Word-Concept Clusters in a Legal Document Collection

T.D. Gedeon [1], **R.A. Bustos** [1], **B.J. Briedis** [1], **G. Greenleaf** [2] and **A. Mowbray** [3]

[1] Department of Information Engineering
School of Computer Science Engineering
The University of New South Wales
Sydney NSW 2052, Australia

School of Law
[2] The University of New South Wales
[3] University of Technology, Sydney

http://www.cse.unsw.edu.au/~tom

Abstract. For very large document collections or high volume streams of documents such as information resources on the web, finding relevant documents is a major information filtering problem. Traditional full text retrieval methods can not locate documents which use specialised synonyms or related concepts to the formal query. This is particularly a problem in legal document collections, since lawyers use normal words with specialised meanings which vary subtly between legal sub-domains. We use a neural network approach to learn synonyms and related clusters of words defining similar concepts from a sample document set. We demonstrate that our clusters of words are qualitatively useful, in the legal domain in particular, and can thus be used for high throughput information filtering to find documents likely to contain concepts relevant to a user's information need.

1 Background

The Australasian Legal Information Institute (AustLII), was established by the University of New South Wales and the University of Technology, Sydney. Funding for 1995 was provided to Greenleaf Mowbray and Gedeon by the Department of Employment, Education and Training, and supplemented by the two Universities. Further funding has been received from the Law Foundation of NSW for 1996, and from the Australian Research Council for 1996-1998 to Gedeon Greenleaf and Mowbray. The work reported in this paper was supported by the latter grant.

The high volume use of the legal materials available via the internet on AustLII provides an invaluable research opportunity in information filtering, retrieval and index generation, particularly for neural networks which require large numbers of instances for training. AustLII the World Wide Web site (http://www.AustLII.edu.au)

came up on the web at the beginning of July 1995, and by August of 1995, AustLII was averaging 4,000 hits per work day. By the end of August 1996, the AustLII site was averaging 38,000 hits per work day.

2 Introduction

The problem domain is the provision of sophisticated access to legal information via AustLII, which allows the modelling of the complex interconnections possible between sources of information, which does not require expensive expert intervention to maintain, and is adaptive to user needs.

Hypertext meets the first two criteria, our aim is to use neural network and other AI learning techniques to discover useful connections [3] based on the document collections themselves, and to maintain and enhance the hypertext structure [4] based on observation of user interaction with the AustLII internet resource.

Users face a difficult task when formulating queries for boolean retrieval: words must be selected that will retrieve the documents wanted, but fail to retrieve unwanted documents. Blair [1] has suggested that this is an unreasonable expectation of users and that retrieval performance of boolean retrieval systems is seriously limited as a result. In situations where high recall is desired (as for most legal tasks) we can add words to the query that will have the least negative effects on precision.

3 Neural network experiment

We create a network consisting of an input and output node for each word, connected by hidden units. Training patterns are generated using each document in the collection. One input only is activated for each input vector, corresponding to a word that occurs in the document under consideration.

The corresponding output vector consists of the word frequencies of all of the words occurring in the document. A pattern is generated in this way for each word in each document. The network is trained on these patterns, and the back propagation algorithm is used to generalise an output vector of word activation terms most similar to the training examples for the given input.

The word activation values can be ranked in descending order to discover the most important related words.

3.1 Preliminary work

This paper continues a preliminary experiment using INSPEC (computer science: neural networks) abstracts [5].

The aim is to use on-line legal data from AustLII, and also address the issue of sub-dividing large documents for retrieval [9].

3.2 Network topology

One hundred words were selected from the collection using a cumulative [2] inverse document term weight [8] method, and an input and output unit created for each. Varying numbers of hidden units were tested over 700 epochs, and on the basis of performance a network with 10 hidden units was constructed.

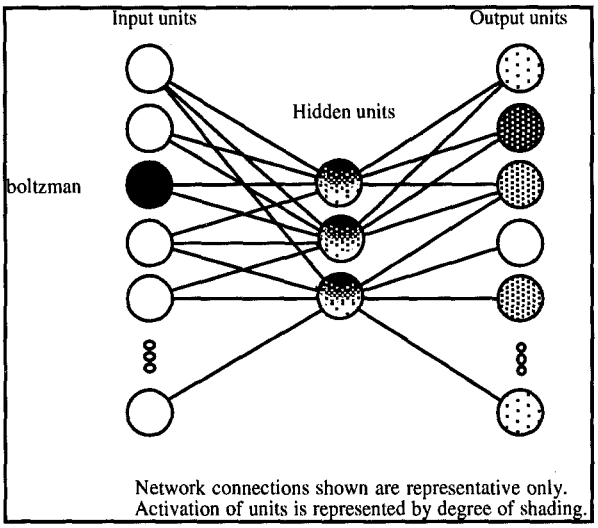

Fig. 1. Network schematic

3.3 Pattern Generation

Every occurrence of an indexed word in the document collection generates a training pattern. Input vectors can be described as input categories, since only a single word unit is activated for the pattern. This unit is activated with a magnitude of 1. The corresponding target output vector for each category is the document word frequency profile of the document containing the input word.

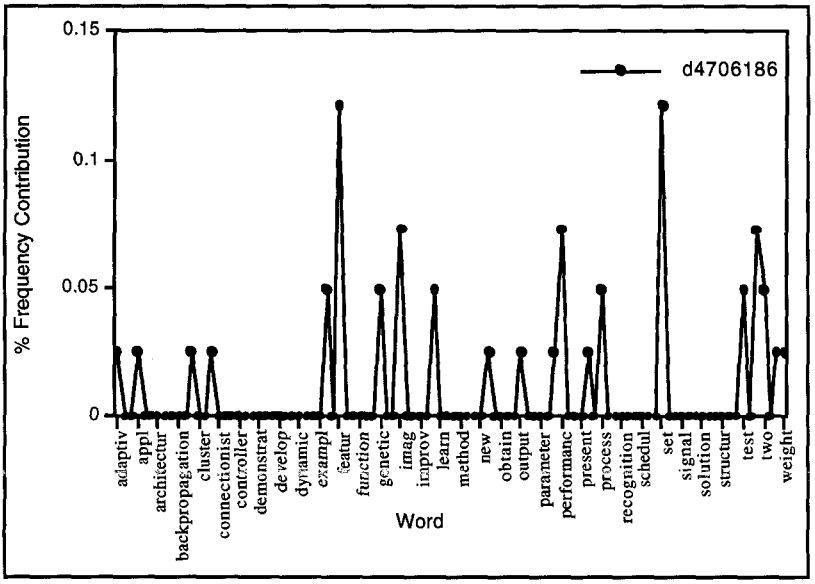

Fig. 2. A document word frequency profile

Document profiles were calculated by normalising the word frequency of each indexed word. The 2,191 abstracts generated 10,220 patterns.

The 621 legal documents (web pages on AustLII) generated 34,128 patterns, from which 22,760 were used for training and 11,368 for validation. The documents were Residential Tenancies Tribunal cases dealing with rental bonds. This data set was chosen because the cases are short and small in number, and are thus similar to our previous work using computer science abstracts [5].

3.4 Inherent error

This task required that the network generalise from the training set which includes multiple target output vectors for each input category. As a result, the training set could be considered to contain an amount of 'inherent' error, since the network would be unable to find a set of weights that exactly satisfies all of the mappings from the input categories to the document profiles.

To quantify this error so that the learning performance of the network can be evaluated, the minimum inherent error was estimated by calculating the total sum of squares of the difference between the maximum and minimum values for each term of the pattern vectors.

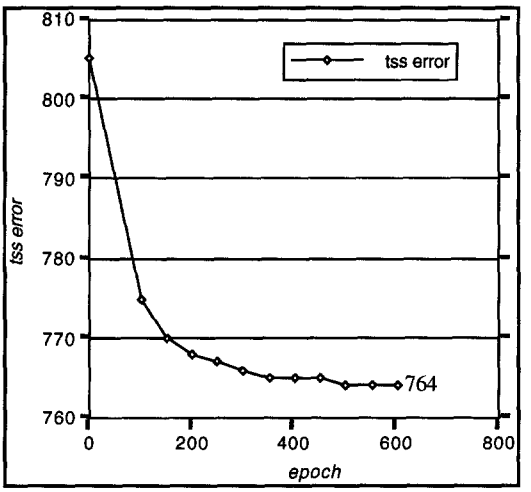

Fig. 3. Total Sum Squares error during training

For each category i with target vector $V_i[t_1, t_2, \ldots, t_n]$, with terms $V_i t_j$ the minimum inherent error is:

$$\sum_{j=1}^{N} \left(max\left(V_x t_j\right) - min\left(V_y t_j\right)\right)^2 \qquad (1)$$

where x, y range over the vectors for each category (input term), and N is the number of terms. Clearly, this method will underestimate the inherent error except in the

degenerate case where all the values are clustered together with just one outlier. However, the calculated value of 256.7 for our data set is an indication of the minimum error possible on this data set. Considering the minimum inherent error, the final TSS of 764 for 10,220 patterns appears acceptable. Networks with varying numbers of hidden units produced similar TSS error profiles.

3.5 Clusters Generated

Clusters are generated by activating a category and ranking the output word units by activation. The most highly activated word units were selected for each input category.

word | technique	Neural Network	Total Co-occ.	Ave. Co-occ.
connectionist	solution	model	model
	problem	fuzzy	fuzzy
	nonlinear	author	author
	filter	learn	learn
fuzzy	perform	control	controller
	learn	method	function
	computational	model	rule
	feature	rule	control
nonlinear	example	model	model
	input	method	filter
	connectionist	control	method
	neuron	problem	neuron

Table 1. INSPEC Comp.Sci. example clusters for words: action, consent, sign

The most striking observation is the similarity between the co-occurence measure cluster contents for a specific word, and even more the similarity of cluster contents for different source words. Note that none of the neural network derived clusters for each of the source words has any overlap.

word | technique	Neural Network	Total Co-occ.	Ave. Co-occ.
action	sum	tenant	tenant
	item	landlord	landlord
	pay	premise	premise
	provide	rent	act

consent	paid	tenant	compensation
	circumstance	landlord	claim
	follow	premise	clean
	item	agreement	tenant
sign	agreement	tenant	said
	pay	landlord	show
	set	premise	rent
	july	agreement	premise

Table 2. AustLII Legal example clusters for words: action, consent, sign

The network clusters are reasonably good, with the words found having some plausible conceptual relationships in the context of legal documents. The total co-occurence statistical measure is essentially useless in this data set, with most input words words producing very similar clusters. The average co-occurence clusters showed less of this effect, however, the words in the clusters are qualitatively less satisfactory than the network produced clusters.

We examined the document vectors we produced from the documents, below is a representation of the vectors using a 4 point scale.

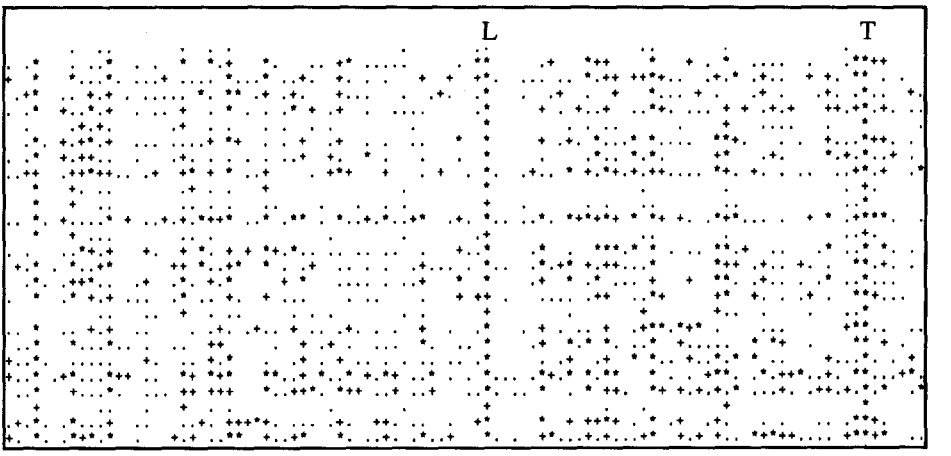

Fig. 4. Compressed representation of a few document vectors

Some words clearly occur in most documents and hence appear as a solid column in Figure 4. In particular, two words are identified. The letter L shows the location in the document representation of landlord, and T shows tenant. This explains the failure of the total co-occurence measure. It is worth noting that the neural technique has shown itself to be robust in this situation, and still produces useful clusters of words.

Issues remaining to be resolved are:

- whether the two statistical techniques can be improved to provide more reasonable comparison for the the neural network;
- improvement of the word selection technique, in that clearly the inclusion of landlord and tenant in the 100 words used to train the neural networks was not ideal in this case; and
- the contribution of the method used for scaling word importance in the production of the document vectors.

4 Results

The examples shown in the previous section are clearly not English synonyms. The backpropagation procedure which produced these words based on the nonlinear nature of artificial neurons is sensitive to the statistical distribution of the collection frequency data the network has been trained on.

The final decision on their usefulness remains to be tested in practice to determine whether the collections of words has value in denoting concepts. Note that Kumar and Lindley [6] have shown that even trigram information traces are suitable for hypertext information retrieval. We retain more information than they do. We believe that our clusters are adaptive to, and reflect higher order statistical dynamic information about the words in the specific (sub-)collection.

To test this assertion, we have performed a simple first order statistical analysis of the clusters, to determine if the network was producing clusters of greater complexity than local word co-occurrence. Two methods were used, clusters calculated using average word co-occurrences and total co-occurrences [5].

5 Conclusion

The clusters produced in this experiment can not be considered English synonyms. The clusters generated clearly have semantic associations that are specific to the document collection used to generate the patterns. These associations are close to being synonyms to the respective source words on a continuum from *synonyms* to *orthognyms*, the latter being two or more words which are in some sense orthogonal in meaning such that they denote a distinct new concept. For example, *artificial intelligence*. Here the concept is related more clearly to the latter word. The opposite is true for *traffic jam.*

The examples shown are plausible clusters depending on the source documents' origins. Notwithstanding that the results are clearly significantly different to those produced using simple first order statistical techniques [5], further qualitative analysis using a comprehensive domain thesaurus is required to understand more fully the semantic value in these clusters.

It is apparent that words occurring with high frequency in the collection are more often included in clusters, than those occurring only infrequently in the collection. A possible solution to this problem would be to use an alternative method when generating training patterns. The relatedness function proposed by Wilks [8] takes into account word frequency and could be modified to apply to single documents rather than

the whole collection. It may also be possible to speed the network learning and improve generalisation by scaling the network training patterns. The danger in this approach is that the already noisy relations inherent in the data may be obscured.

The technique described here has possible practical application to off-line processing of retrieval collections, and with further development, automated generation of synonyms that are domain specific [2]. Thesauri are useful to augment users queries, however the high costs of maintenance means that they can rarely be truly domain specific. Query enhancement strategies to improve information retrieval will become more practical when such thesauri are more readily available. The techniques described here apply back propagation neural networks to this problem in a way that has not been reported elsewhere.

References

1. Blair D.C. *Language and Representation in Information Retrieval*, Amsterdam, Elsevier, 1990.
2. Bustos, R.A. and Gedeon, T.D. "Learning Synonyms and Related Concepts in Document Collections," in Alspector, J., Goodman, R. and Brown, T.X. *Applications of Neural Networks to Telecommunications 2*, pp. 202-209, Lawrence Erlbaum, 1995.
3. Gedeon, T.D., Johnson, L. and Mital, V. "Neural Networks for Information Retrieval," in Mital, V. and Johnson, L. *Advanced Information Systems for Lawyers*, pp. 268-277, Chapman & Hall, 1992.
4. Gedeon, T.D. and Mital, V. "Information Retrieval in Law using a Neural Network Integrated with Hypertext," *Proceedings International Joint Conference on Neural Networks*, pp. 1819-1824, Singapore, 1991.
5. Gedeon, T.D. and Bustos, R.A. "Word-Concept Clusters in Document Collections," *Proceedings Australian Document Computing Conference*, pp. 21-24, Melbourne, 1996.
6. Kumar, V.R. and Lindley, C.A. "Improving Decision Support Through Hypermedia", *Proceedings, 3rd ACM Golden-West International Conference on Intelligent Systems,* Kluwer Academic Publishers, Las Vegas, 1994.
7. Salton, G. *The SMART Retrieval System - Experiment in Automatic Document Processing*, Englewood Cliffs, Prentice-Hall, 1971.
8. Wilks, Y., Guthrie, L., Guthrie, J. and Cowrie, J. "Combining Weak Methods in Large-Scale Text Processing" in Jacobs, P.S. *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, at pp. 35, 1992.
9. Zobel, J., Moffat, A., Wilkinson, R. and Sacks-Davis, R. "Efficient Retrieval of Partial Documents," *Information Processing and Management*, vol. 31, no. 3, pp. 361-377, 1995.