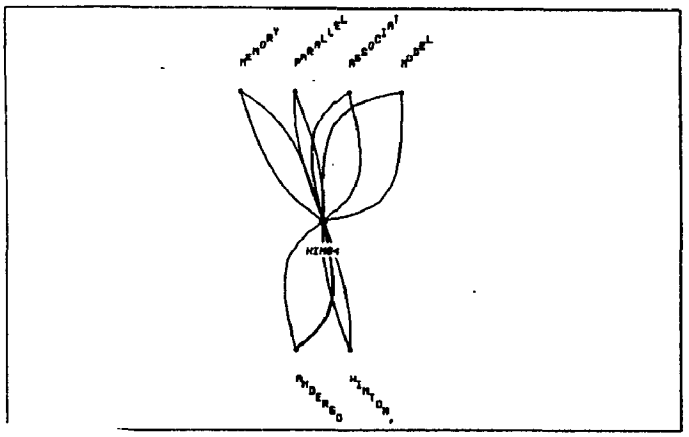

Figure 3: Network corresponding to one document

used. For example, a link from a document node (in the middle level) to a keyword node (in the top level) goes clockwise, around to the left.

5.2 Mapping the IR domain into a connectionist representation

As with most connectionist systems, AIR users a weighted graph as its basic representation. In the experiments to be described below, the only information used about a document is its title and authors. Titles are obviously a very limited sample of free text, but the methods to be described promise to "scale up" to larger samples of text, such as abstracts or full text of the documents. Attributes other than other author's names (such as publication data) could also be easily introduced (see Section 6.5).

Unlike most connectionist systems however, AIR does not begin from scratch (i.e., with a randomized network) but from a network constructed from the initial representations of the documents it contains. As users query and interact with the system AIR changes this representation, modifying the representation of documents, keywords and authors.

Figure 3 shows that portion of the network corresponding to a single document. The initial topology (again, this will be changed through adaptive mechanisms) of AIR's network is tripartite, with the top level corresponding to keywords, the middle level of nodes corresponding to documents, and the bottom level of nodes corresponding to authors.

Each citation first causes a corresponding document node to be generated. One author node is generated (if it doesn't already exist) for each author of the document. The basic questions of indexing arise with the title: what term nodes should be generated from the words in the title? AIR's initial indexing is again about as straightforward as possible: basically every word in the title becomes a term node.[5]

Symmetric links are then formed from the document to each of its keywords and back, and from the document to and from each of its authors. Weights are assigned to these links according to what is known in IR as an *inverse frequency* weighting scheme [29]: the sum of the weights on all links going out of a node is forced to be a constant. This has the desirable effect of making a node with high out-degree (i.e., many *out neighbors*) have a low connection weight to each of them, while a node with only a few out-neighbors is relatively strongly connected to those.

There is a second, independent motivation for using inverse frequency weights, however. One way of insuring that the weight from a keyword to the documents it indexes is inversely proportional to the number of such documents is to make the total associativity from an index term constant. If this constant is made unity, the resulting network has the very satisfying property of conserving activity. That is, if a unit of activity is put into a node and the total outgoing associativity from that node is one, the amount of activity in the system will neither gain or diminish. This property is obviously helpful in helping to control the dynamics of an associative network.

Notice that while every link will have an inverse the two links will, in general, have different weights. Allowing asymmetric connection strengths makes AIR fundamentally different from many other connectionist systems (notably the Boltzman Machine [14,15]).

The initial network is constructed from the super-position of many such documents' representations. Most of the experiments to be described in this report used a network constructed from 1600 documents, forming a network of approximately 5,000 nodes. The next section contains more details about the composition of this network.

6 Supporting open-textured legal concepts

It is proposed that these representations can be particularly useful for the representation of legal concepts. A key feature of legal concepts is their *open texture*: the inherent ambiguity of natural language used in the Law permits inherent indeterminancy in the classification of fact situations [34,33]. A criterial representation for these concepts, in terms of necessary and sufficient conditions, is rarely available. In fact, the adversarial system of the Law is designed exactly to deal with this fundamental ambiguity. Also, legal concepts are inherently dynamic. The meaning of "fair use," for example, must be allowed to evolve as the term's usage adapts to a constantly changing world (in this case in the form of emerging technologies).

AIR brings two basic mechanisms to bear on the problem of CIR. First, AIR's associative retrieval method relies upon many, many pieces of relatively weak information. This makes it both stable and robust, in that small perturbations in either the user's query or the document's description is not critical. Also, AIR's adaptive mechanisms move to solidify initially "serendipitous" retrievals, while culling out inappropriate ones. The following example will illustrate both mechanisms.

This section will higlight some of AIR's characteristics that seem most valuable towards the retrieval of open textured legal concepts.

6.1 Variable recall/precision

Figure 4 shows the most typical perspective of the IR problem based on this assumption. Consider the universe of all documents contained in an IR system, and also consider two subsets of this universe: **RET**, the set of documents retrieved by a particular query; and **REL**, the set of relevant documents that *should* have been retrieved by that query, as judged omnisciently. Abstractly, the goal of an IR system is to make **RET** match **REL**.

[5] There are a few refinements, however. First, a "noise word" list of extremely common words (e.g., the, of, and) is maintained; these are not indexed. The noise word list used is very small (134 words) but using even this small list reduced the set of nodes significantly. Second, pluralized nouns are changed to their singular form. This procedure is somewhat sophisticated (more than just trailing S's are removed) and seems adequate. Punctuation, any numbers less than 100 (the theory being that numbers like 2001 may be meaningful) are also removed, and all words are then capitalized.

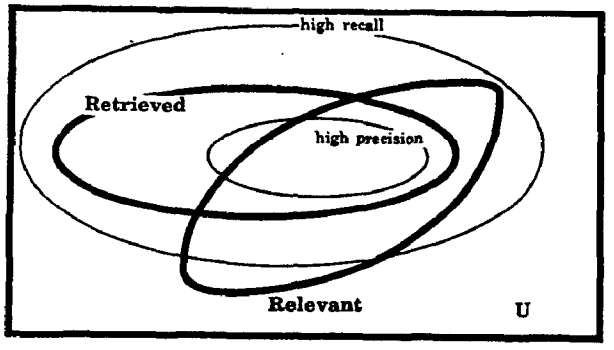

Figure 4: Relevant vs. retrieved documents

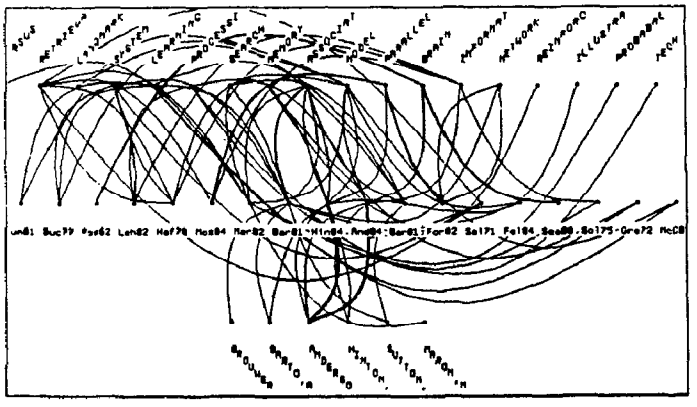

Figure 5: High recall query

Notice there might be two approaches to retrieval by the IR system. The first approach would be to retrieve a small, conservative set of documents that were very likely to be relevant. Unfortunately, there would also be many other relevant documents not retrieved. This is called a *high precision* retrieval. The other approach is to retrieve a large set of documents that almost certainly encompass all relevant documents. The difficulty with this approach, however, is that it also retrieves many irrelevant documents. This is called a *high recall* retrieval. Real-world IR systems typically attempt to compromise between these two extremes. Also, users vary widely as to what type of retrieval they prefer. Typical users tend to want high-precision retrievals; rarely do they need or expect to get everything relevant to their topic. Lawyers, on the other hand, tend to need high-recall retrievals [30]. If they miss one case "on point" that their opponent may discover, it is a costly mistake. They are willing, therefore, to wade through many irrelevant documents if they can be assured of retrieving all relevant documents.[6] The point is that variable selectivity is a critical feature for IR systems.

AIR provides a very convenient mechanism for varying from high precision retrievals to high recall retrievals. Figure 5 shows the result of the same query used in Figure 2, but with a single parameter having been lowered.[7] This shows simply that AIR can respond readily to varying user requirements for high precision or recall. Traditional IR

[6] It is worth noting that the evaluation by Maron & Blair mentioned in Section 3 was designed especially to meet the high-recall requirements of lawyers. Despite this fact, the resulting IRS is very far from perfect.

[7] *SIGNIFICANT-ACTIVITY* is the parameter of the system that sets a threshold above which nodes are considered to be "significantly" active.

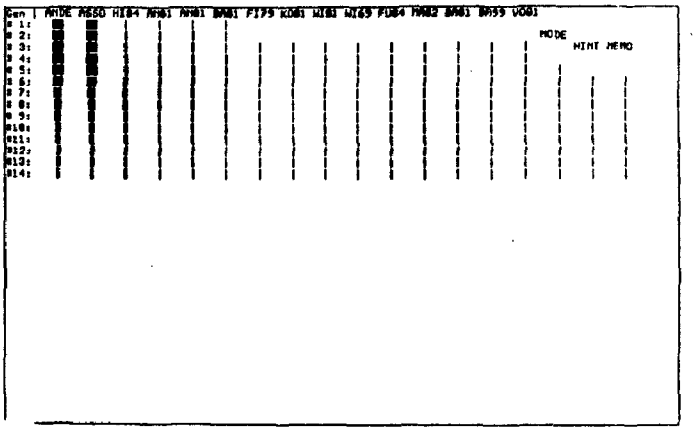

Figure 6: Bar graph view of activity

systems with weighted links can accomplish some but not all of this same behavior.

6.2 More communicative user interface

A more significant difference is that the "input-output channel" from and to users has been widened by the AIR system. Typically, queries to IR systems are composed of keywords; it is also common to be able to specify authors of interest. But AIR also allows specification of *documents* in a query. The provision of this sort of this "query by example" seems a very useful extension language.[8]

The result of AIR's retrieval is even more uncommon. The traditional result of an IR query is only documents (or more typically, citations to or proxies of documents). While this is AIR's major output as well, the system also provides keywords and authors. Keywords retrieved in this manner are considered *related terms* that users may use to pursue their searches. Retrieved authors are considered to be closely linked to the subject of interest.

It could be argued that these keywords and authors have no intrinsic value but are useful only to the extent that they ultimately lead to relevant documents. However, there are many ways in which a user might find related terms and centrally involved authors a valuable information product in their own right. For example, if a user wants to pursue his or her search in other information systems (such as a traditional library), these additional cues can be very useful. The fact that users had no more difficulty judging the relevance of keywords and authors than they did judging documents supports this view (see [3] for details of these experiments).

6.3 Generalized Boolean queries

AIR's feature query language is somewhat unusual in that it is not strictly Boolean. A query is specified by mentioning a set of *features* (keywords, authors, documents), or their negation. No provision is made for the traditional Boolean connectives AND and OR.

Figure 6 suggests that they are not missed, but this representation requires some explanation first. Here, the temporal dimension of AIR's retrieval process has been made explicit. Nodes in the network now

[8] Mike Moser was the first to point this out [24].

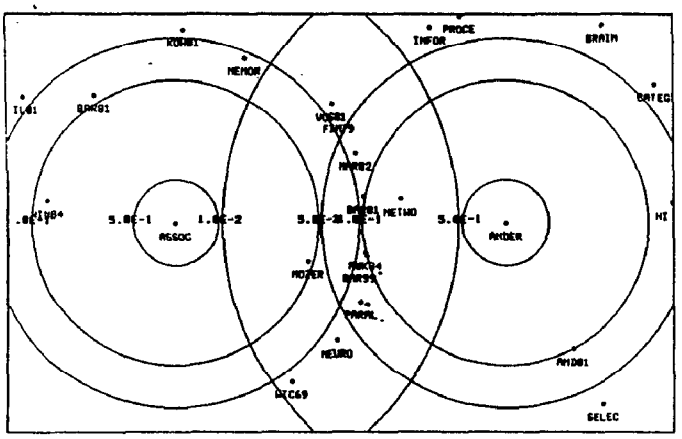

Figure 7: Concentric view of propagation

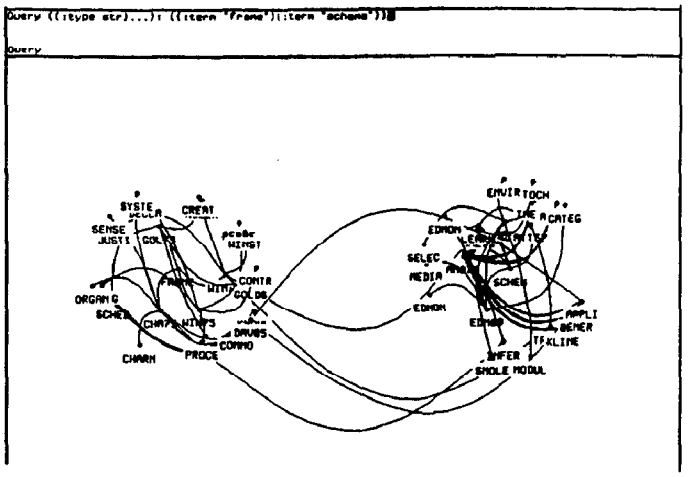

Figure 8: Concentric view of propagation - with links

run across the horizontal axis, and time steps of the model are listed vertically. The filled rectangles below each node represent the activity level of that node at each time step. The typical progression is for the node to begin at low activity levels (drawn as thin rectangles), to swell to higher activity levels (drawn as fatter rectangles), and finally to dwindle away to inactivity. Notice that the column labels for each node do not occur all on one line, but in staggered groups. Because only a small fraction of the network becomes active during any one query, nodes are only included in this representation as they become active. Those nodes active during the first time step of the model are labeled on the first line, those that first become active on the second time step are labeled on the second line, and so forth.

Beginning at the left, the first two nodes (ANDERSON,JA and ASSOCIATIVE) were mentioned directly by the query and so are not considered part of the retrieved set. The next node corresponds to a document that satisfied the *conjunction* of the two query features. The next set of three documents are in the *disjunction* of the two features, and the remaining nodes are not directly associated with either of the query features but were activated by other, indirect associations.

There are several things worth noticing in this representation. First, notice that the retrieved nodes are sorted from left to right from those most obviously implicated by the query to nodes less directly implicated. The first nodes on the far left are the query nodes. Next are those nodes directly associated with one or more of the query nodes. Then, in "serendipitous order," are nodes somehow indirectly related to one of the query nodes. Second, it is important to know that AIR will use maximum activity levels of nodes to determine whether they are retrieved; in this representation this corresponds to the fattest box in each column. Notice that this point occurs later and later in more and more remotely connected nodes.

The point is that *the difference between AND and OR is a matter of degree*; this insight goes back to von Neumann. A user searching for relevant documents does the best he or she can to describe *features* of the set in which they are interested, but a Boolean language may require more precision from this description than is appropriate. AIR's feature language asks for less rigor, but generates retrievals that turn out to be a natural generalization of Boolean-like languages.

Figure 7 shows a third way in which AIR's results can be represented. In this display, each query node becomes the center of a set of concentric circles. The rings correspond to progressively smaller activity levels, drawn on a logarithmic scale. As nodes are retrieved, they are drawn close to the query node whose activity first reached them, at a distance proportional to their activity level. The point of this representation is to suggest the interactions between the spreading activation waves from multiple point sources (i.e., query nodes). A node can be significantly active because it is very close to one of the query nodes or because it is somewhat close to several query nodes. This representation distinguishes between those two situations. Figure 8 is basically the same representation, but the logarithmic circles have been removed and the lengths between nodes have been added. A different query — FRAME & SCHEMA — is shown to provide an example with non-obvious connections between the query features. It is also drawn to a slightly different scale in order to highlight with long links the significant interactions across the clusters immediately surrounding the query nodes.

6.4 The extensibility of the representation

One key advantage of connectionist representations is that it proves quite easy to encode additional, new information. Because connectionist links correspond to simple associations, and because association is such a basic, common, generic relationship in the world, representation becomes particularly straight-forward. In AIR, for example, it took no great insight to determine how documents might be initially associated with keywords and authors. [9]

Similarly, it is straight-forward to incorporate many new forms of information, some of which are shown in Figure 9. For example, adding critical citation information is a very natural and promising extension to AIR's current representation. In the law, this type of information is often referred to as Shepard's index, and its use in legal CIR systems has been investigated [32]. Eugene Garfield has pioneered the use of this information in the sciencific literature with his *Science Citation Index*. In AIR, citations will be represented as links from the *citing* document to the *cited*; reciprocal links will also be valuable. Other possible sources of information that could be incorporated into AIR's knowledge base just as easily include:

- *thesaurus* relationships among keywords (e.g., broader term, narrower term, related term relations). This is an especially impor-

[9] Notice how much more difficult the task becomes if the links are *labeled*, as in semantic networks.

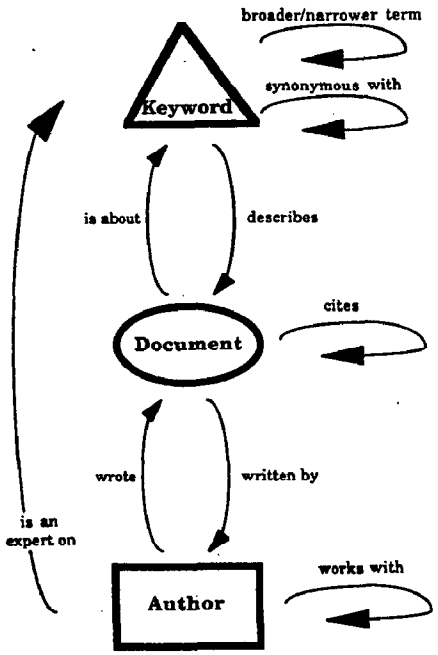

Figure 9: Available information

tant form of information; see the next section;

- *colleague* relationships among authors (e.g., common research institutions, common educational histories);
- *expertise* relationships between authors and keywords (e.g., personal descriptions of research interests, often used for selective-dissemination-of-information (SDI) applications)

As with the initial indexing information, some but not all of these *syntactic* facts will prove significant. This data is merely a plausible basis on which to begin retrieving documents. It is the *adaptive* side of the AIR system that will move to codify the beneficial relationships into permanent structures while pruning disfunctional ones.

6.5 Integrating symbolic information into a sub-symbolic representation

Our current work is aimed at just the sort of extension described above, viz., integrating thesaurus information. The reason this is an especially important type of knowledge to add is that it functions as a critical experiment towards understanding the relationship between symbolic and sub-symbolic representation techniques (see Section 4). Use of such taxonomic information is a critical componenet of the WESTLAW key numbering system. The semi-automated construction of such structures has been investigated [8], suggesting more such information may become available as the task of generating it becomes less onerou.

It would be possible to represent thesaurus information as simple weighted associations. However, a great deal would be lost in the translations. To say that one keyword is a `BROADER_TERM` than another keyword is to say more than that there exists a weighted association between them, whatever the weight; the same is true of the inverse `NARROWER_TERM` relation. It is true that the `RELATED_TERM` relation is well-captured by simple weighted associations, but this merely serves to highlight the additional *semantics* attending the hierarchic relations.

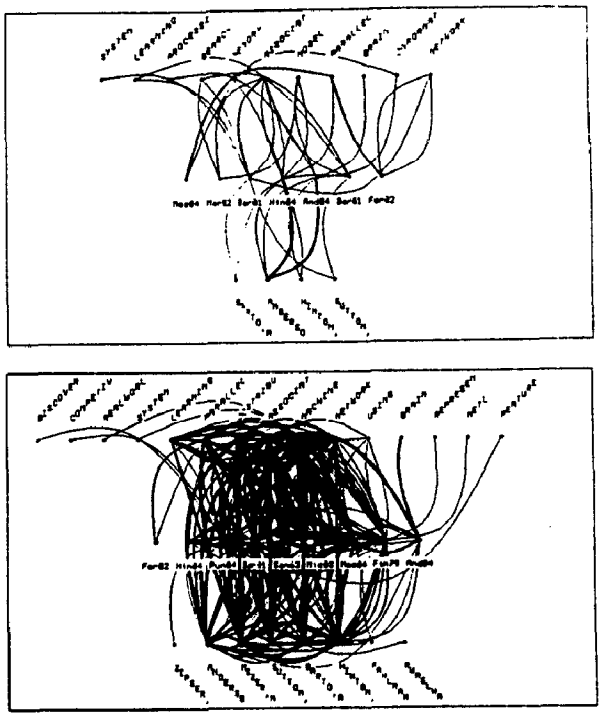

Figure 10: The symbol `ASSOCIATIVE` - before and after

Representing such semantic information, symbolically, has been a central concern for AI almost from the beginning, and it is arguably the area in which most progress has been made. The `BROADER_TERM` / `NARROWER_TERM` relations, for example, are simply a special case of the `IS_A` relation used by many knowledge representation languages [5]. We intend, therefore, to make use of these symbolic representations for the thesaurus information. The question then becomes one of getting this symbolic information to interact in a meaningful way with the sub-symbolic information captured by AIR's existing connections.

We are investigating several mechanisms for integrating the two representations. For example, it is possible to selectively propagate activity along `IS_A` links, similar to the marker passing procedure used in Fahlman's NETL system [10]. This quasi-logical processing would augment and interact with the *uniform* propagation along weighted links currently used by AIR. A second possibility is to add strict Boolean operations to the query language (as opposed to the generalized Boolean operations implicit in AIR's current query language). These would result in (logical) set operations performed *post hoc* on intermediate document sets retrieved in the standard fashion. A third possibility is to use symbolic and sub-symbolic search procedures redundantly, with the two forms of inference interacting via a "blackboard" architecture [9].

We believe that the most important form of interaction between the symbolic and sub-symbolic forms of knowledge representation, in the context of CIR and more generally, will be the *induction* of symbolic structures from sub-symbolic ones.

Figure 10 shows the sort of symbols built by the AIR system. In Figure 10.A the net shown is AIR's response to the query `ASSOCIATIVE`. In this initial query, the set of nodes retrieved depends only on *syntactic* information contained in documents of the collection (viz., what words

occurred in what documents). Notice that this retrieval contains both "serendipitous" (i.e., reasonable, appropriate) elements, such as the keywords LEARNING and ASSOCIATIVE, and the author J.A. ANDERSON and inappropriate components (e.g. the keywords SYSTEM and VERSUS and the author M. MARON). Figure 10.B shows the response of AIR to the same query after several learning trials.[10] Not only have the inappropriate responses been culled, but the retrieval has also been extended to new nodes which were neighbors of appropriate parts of the initial retrieval set.

Once a number of users have refined the sub-symbolic structure corresponding to this symbol, ASSOCIATIVE could now be used as an atomic element of a symbolic system, or as an access point into the much richer sub-symbolic network.

To see how I imagine such a symbol interacting with a conventional AI expert system, imagine that ASSOCIATIVE was replaced by a nice medical term like FEVER. Using such a symbol, a medical expert system could proceed as usual, except that instead of having FEVER resolve only into a simple character string, it would correspond to a node in an associative net as well. When the expert system got stuck, the hybrid system could rely on connectionist processes (like spreading activation search) to help provide new options. Instead of being hollow, the symbol FEVER will have a rich set of connections to other symbols.

7 Conclusion

This paper has argued that connectionist representations offer a fundamentally different approach to the problems of conceptual information retrieval. By relying on the combined evidence of many weak, syntactic clues and providing a learning mechanism that can codify the most salient of these, such systems promise to provide "serendipitous" retrievals that are much better than random. We have also argued that this low-level *sub-symbolic* representation should be viewed as a complement to, rather than a replacement for, more *symbolic* representations proposed by others. Finally, we hope to have shown what a rich, important area the Law, particularly conceptual information retrieval, is for future research in AI.

Acknowledgements

I am grateful to Charlie Saxon for introducing me to the Law as a problem domain. I also thank Dan Rose for reviewing a draft of this paper and providing many useful comments.

References

[1] L.E. Allen. Language, law and logic: plain legal drafting for the modern age. In B Niblett, editor, *Computer science and law*, Cambridge University Press, Cambridge, England, 1980. normalization.

[2] J.A. Anderson and G.E. Hinton. Models of information processing in the brain. In G.E. Hinton and J.A Anderson, editors, *Parallel models of associative memory*, Lawrence Erlbaum Assoc., Hillsdale, NJ, 1984. distributed representation; activity patterns.

[3] R.K. Belew. *Adaptive information retrieval: machine learning in associative networks.* PhD thesis, Univ. Michigan, CS Department, Ann Arbor, MI, 1986.

[10] Again, the reader is referred elsewhere for the details of AIR's learning mechanism. [3]

[4] D.C. Blair and M.E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 1985.

[5] R.J. Brachman. What isa is and isn't: an analysis of taxonomic links in semantic nets. *Computer*, 1983.

[6] B.G. Buchanan and T.E. Headrick. Some speculation about artificial intelligence and legal reasoning. *Stanford Law Review*, 1970.

[7] Benjamin Nathan Cardozo. *The nature of the judicial process.* Yale university press, New Haven, CN, 1921.

[8] Constantino Ciampi, Deirdre Exell Pirro, Elio Fameli, and Giuseppe Trivisonno. Thes/bid: an expert system for constructing a computer-based thesaurus for legal informatics and computer law. In Charles Walter, editor, *Computing Power and Legal Reasoning*, pages 375–412, West Publishing Co., 1985.

[9] L.D. Erman, F. Hayes-Roth, V.R. Lesser, and D.R. Reddy. The hearsay-ii speech understanding system. *Computing Surveys*, 1980. focus of control; multiple levels of abstrac.

[10] S.E. Fahlman. *NETL: a system for representing and using real-world knowledge.* MIT Press, Cambridge, MA, 1979. virtual copy; marker passing; primitive link-types; parallel network; instance v. role.

[11] Grayfred Grey. Law & technology conference expert system workshop. In Charles Walter, editor, *Computing Power and Legal Reasoning*, pages 621–626, West Publishing Co., 1985.

[12] C.D. Hafner. *An information retrieval system based on a computer model of legal knowledge.* PhD thesis, University of Michigan, 1978. LIRS; negotiable instruments; semantic nets.

[13] Michael Heather. Demand-driven model for a half-intelligent system. In Charles Walter, editor, *Computing Power and Legal Reasoning*, pages 69–103, West Publishing Co., 1985.

[14] G.E. Hinton and T.J. Sejnowski. Learning and relearning in boltzman machines. In J.L. McClelland and D.E. Rumelhart, editors, *Parallel distributed processing: Explorations in the microstructure of cognition*, chapter 7, Bradford Books, Cambridge, MA, 1986.

[15] G.E. Hinton, T.J. Sejnowski, and D.H. Ackley. *Boltzman machines: Constrait satisfaction networks that learn.* Technical Report CMU-CS-84-119, Dept. Computer Science, Carnegie-Mellon University, 1984. annealing.

[16] W. James. *The principles of psychology.* Dover, New York, 1890.

[17] W. James. *Psychology (Briefer course).* Collier Books, New York, 1893.

[18] Robert Krovetz. The use of knowledge representation formalisms in the modeling of legal concepts. In Charles Walter, editor, *Computing Power and Legal Reasoning*, pages 275–317, West Publishing Co., 1985.

[19] L.T. McCarty. The taxman project: towards a cognitive theory of legal argument. In B Niblett, editor, *Computer science and law*, Cambridge University Press, Cambridge, England, 1980. prototype deformation.

[20] J.L. McClelland, D.E. Rumelhart, and PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume II: Psychological and Biological.* Bradford Books, Cambridge, MA, 1986.

[21] W.S. McCullough and W.H. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 1943.

[22] J.A. Meldman. *Decision support systems for legal research and analysis.* Technical Report 1039-79, Sloan School, MIT, Cambridge, MA, 1977. legal reasoning.

[23] M.L. Minsky and S. Papert. *Perceptrons: An introduction to computational geometry.* MIT Press, Cambridge, MA, 1969. perceptron.

[24] M.C. Mozer. *Inductive information retrieval using parallel distributed computation.* Technical Report, Inst. Cognitive Science, UCSD, La Jolla, CA, 1984.

[25] Allen Newell. Physical symbol systems. *Cognitive Science*, 4:135–183, 1980.

[26] F. Rosenblatt. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms.* Spartan Books, Washington, D.C., 1962. perceptron.

[27] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning internal representations by error propagation.* Technical Report ICS-8506, Institute for Cognitive Science, UCSD, La Jolla, CA, 1985. feed-forward networks; semi-linear propagation; feedback propagation.

[28] D.E. Rumelhart, J.L. McClelland, and PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I: Foundations.* Bradford Books, Cambridge, MA, 1986.

[29] G. Salton and M.J. McGill. *Introduction to information retrieval.* McGraw-Hill, Inc., New York, NY, 1983.

[30] J.A. Sprowl. Lexis v. westlaw: computer-assisted legal research comes of age. *Illinois Bar Journal*, 1979. LEXIS; WESTLAW.

[31] R.K. Stampner. Legol: modeling legal rules by computer. In B Niblett, editor, *Computer science and law*, Cambridge University Press, Cambridge, England, 1980. structure; data normalization; action; definition.

[32] C. Tapper. Citations as a tool for serching the law by computer. In B Niblett, editor, *Computer science and law*, Cambridge University Press, Cambridge, England, 1980.

[33] Anne v.d.L. Gardner. *An artificial intelligence approach to legal reasoning.* PhD thesis, CS Dept., Stanford University, Palo Alto, CA, 1984.

[34] Anne v.d.L. Gardner. Overview of an artificial intelligence approach to legal reasoning. In Charles Walter, editor, *Computing Power and Legal Reasoning*, pages 247–274, West Publishing Co., 1985.

[35] Terry Winograd. *Language as a cognitive process - Syntax.* Volume 1, Addison-Wesley, Reading, MA, 1983.