Evaluation of Connectionist Information Retrieval in a Legal Document Collection

R.A. Bustos and T.D. Gedeon
School of Computer Science and Engineering
The University of New South Wales
Sydney, NSW 2052
Australia
E-mail: {robertb, tom} @cse.unsw.edu.au

Abstract

This paper describes the evaluation of a spreading activation model for information retrieval. Different configurations of this model were applied to the problem of discovering cross references for a hypertext version of a legal research textbook. The `cirs` (Connectionist Information Retrieval System) is described. Six different weighting schemes for links between neurons were evaluated, using human indexed cross references.

1. Methodology

The aim of this experiment is to explore the effectiveness of spreading activation models of information retrieval when applied to a practical problem. A general model with a graphical user interface has been implemented in the `cirs` system (Connectionist Information Retrieval System, Bustos, 1995). `Cirs` allows users to interactively modify parameters of the system including indexing word selection criteria, network parameters, network models and query generation.

Using cirs, six different weighting schemes for links between neurons were evaluated, using a legal text document collection with human indexed cross references. The evaluation is carried out using the user oriented recall and precision measurement method [Savoy 1994]. The network models are described in full below. Recall and precision values for each model were recorded for 12 queries, selected to display varying levels of performance, from the 73 available. These results are in section 3 below.

The best precision / recall values over the 100 iterations of the network were selected on the basis that such a test would simulate the behaviour of a user operating the retrieval system, having the opportunity to decide whether or not to continue to search for more references at any point. The aim of this experiment was to compare the user oriented performance of the different models. Clearly, users would not have the prior knowledge of ideal retrieval results with which to measure performance and to help decide when to halt the search, however it is suggested users would approximate this by considering the number of relevant items, and their relative positions in the retrieved list. Parameter tuning heuristics, discovered by theoretical analysis and practical experience are in section 4 below.

2. Network Models

This section describes the six different network models which were implemented in `cirs`. Each of the models (including Smart Boolean) are implemented by defining the nature of the connections between word and document units.

2.1. Smart Boolean

Boolean full text retrieval systems rely on users' ability to choose appropriate words for queries, a task that they (usually) approach with knowledge of approximately how common particular words are in the domain of the search. Using these intuitions, users select words that they think will occur sufficiently often to retrieve relevant documents, yet not be so widespread as to cause retrieval of unwanted documents.

The present evaluation task requires that queries be constructed automatically from tutorial questions written to test understanding of the textbook that forms this document collection, without the intervention of a human user. One strategy would be to use every word from each tutorial document and create links of constant weight between each word and the document that it occurs in, and consider retrieved any document neuron that received any activation. This process most accurately models the boolean retrieval method but fails to take into account the users' ability (and interest) in being more selective when choosing words for queries.

A weighting strategy as described above would demonstrate unrealistically poor performance, so modifications were implemented to make the model a more realistic simulation of the combination of users domain knowledge than the simple boolean retrieval method. Links from words to documents are created with constant weights, but an activation threshold is applied, with only document units of greater activation than the threshold considered to be retrieved. In this way, document with more of the query words will be retrieved first, thus performing a crude document relevance ranking.

Significantly, this retrieval method is the only one that does not have connections from documents to words, since boolean retrieval systems do not retrieve on the basis of shared words in any way. Inhibition between units is another feature of the other network models that is not included, as the traditional concordance implementation of boolean information retrieval does not include any facility such as this.

2.2. Binary Networks

The binary network model is a simple enhancement of the boolean network, by adding an extra link of constant weight in the opposite direction, from document neuron to word neuron, whenever a word occurs in a document. There are now two connections, which can also be described as a single bi-directional connection between the word and document units.

The binary network is a more powerful way of representing the relations between words and documents than the smart boolean method, since activation flowing from document to word neurons over the additional links can activate further new words, thus capturing context and eventually activating other document neurons. The use of the activation threshold maintains the retrieval ranking mechanism, and competition within the word and document unit pools limits the activation, with the effect of focussing the retrieval and reducing the effects of low frequency noise words.

2.3. Word Frequency Networks

Binary connections are, nevertheless, a relatively coarse way of representing relations between words and documents. The symmetric word frequency weighting method connects words and documents with bi-directional links, that are proportional to the word frequency of the indexed word in the given document. Weights are normalised by the total frequency of the word, thus reducing the relative effect of words which are highly frequent in the overall collection.

The frequency network is an improvement over the binary net, since the proportionality of the connection weights to document neurons allows for greater detail of this relation to be represented and used for retrieval. As the links in this method are bi directional, properties of the binary network method such as context capturing and use of inhibitory pools remain.

An asymmetric version of this interconnection method is also included. Weights from documents to words are calculated as above, thus the semantic meaning of documents is represented based on word frequencies in the document. However, the links from words to documents are proportional to the number of occurrences of the word in the document, normalised by the number of occurrences in the whole document collection. First described in [Rose and Belew, 1991], this interconnection method more accurately models the asymmetric nature of the relation between words and documents: a particular word may be of disproportionate value in describing the content of a particular document, however it does not follow that the word should retrieve that document any more readily than any other, as there may be several documents each containing this word with varying combinations of other terms that could influence strength of connection. The increased importance of the word in describing the content of the document becomes important however when activation begins to flow backwards from document to terms, when it is desirable that terms be activated in proportion to their semantic value for describing document content.

2.4. IDTW Networks

These networks with connection weights proportional to the inverse document - term weights of the document. This weighting method was developed by [Salton 1983] for vector retrieval, and takes account of the document frequency of words to assign term importance according to the discriminatory power of the word. Thus, words occurring very frequently in many documents, and those occurring only occasionally in very few documents are assigned less weight. The equation for the weighting method is:

$$TW_{ij} = \sqrt{\frac{F_{ij} \times \log\left(\frac{\mathrm{N}}{\mathrm{DF_j}}\right)}{\sum_{\mathrm{k=1}}^{\mathrm{T}}\left(F_{ik} \times \log\left(\frac{N}{DF_k}\right)\right)^2}}$$

where:
$N =$ Number of documents

$T =$ Number of terms

F_{ij} = Frequency of term j in document i.

DF_j = Document frequency of term j.

As with the word frequency networks, symmetric and asymmetric networks are constructed. The symmetric network consists of weights in both directions proportional to the IDTW value, whilst the asymmetric net uses IDTW values for connections from document to word neurons, and normalised document frequency values from words to documents (as was the case for the frequency networks described above).

Calculation of IDTW values for connection weights is significantly more computationally expensive than simply using normalised word frequencies values. This weighting method better models the relation, however, because the frequency of words over the whole collection is taken into account when assigning importance to terms in documents.

3. Network Evaluation

This section describes evaluation of the different network models using the user oriented recall and precision approach described above, in the Methodology section. Results presented are the best recall / precision points for each query, for each method (Tables 1 - 60.

3.1. Recall /Precision Performance Comparison

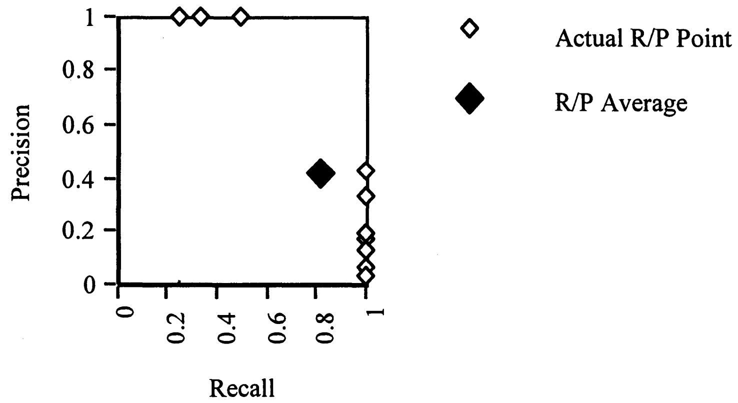
The 13 example queries displayed in these tables are a subset of the full collection of 73 queries, and were selected to display a wide variety of the properties of the collection, including varying query length, differing use of context in the query, and differing tutorial question style. Accordingly, these queries demonstrate the varying levels of performance of the models.

3.2. Smart Boolean Network

The Smart Boolean network demonstrated surprisingly good performance and was one of the most efficient networks, both for network generation time and cycling time. The activation threshold was set to 0.1. This parameter can have a dramatic effect on precision values for the retrieval method, since the combination of the activation threshold and number of cycles run effectively determines the number of items in the retrieved list.

Table 3.1. Boolean Network Results

Query	0	1	2	3	4	5	6	7	8	9	10	11	12
Recall	1.000	1.000	0.333	0.250	1	1.000	1.000	1.000	1.000	1.000	0.500	0.500	1.000
Precision	0.167	0.429	1.000	1.000	0.333	0.188	0.061	0.133	0.063	0.063	1.000	1.000	0.036
No. Ideal	1	3	3	4	1	3	2	2	2	6	2	2	7

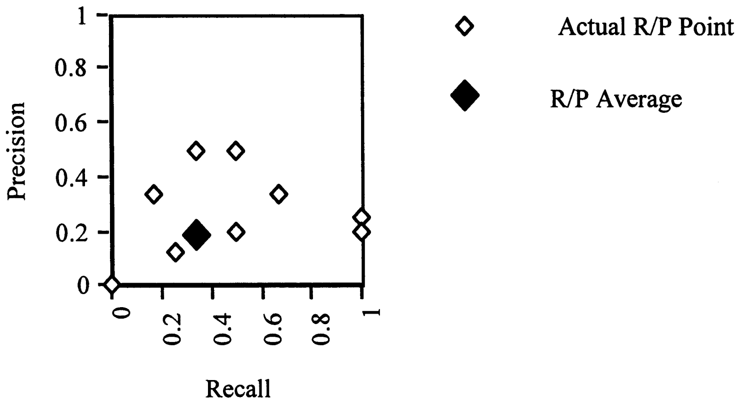


3.3. Binary Network

The binary network demonstrated the poorest performance and was the hardest to set the parameters well. Greater decay was required since the stronger interconnections between words and documents caused a very rapid increase in network total activation. Increasing decay has the effect of widening the separation between the activated units and units which are not activated, since units not receiving enough activation to overcome the decay end up being deactivated. The effects of inhibition, excitation and external input are also reduced by lowering these parameters.

Table 3.2. Binary Network Results

Query	0	1	2	3	4	5	6	7	8	9	10	11	12
Recall	0	0.667	0.333	0.250	1.0	1.000	0	0	0	0.167	0.5	0.5	0
Precision	0	0.333	0.500	0.125	0.25	0.2	0	0	0	0.33	0.5	0.2	0
No. Ideal	1	3	3	4	1	3	2	2	2	6	2	2	7



The crudeness of the associations between words and documents, when defined using only constant magnitude weights means that the first retrieval of the network is not particularly accurate, which greatly effects the final results.

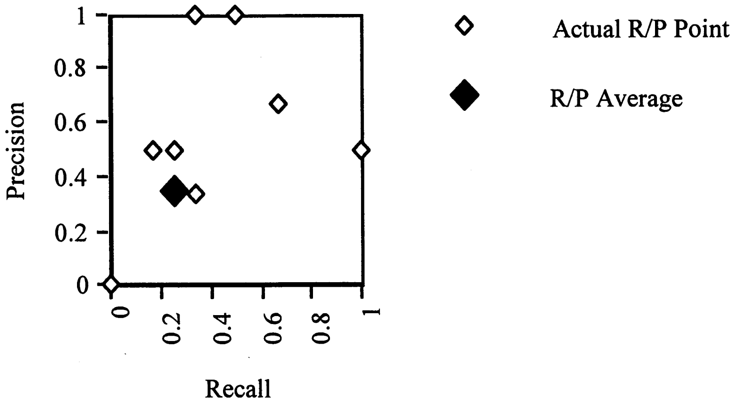
With greater levels of total activation, retrieval results also end up being more dependent on the network topology than the query, so the same units keep on getting activated since they have the largest number of connections.

3.4. Symmetric Frequency Network

In general, the symmetric networks showed significantly worse performance when compared to the asymmetric versions. The normalised connections of this model made the task of setting parameters for the model easier, without the problem of rapid activation growth. However, after a number of cycles, the activation often did reach the maximum value and began to be scaled.

Table 3.3. Symmetric Frequency Network Results

Query	0	1	2	3	4	5	6	7	8	9	10	11	12
Recall	0.000	0.667	0.333	0.250	1.000	0.333	0.000	0.000	0.000	0.167	0.500	0.000	0.000
Precision	0.000	0.667	1.000	0.500	0.500	0.333	0.000	0.000	0.000	0.500	1.000	0.000	0.000
No. Ideal	1	3	3	4	1	3	2	2	2	6	2	2	7



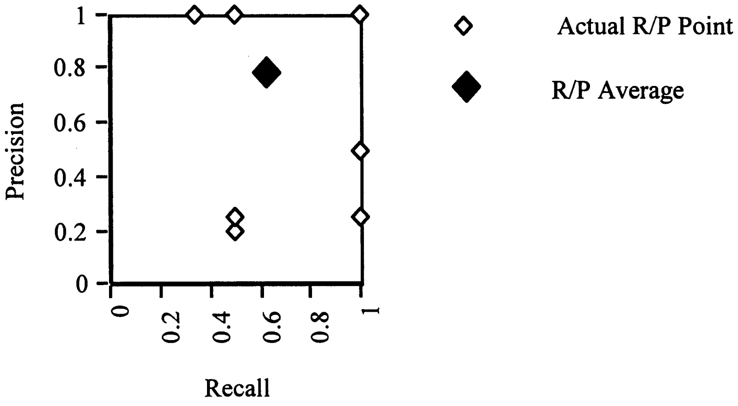
A significant feature of this network model is that the connections are not document size independent, so larger documents (with more connections) tended to be retrieved first.

3.5. Asymmetric Frequency Network

Performance improved significantly when using the asymmetric version of the frequency mode. Normalised connections again made the task of setting network parameters and easier one. It is interesting to note the performance with this model is generally at a lower level of recall than the Smart Boolean method, but with much improved precision.

Table 3.4. Asymmetric Frequency Network Results

Query	0	1	2	3	4	5	6	7	8	9	10	11	12
Recall	1.000	1.000	0.333	0.333	1.000	1.000	0.500	0.500	0.500	0.333	0.500	0.500	0.500
Precision	1.000	1.000	1.000	1.000	0.500	0.250	0.200	1.000	1.000	1.000	1.000	1.000	0.250
No. Ideal	1	3	3	4	1	3	2	2	2	6	2	2	7

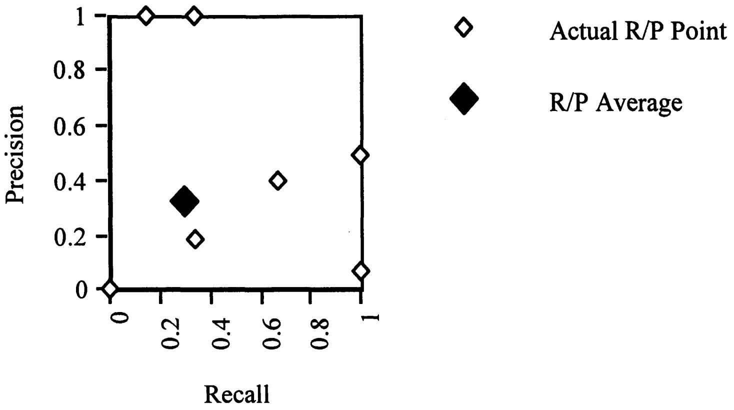


3.6. Symmetric IDTW Network

As with the other symmetric model, the Symmetric IDTW network shows disappointing performance, often failing to find any of the references for given tutorial. Greater effort was thus applied to finding the appropriate parameter settings for the asymmetric version of this model.

Table 3.5. Symmetric IDTW Network Results

Query	0	1	2	3	4	5	6	7	8	9	10	11	12
Recall	1.000	1.000	0.333	0.25	0.000	0.000	0.000	0	0	0.333	0.333	0	0.143
Precision	0.500	0.067	1.000	0.125	0.000	0.000	0.000	0	0	1.000	0.182	0	1.000
No. Ideal	1	3	3	4	1	3	2	2	2	6	2	2	7

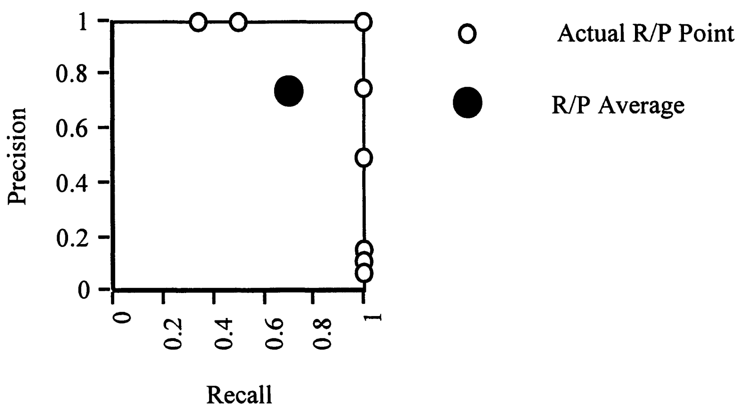


3.7. Asymmetric IDTW Network

This result is a significant improvement on the Smart Boolean network results, and is the best results over all the models. The same parameters that were applied to the model as were used in all the other networks; same 13 queries, activation

Table 3.6. Asymmetric IDTW Network Results

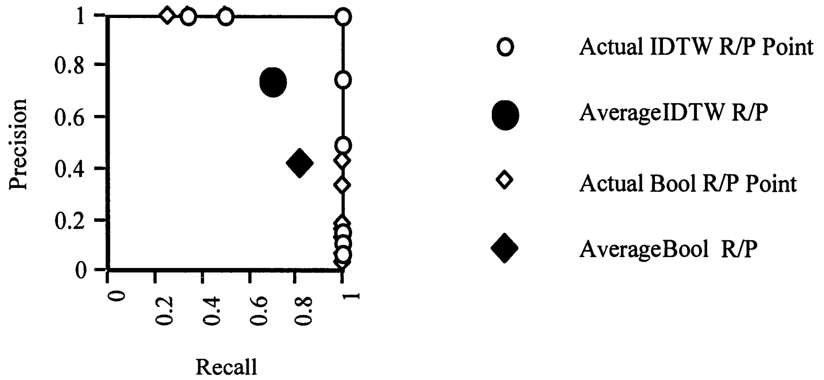
Query	0	1	2	3	4	5	6	7	8	9	10	11	12
Recall	1.000	1.000	0.333	0.500	1.000	1.000	0.500	0.500	0.500	0.500	0.500	1.000	0.429
Precision	1.000	0.750	1.000	0.500	0.500	0.150	0.250	1.000	1.000	1.000	1.000	0.286	0.250
No. Ideal	1	3	3	4	1	3	2	2	2	6	2	2	7



threshold and number of cycles, with the exception of those from the binary model, which required different settings due to the non-normalised nature of the word-document connections. Section 4 examines this network in a more qualitative light, and the 12 queries used here are tested with varying parameters for query length and sensitivity to stopword removal and stemming.

3.8. Comparison of Smart Boolean and Asymmetric IDTW over Full Collection.

The following scatter graph shows a comparison of the best retrieval results of the Smart Boolean and Asymmetric IDTW networks over the set of 13 testing queries used for the single network scatter graphs.



These results confirm the impression that the asymmetric IDTW spreading activation model appears to find a similar number of the references, however with greater precision.

4. Parameter Tuning Heuristics

This section records some observations during the parameter tuning process. The selection of network and word indexing parameters is an important, yet largely unguided process. A method of selecting indexed words was discussed in [Bustos and Gedeon, 1995], and it may be possible to use other machine learning techniques to discover correct network settings. Decisions about parameter settings were made with consideration of factors including machine resources, observed total activation levels and apparent retrieval performance (ie informal examination of results lists).

4.1. General Techniques

In general, parameters were adjusted to slow growth in total activation, since rapid growth often led to oscillation, and the negative effects of scaling to a total activation value which are described below.

The external input (*estr*) parameter can be decreased to reduce the effect of the initial query input on documents retrieved. A very high value for this parameter makes the retrieval process more closely resemble the underlying vector retrieval relations embodied in the weights, and total activation assumes a high value after just a single cycle.

The excitation parameter (*alpha*), may be increased to hasten the effect of documents activating other documents through the word units. Increasing alpha in combination with reducing the external input strength parameter emphasises the context capturing ability of the spreading activation model, however this can introduce problems of rapid activation growth and difficulty in achieving network convergence.

4.2. Decay and the Effects of Scaling to a Global Maximum Activation

It is necessary to prevent unbounded growth of total network activation, since effectively results in retrieval of all the documents in the collection, and thus very poor precision. Two approaches are built into `cirs`: a global decay term, and scaling to a constant total activation. Provision of a decay term was a strategy implemented in [McClelland and Rumelhart, 1988], while scaling of total activation was an extension examined by [Gedeon and Mital, 1991].

The inclusion of a decay term modifies the network algorithm to create a tendency for neuron activation to decrease, if the neuron does not have continuing support from neighbouring units. This has the effect of 'focussing' retrieval in the sense that low frequency connections are reduced further and activation growth (retrieval) proceed according to the most strongly connected units. Accordingly, use of a decay term removes some of the benefits of the accumulation of many smaller connections to retrieve documents, a feature of the spreading activation models.

Scaling the total network activation to a constant value allows control of the activation levels whilst maintaining the effects of low connection weight associations between words. This method is implemented in `cirs` with a single parameter for the total activation of document neurons, from which a proportional value for word units is calculated. When the total activation of a pool of unit exceeds these values, the activations are proportionally scaled down to the given total.

While this scaling has the desired effect of controlling activation, under conditions of total activation significantly greater than the scaling parameter, retrieval results become dependent on network topology only and the effect of the initially activated query words is lost. Activation is progressively spread over all the units in the pool, and in combination with a retrieval activation threshold, this has the effect of reducing the length of the retrieval list as fewer units reach the threshold.

5. Commentary and Conclusions

The aim of this work has been to implement and evaluate spreading activation models of information retrieval. The `cirs` system was designed and coded in object oriented C++ for this purpose and evaluated with the user oriented precision and recall approach.

The document collections used for this section are larger than any published study of spreading activation models, using more documents and indexing more words than any of the other reported systems. A number of weighting schemes were examined and the asymmetric approach to frequency, mentioned in [Rose and Belew, 1991] was implemented and evaluated, and extended to IDTW.

This large scale evaluation has achieved results significantly different to those expected from smaller scale implementations, since improved precision appears to be common result from use of the models. Expected improvements to absolute recall were not recorded, perhaps supporting the views of those [e.g., Blair 1990] who highlight the limitations of using word occurrence information to capture document meaning.

The specialised task of link generation for hypertext systems was examined, and problems with the large number of free parameters in the spreading activation model highlighted. Although beyond the scope of the evaluation project attempted here, the application of machine learning techniques to find appropriate parameter settings may be an area of research meriting future attention before these methods become practical enough to replace dominant retrieval technologies such as the boolean concordance.

The implementation of a network model combining connection weights proportional to word frequency in the document for document to word links, and weights proportional to word frequency normalised by total word occurrence for the whole document collection for word to document links, has not been described elsewhere.

The document collection view is shown in Exhibit 1.

References

Blair D.C. 1990 *Language and Representation in Information Retrieval*, Amsterdam, Elsevier.

Bustos, R.A. 1995 *Neural Networks in Law*, Thesis, School of Computer Science and Engineering, The University of New South Wales, Sydney.

Bustos, R.A. and Gedeon, T.D. 1995 "Learning Synonyms and Related Concepts in Document Collections," in Alspector, J, Goodman, R and Brown, TX

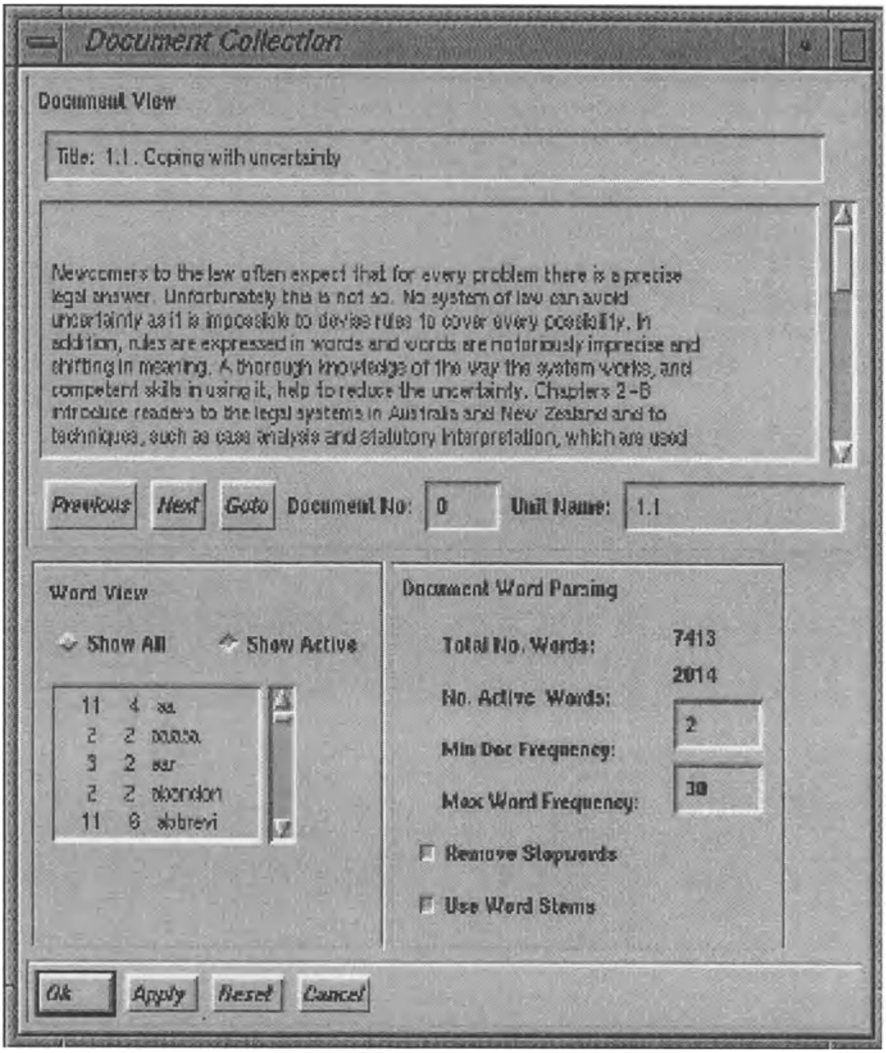


Exhibit 1. Document collection view

Applications of Neural Networks to Telecommunications 2, pp. 202-209, Lawrence Erlbaum.

Gedeon, T.D. and Mital V.J. 1991 "An Adaptive learning network for information retrieval in a litigation support application", *Proc AMSE International Conference on Neural Networks,* vol. 2, pp. 25-34, San Diego.

McClelland, J.L. and Rumelhart, D.E. 1988 *Explorations in Parallel Distributed Processing*, MIT Press, Cambridge, Massachusetts.

Rose, D.E., Belew, R.K. 1991 "A connectionist and symbolic hybrid for improving legal research", *International Journal of Man Machine Studies*, vol. 35, pp. 1-33.

Salton, G. 1983 "On the Representation of Query Term Relations by Soft Boolean Operators," *Proceedings of the Association for Computational Linguistics (ACL) European Chapter 2nd Conference*, pp. 116-122 Geneva.

Savoy, J. 1994 "Searching Information in Legal Hypertext Systems," *Artificial Intelligence and Law* , vol 2., pp. 205.

Wong, S.K.M., Cai, Y.J. and Yao, Y.Y. 1993 "An Application of neural Networks in Adaptive Information Retrieval" VOL 2, *Proceedings IEEE International Conference on Neural Networks*, San Francisco, CA , March 28-April 1.