AN ADAPTIVE LEARNING NETWORK FOR INFORMATION RETRIEVAL IN A LITIGATION SUPPORT ENVIRONMENT

T.D. Gedeon and **V. Mital**
Brunel University
Department of Computer Science
Uxbridge, Middx. UB8 3PU

Abstract:

Complex litigation can involve thousands of documents that bear greater or lesser relevance to the issues in contention. Studies show that full text retrieval in which documents are indexed by all non-trivial words, perhaps with some statistically determined weighting and probabilistic matching, gives extremely poor *recall*, although users imagine otherwise. This is dangerous in a litigation situation. Binary subject indexing is sometimes added to full text retrieval but there is a trade-off between generality of subject labels and how likely the user is to remember a particular label. The problems arise from the facts (a) that legal concepts are open-textured and cannot be readily classified, and (b) a large number of concepts are expressed using a small number of technical terms. It is felt that a symbolic representation of legal concepts is inadequate in the information retrieval context and that subsymbolic features must be discovered, rather than assigned. Consequently, much attention is being focused on neural networks.

We use a network in which documents are linked to significant words found in them, the weight of links being determined through an automatic text analysis based on a normalised word frequency measure. There is no intermediate layer, allowing us to construct the initial state of the network rapidly, and to readily accommodate new documents as they arrive over a period of time. This is a peculiar requirement of the litigation support application environment. Other researchers in legal information retrieval have suggested that *symbolic* links reflecting properties such as explicit references/refutations between documents be inserted into such a network. Our research eschews direct manual intervention in the form of symbolic links for both practical and theoretical considerations. Instead, particular emphasis is paid to the method in which the user can provide feedback based on the value s/he attaches to the documents retrieved in response to a query. Here, we deal with a peculiar problem: generally, the user can say that a document in a retrieved set is relevant or not; s/he can shed little light on which relevant documents were *not* retrieved in response to the particular query.

The network has been tested on abstracts of reports of legal cases, material that is known to have a concept to word ratio similar to documents authored be parties to a litigation. The advantage is that we have some authoritative guidance as to the relevance of text units to concepts. Results show improvements on recall figures obtained by vector space retrieval. In the latter, the totality of m distinct terms forms an ordered set for which the presence or absence of terms in a particular document characteristics that document as a m-dimensional vector. As vector space retrieval too relies on weights derived from a word frequency analysis of text, the results suggest that the more complex associative nature of neural networks is discovering meaning.

Introduction:

It has long been believed that some of the problems with Full Text Retrieval (FTR) systems lie not with some inherent flaw in using queries consisting of textual expressions, but the way in which the queries are (1) formed by the user, and (2) matched with documents. A term is either present in a query or is not. Similarly, either a document matches the query or it does not.

There is sometimes provision for the retrieval of partial or ranked matches, but this can be criticised because all documents with matches for, say, three out of five query terms are lumped together, irrespective of the relative significance of the terms. Therefore, it has been advocated that the words and lexical items by which documents are indexed be weighted (Bing, 1989). It may be noted that this differs from the ordinary practice in FTR, indexing documents by means of all non-noise (non-function) words, giving equal weight to all. The user too could specify the relative importance or weight to be attached to each element of the query. Theoretically, this would allow an unlimited number of documents to be retrieved and yet not overwhelm the user because the capability of browsing through the retrieved set in the order of ranking would be present. However, there have to be procedures in the system which would cause the retrieved documents to be ranked in much the order of importance which the user would attach to them. In other words, the ranking has to have some semantic significance.

It has been suggested that statistical measures, such as relative word frequencies, have some relationship to the semantic significance. This would give us the ability to rank indexing words in order of importance. However, words do interact and there has to be some means of measuring the overall semantic significance of a collection of differently weighted words. For this, it has been proposed that both the request and the words indexing a document should be treated as vectors (Salton, 1971). A trigonometric property of pairs of vectors, namely, the cosine of the angle between the query vector and a document vector, is used to judge the similarity between a query and a document. The smaller the cosine, the greater the similarity.

Much practical work has been carried out along this line (Blair, 1990), albeit not specifically in a legal domain. Experiments have not been too encouraging. The performance of vector-based systems does not appear to exceed that achieved by manually adding subject indices and pre-ranking them (Bing, 1989). Rose and Belew (1989) have countered this by pointing out that in traditional vector-based systems there is only one link between a word and a document, and the weight of this link alone is measured. In a connectionist device - a neural network - there can be many paths between a document and a word: some direct, some through other documents or words. It is proposed that this extra connectivity can discover and capture the semantic significance at a sub-symbolic level.

Working from behaviour:

Most concepts in the real world - particularly in law, when human affairs are being reasoned about - are such that necessary and sufficient conditions cannot be stated for the application or occurrence of concepts. When it comes to categorising a set of facts as being an occurrence of one concept rather than another, this absence of necessary and sufficient conditions is a handicap. Systems based on symbolic processing attempt to overcome this handicap in a number of ways, e.g. providing numerical or quasi-numerical weighting and scoring to resolve conflict between conflicting hypotheses, or reasoning within a restricted domain and letting the human user deal with the uncertainties at the boundary.

A rather different line of attack is taken in the connectionist view. Complex computations are done, not by a set of instructions operating on a representation of entire problems, but by massively parallel (actually or notionally) computations in interconnected networks of units or neurons.

Neural networks are potentially usable where necessary and sufficient conditions for the application of concepts are absent. What is needed is reliable and representative data in the form of input-output sets - i.e. cases showing the occurrence of certain situational features or events as well as the categorised concepts which exist by virtue of the situational features. From such data a neural network system can make generalisations and potentially come to classify cases which are not part of the training set. Clearly, there is no explicit symbolic representation of a concept. What we have is a distributed computational process which reproduces the behaviour of a human with respect to the concept.

Querying a network:

Once a network of units has been constructed, a query can be presented to the system by activating those units which correspond to features of the current situation. The act of specifying some of the units activates, to some degree, those connected to it. The latter in turn spread the activation further. The links between the units are weighted in accordance with some (possibly pre-ordained) criteria. The degree of the effect that the activation of one unit has on any adjoining units is dependent on the weight assigned to the connecting link(s). The effect of the incoming activation is aggregated in each unit in accordance with built-in rules.

In a sense, it may be said that the activation of one unit is evidence for the activation of connected units, the weight of the links having correspondence to the associative or suggestive value of the evidence. Whatever the mechanism of activation, the eventual stable state of the network can be said to represent an approximation of some concept associated with the co-occurrence of the units which were initially activated through the specification of the query. Those units which are active to at least some desired level are treated as being part of the answer to the query. Alternatively, all those units which at any time reach an activation threshold may cumulatively be added to the answer.

Interactive networks:

Belew (1987) has done some of the pioneering work in the use of neural networks for information retrieval and has settled for an interactive network of the type illustrated in Figure 1 overleaf. This example appears to have two layers of neurons, although the situation is slightly more complex. In feed-forward networks, the boundaries of each layer are easily distinguishable - some have input signals from the outside world, some send results out, while others are for internal processing only. There is no such clear-cut distinction between layers in an interactive network because the units are not so well segregated. It is possible to (notionally) divide a particular network into layers in a number of ways.

A unit in the interactive network may receive inputs from the outside world, receive signals from other units and also emit results to the outside world. The flow of information is not unidirectional. Instead, it goes back and forth between connected units (resonates). At each cycle of computation, the state of the network is changed by units receiving incoming signals, processing them according to some built in rules, and sending forth signals to other units. The network is said to have stabilised when the magnitude of activation of all units does not change from one cycle to the next. Alternatively, it may change, but only recurrently in that the same values are eventually repeated. When the network has stabilised, outputs are treated as solutions/answers to the query presented by the inputs. The thesis that the network contains knowledge as to the relationships between queries and their answers applies here just as much as for feed-forward networks. The collective result of the network - shown as signals on the outputs - may be considered to be the system's purported answer or solution to the query or problem posed by the signals sent from the outside world via the inputs.

Figure 1 Basic interactive network

Augmenting an interactive network:

In Belew's work, the network is set up so that each neuron represents either a word or a document (Belew, 1987). Links are weighted in correspondence to a certain word frequency measure derived from automatic analysis of the text. This means that the network is usable, to some extent, immediately on creation - rather than only after extensive training, as would be the case for feed-forward networks. Training and learning is, of course, likely to change these weights. This is where, in our view, considerable problems are likely to arise, at least in the context of the application that we are concerned with: supporting litigation by allowing a large collection of documents to be incrementally enlarged without deterioration of performance.

Learning during use means that the weights of links, which are originally set by relative word frequency measures, are likely to have changed by the time new documents come to be added to the information base. Adding new documents will affect the relative word frequencies of even the existing documents. It is therefore not clear whether the whole network will have to be reinitialised with weights corresponding to new word frequencies, or, whether only the new documents will be connected by such links, while the links of the previously existing documents will keep their status quo in spite of the changed state of play. Either alternative is unsatisfactory. As shall be seen next, we have taken this problem into account when designing our network architecture.

System Overview:

The structure of the network that we are using is shown in Figure 2. All words, other than so-called noise words or position-holders, such as prepositions and articles are made units in the network. All documents are also units. Every word has links to each document to which it is 'significant'. Significance is judged by a standard technique in automatic analysis of text for information retrieval: forming a numerical measure of the frequency with which a word occurs in a document, normalised by (a) the frequency of the word's occurrence in the document collection as a whole, and (b) the size of the particular document. Those words which have either too low or too high a word frequency measure are discarded as either being too parochial or too ubiquitous to be of much use in discriminatory retrieval.

Figure 2 Network designed for information retrieval

Each word is linked to every document to which it is significant through a bidirectional link created with a weight corresponding to the frequency measure, these links are called textual-associative (T-A) links. Additionally every word is connected to every document by latent (L) links which are all initially given a weight of zero. These are also bidirectional, in the sense that weights are the same in both directions.

The weight of T-A links is, as mentioned, determined by the results of the automatic text analysis, using the number of times that the word occurs in the document as well as the number of times it occurs in the entire document collection as a whole. The weight is a increased in proportion to the former measure, the hypothesis being that the more often a word is found in a document, the more representative it is of the document. The latter measure causes a decrease in the weight since a common word is unable to adequately distinguish between documents. To avoid the danger that these measures would be distorted by the variation in the size of documents: the fact that a ten page document contains a word five times does not mean that the word is more representative of this document than of a one page document in which it occurs once.

Queries are restricted to specifying words which act as concept micro-features. Alternatively, a query may specify at most one document. A query may not contain both words and documents because this is of dubious significance in the context of conceptual information retrieval. This is because a document may contain, or be relevant to, a number of concepts. If, say, three words and one document were to be selected, does it mean that the concept signified by the co-occurrence of the three words, together with all the concepts referred to in the document, are part of the query? The reason we allow one document alone to be in queries is that then the documents contained in the answer can be thought to be (in some sense) similar to the one in the query in a multidimensional concept space.

The set of retrieved documents itself contains rankings and gradations according to the strength of activation. Consequently, relevance is not an all-or-nothing quality, but is relative.

Operation of the network:

As described above, the network consists of two types of bidirectional links. The T-A links use a word frequency measure, while the L links weights are derived from training during actual use by a user. T-A links are always positive as, obviously, there is no meaningful measure of how many times a word does not occur in a document. During learning, L links are modified based on user responses to documents supplied as answers to queries. The user may respond as follows: relevant, marginal, and irrelevant. The L links to a document from the words forming the query are increased in weight for ‘relevant’ user responses, and reduced for user assertion that the document was ‘irrelevant’.

Exclusionary relation between concept and document:

The network can learn to assign negative relevance of words to documents (in the context of concepts) because the joint effect of the T-A and L links may become negative. This is significant, a word may occur in a document following an exclusionary declaration about a concept. An example would be a statement in the text that ‘we will not consider negligence’. The word negligence does occur, but only so that the concept(s) related to it are immediately excluded. Therefore, this document should not be retrieved when the relevant concept is specified in the query. We allow the network to learn about this. We considered including a possible fourth user response during training to explicitly indicate that rather than being ‘irrelevant’ the document has a exclusionary connection to the query concept. However, we believe that this would be perilously close to adopting a symbolic computation approach, with all the attendant problems of exhaustive assignment of symbolic connections (Rose & Belew, 1989). We would prefer for this information to be discovered at the subsymbolic level. If a word and document pair are sufficiently often noted by the user as not relevant, then eventually the network can discover that the concept is excluded even though it is mentioned.

Vector retrieval is a byproduct:

The first cycle of processing in the network is special, in that the L links are not used. This allows us a baseline of a non-banded or non-discrete valued (continuous) version of vector retrieval as the first set of outputs on the document units. Vector retrieval is a well known, albeit simplistic, method of retrieval based on statistical affinities between words and documents (Salton, 1971; Bing, 1989). This is useful as an interim result, and is very fast - important in a practical system.

The document set produced as an answer to the query can be built up in a number of ways:

(a) The above mentioned interim set of vector retrieval.
(b) The document units active when the net has settled, in the sense that the change in activations has a low amplitude.
(c) The document units activated in a short repetitive cycle.
(d) All those above a threshold activation at any time during processing.

Activations of documents and words are limited to the range between 0 and 1. The ceiling could obviously be any arbitrary value, the floor value of 0 is significant in that an activation of 0 implies that that word or document unit is not relevant (is excluded from) the query, and is therefore effectively isolated from influencing the rest of the network. Subsequent incoming signals may of course increase its activation and reinstate it.

Scaling:

Activations coming into units are adaptively scaled in a global as well as a local fashion. This is to guarantee that the network will settle within a reasonably short period, or oscillate in a restrained fashion only.

Global scaling means that the total energy in the network is measured and every unit’s activation is reduced pro rata to bring the energy to a pre-specified level. Global scaling of activations indirectly affects approximately how many documents are to form part of the answer to the query, as well as how many words other than the ones input by way of query are to be considered internally by the network. In other words, scaling affects the scope of the spread of activation. As a simplified example, values of 3 and 10 are chosen for scaling factors of document units and word units respectively. Then the total activation energy is scaled down to accord with these factors. The total could be distributed so that 3 documents and 10 words all have activations of 1, with all the rest 0. This is unlikely, we find that a more wide spread distribution is usual. However, this is not so wide as to contain a lot more than 3 documents and 10 words significantly active. In other words, the spread has been limited to a certain extent. Of course, which documents/words are active will be decided interactively by the network.

Local scaling is a further scaling. It looks at the highest activation in the network. That is reduced to bring it down to a the permitted maximum. However, the other units are reduced in a ‘socialist’ fashion in that the percentage reduction is proportional to the local activation energy.

Learning:

As mentioned, the network in its initial state with all L links at 0 still produces useful results. The first cycle produces a vector-retrieval. After the network has settled into a steady state, the result signifies the output of the network’s approximation of a concept formed by the query words. This concept may not have been exactly what the user had in mind, as would later be evinced by his responding that some of the output documents are ‘irrelevant’. This is where learning comes in.

We can incrementally adjust the network by utilising the user’s explicit behaviour. The error or correctness signified by the user’s responses is readily interpretable into blaming or rewarding particular parts of the network because there are direct word to document relationships. The network is changed by modifying the L links by attaching extra weight or reducing the weight of the connections between the query words and the impugned or approbated documents.

Implementation status:

We have implemented the first version of the network on an Apple Macintosh computer using Lightspeed Pascal. We may note that our intention is to integrate the neural network to a hypertext system (Southam, Mital and Thomas, 1991) which also runs on a Mac.

The initial tests have been done using only 19 documents containing approximately 21,000 words. From this, we obtained 3,700 non-noise (non-function) words. From the culling which followed the analysis of normalised relative word frequency distribution we were left with 569 significant words, each of which is represented by a unit in the network, as is each of the 19 documents.

The setting up of the initial network, from calculating word frequencies to inserting and establishing the weights of the T-A links, takes under one minute.

Each cycle of spread of activation takes from 6 to 8 seconds. Therefore, the vector-retrieval answer is obtained almost immediately. Further, the network usually reaches a steady state (quiescence or relatively minor oscillations) in 5 to 12 cycles.

The initial tests have been done with abstracts of reports of cases relating to the law of torts. So far, the authors themselves have been assessing the quality of retrieval and checking it against the standard classifications in legal treatises. In comparison to vector retrieval, the settled network produces more of those documents which are listed in standard works as being relevant to the query. We propose to carry out more exhaustive tests with independent users.

Discussion:

The question which must be answered is why maintain the two sets of links L and T-A?

The major reason is to allow the inclusion of new documents at will into the network without degrading the performance. For this we want to retain the L link information which the network has learnt. Certainly the addition of a number of documents will change the word frequency measure of some connections between words and a particular document to the extent that, if T-A links were all we had, the document would not be presented to the user although it might be on the basis of the same query using the original document set. Nevertheless, the significance of the user selecting a document as either ‘relevant’ or ‘irrelevant’ to the concept he had in mind when making a query in the smaller set is not diminished and should be carried forward. If a document was said by the user to be relevant to a concept derived from the words making up a query it should remain so even though the additions change the word frequencies. Even more importantly - because of the much longer time taken to learn this - if a document is learnt to be irrelevant to a concept, it should not become falsely relevant in the light of new documents.

However, the relative importances of the L and T-A links when the document set changes is a matter which we are still investigating. We are currently looking at the susceptibility of the network in the face of sizable changes to the information base.

Combining symbolic and sub-symbolic links:

Another question that arises is the the efficacy or sufficiency of the connections between the various units. If one relies on statistical analysis alone, it may be that two words which any informed person would say are closely related have no direct linkage,or that one case is known to support another explicitly, yet activation is not propagated alone a heavily weighted link between them. Researchers have addressed this point.

The FLEXICON (Fast Legal Expert Information Consultant) system (Gelbart & Smith, 1990) and SCALIR (Symbolic and Connectionist Approach to Information Retrieval) (Rose &

Belew, 1989) add ‘logical’ links to a network created along the above mentioned connectionist lines. These logical links are those which would be suggested by a domain analysis similar to that carried out for any knowledge system premised on a symbolic approach.

In SCALIR, the connectionist links are termed C-links, the logical (or symbolic) links S-links. The former are treated as described for the AIR system. Presently, the S-links are primarily meant to denote the relationships which exist between key numbers in the widely used WESTLAW system from West Publishing. This company has long been publishing a large proportion of the case law reporters in United States jurisdictions and the organisation of concepts implied by the categorisation of the key numbers has achieved the status, almost, of being part of the law. Consequently, all term units in the network which correspond to West key numbers are joined by S-links. Some units correspond to statute sections: S-links are added to acknowledge the structure of the statute. Other S-links represent information such as that one case has been overturned by another. The S-links are distinguished from one another, not by being differently weighted, but by having labels: ‘includes’, ‘refers to’, etc. A semantic network is thus formed.

Activity traverses C-links in the normal connectionist manner: all inputs into a node are multiplied by the weights of the respective links they are passed along. When a S-link is encountered, activity either passes fully or does not pass at all. An interpreter allows only that type of activity which is compatible with the link to be passed on. So, once an activity has been passed along a link of type ‘supported by’, it will subsequently propagate via such links only.

FLEXICON also superimposes links suggested by a symbolic domain analysis on to those derived from statistical analysis of text, but in a somewhat different way. It appears that rather than having logical links contribute to activity propagation at every stage of computation, their role is confined to the act of refining a query prior to its being presented to the network, i.e., before the units corresponding to terms in the query are activated. The developers are clearly more confident than developers of SCALIR that word frequency and co-occurrence measures are adequately representative of the semantic significance of documents. Indeed, FLEXICON proposes to present a statistically determined case profile to the user as a summary or abstract of the case - could it be no more than a vector?

So far, the developers of the above systems have not reported on whether the use of symbolic links has been successful. All that can be said is that the approach is not entirely dissimilar to domain modelling by a knowledge engineer in the usual fashion. This is labour intensive and sacrifices the enormous potential advantages of automatic text analysis for a hybridisation with uncertain benefits.

Conclusion:

The connectionist approach to information has great practical utility. It can be used in conjunction with conventional systems such as fulltext databases, key word retrievers, and hypertext based information systems. Indeed our system has been integrated with a hypertext based legal document assembly system (Gedeon and Mital, 1991).

We believe that the connectionist approach has considerable advantages over the vector retrieval approach as well as several types of probabilistic retrieval models. Further, whether or not one believes that automatic indexing is to be preferred to conceptual indexing, there already are commercial situations where the former is the only one feasible economically.

References:

Belew, R.K. (1987). A connectionist approach to conceptual information retrieval, Proc. First International Conference on Artificial Intelligence and Law, ACM, 116-125.

Bing, J. (1989). The law of the books and the files: possibilities and problems of legal information retrieval, in: G.P.V. Vandenberghe, Ed., Advanced Topics of Law and Information Technology, Hillsdale, NJ, Lawrence Erlbaum, pp.151-182.

Blair, D.C. (1990). Language And Representation In Information Retrieval, Amsterdam, Elsevier.

Gedeon, T.D. & Mital, V. (1991). Integrating connectionist information retrieval with hypertext, Technical Report CSTR-91-8, Brunel University.

Gelbart, D. & Smith, J.C. (1990). Knowledge-based information retrieval systems. Computers & Law, October, 1990, pp. 23-26.

Rose, D.E. & Belew, R.K. (1989) Legal information retrieval: a hybrid approach, Proc. First International Conference on Artificial Intelligence and Law, ACM, 138.

Salton, G. (1971). The SMART Retrieval System - Experiment in Automatic Document Processing, Englewood Cliffs, Prentice-Hall.

Southam, M.; Mital, V. & Thomas, P. (1991). HyperNotary: Legal document assembly using hypertext, Proc. 13th BCS Research Colloquium on Information Retrieval, Lancaster.