Word-Concept Clusters in Document Collections

T.D. Gedeon and *R.A. Bustos*

School of Computer Science Engineering
The University of New South Wales
Sydney NSW 2052, Australia

{tom, robertb}@cse.unsw.edu.au

Abstract

Finding relevant documents in large document collections is a major information filtering problem. In a legal information retrieval system many documents can be found using specialised connections mirroring the hierarchical and symbolically connected nature of legal documents. Nevertheless, many documents are only related to specific user information needs by their conceptual content. Individual words are not reliable indicators of concepts. We report here on an experiment using neural networks to learn clusters of words which are not based on simple statistical co-occurrence.

1 Background

The Australasian Legal Information Institute (AustLII), a legal research facility on the internet, was jointly established by the University of New South Wales and the University of Technology, Sydney. Funding for 1995 was provided by the Department of Employment, Education and Training from its Research Infrastructure Equipment and Facilities Program ('Mechanism C'), and supplemented by the two host Universities. Further funding has been received from the Law Foundation of NSW for 1996, and from the Australian Research Council Large Grants Scheme for 1996-1998.

The high volume use of the Australasian legal materials available via the internet on AustLII, provides an invaluable opportunity for research in information retrieval and index generation, particularly using techniques such as neural networks which require large numbers of patterns or instances for training. By August of 1995, AustLII the World Wide Web site (http://www.AustLII.edu.au) was averaging 4,000 hits per weekday, and 900 hits on weekend days.

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, does not require expensive expert intervention to maintain, and is adaptive to user needs.

Hypertext is a technology which meets the first two of these criteria, our aim is to use neural network and other AI learning techniques to discover useful connections [4] based on the document collections themselves, and to maintain and enhance the hypertext structure [5] 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 the task becomes one of adding words to the query that will have the least negative effects on precision. The addition of synonyms to user query words from a domain specific thesaurus is one way of selecting such terms, whilst maintaining the semantic integrity of the user query. This paper reports an experiment using neural networks to discover word-concept clusters which are not based on simple statistical co-occurrence, and may be useful substitute for a domain thesaurus when none is available. We note the work by Führ & Pfeifer [3] using logistic regression in combining model and description oriented approaches for probabilistic indexing, to discover the relative significance in practice of the various words. Feed-forward neural networks can perform logistic regression, and could eliminate the requirement for the additional heuristic strategies used.

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.

Proceedings of the First Australian Document Computing Conference, Melbourne, Australia, March 20-21 1996.

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 Initial version

This paper reports results of a preliminary experiment using INSPEC (computer science: neural networks) abstracts. The aim is to use on-line legal data from AustLII, and also address the issue of sub-dividing large documents for retrieval [10].

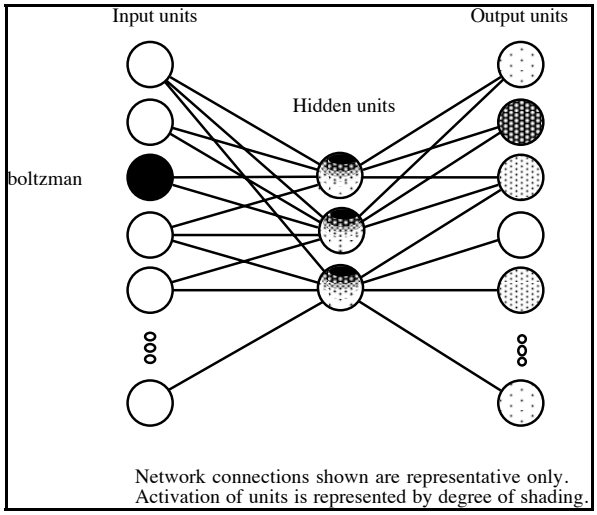

Figure 1. Network schematic

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.

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

3.4 Inherent error

This task required that the network generalise from the training set which includes multiple target output

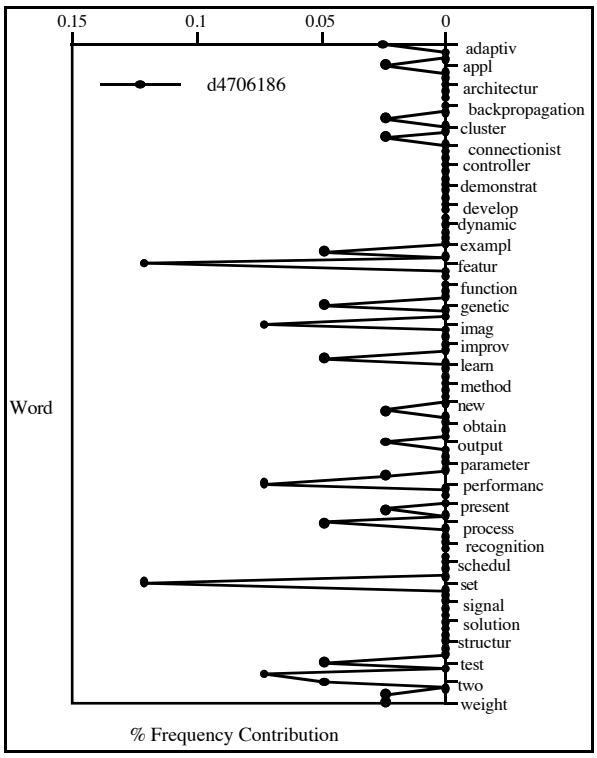

Figure 2. Sample document word frequency profile

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.

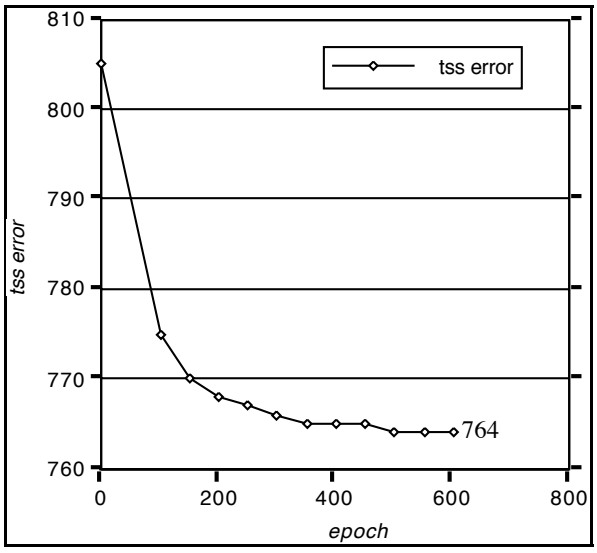

Figure 3. Total Sum Squares error during training

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:

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

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 five most highly activated word units were selected for each input category.

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.

Average co-occurrence clusters were generated by calculating the co-occurrence frequency for each word, and dividing by the number of documents that the word appeared in. The co-occurrence frequency is the number of times that the word appears in the same document as the indexed word under consideration. The five highest average co-occurrence value words were selected to form the cluster.

Total co-occurrence clusters were similarly generated using the co-occurrence frequencies without normalisation for the number of documents each word occurred in. Again, the five most frequently co-occurring words were selected to form the cluster.

The overlap between the network produced clusters and the two statistical methods were calculated by counting the average number of words that appeared in clusters produced by both methods.

nonlinear	example	input	connectionist	neuron	adaptive
adaptive	simulat#	anneal#	filter	design	optimization
shown	anneal	data	solution	given	describe
procedure	experiment	computational	weight	approach	search
cluster	adaptive	structure	propos#	genetic	pattern
dynamic	optimal	recognition	adaptive	develop#	simulat#
information	genetic	number	develop#	example	approach
neuron	local	fault	control	vector	nonlinear
decision	show	perform	experiment	structure	consider

Table 1: Example clusters

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 [7] have shown that even trigram information traces are suitable for hypertext information retrieval. We retain more information than

Network vs Total Co-occurrence	Network vs Average Co-occurrence	Total vs Average Co-occurrence
3.8%	4.0%	19.0%

Table 2: Similarity between Network and First Order clusters of 5 words

These results indicate that the network produced clusters share a single word with the statistically produced clusters in less than 1/5 of the clusters. The similarity between clusters produced by the two statistical methods five times as high.

The clusters of words produced by each of these methods were anonymously labelled by one of the authors, and the other author rated the clusters on a scale of 0 to 5, the score being the number of words which were judged to be relevant to the category word.

	5	4	3	2	1
Network	–	–	12	28	22
Ave. Co-occ.	–	8	18	33	5
Tot. Co-occ.	20	12	19	16	15

Table 3: Cluster ratings

These results are surprising, indicating the network was best at producing 1, 2 or 3 words relevant to a category word. The total co-occurrence measure clusters appear to be equally good irrespective of their size. The average co-occurrence clusters were more similar to the network clusters as expected.

5 Conclusion

The clusters produced in this experiment are not convincing as concept descriptors and could clearly not be considered English synonyms. However, the higher statistical order clusters generated clearly have semantic associations that are specific to the document collection used to generate the patterns.

The examples shown are plausible clusters considering the source documents' origins as records in an abstracts database. Notwithstanding that the results are clearly significantly different to those produced using simple first order statistical techniques, 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 [9] 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] Führ, N. & Pfeifer, U. "Combining Model-Oriented and Description-Oriented Approaches for Probabilistic Indexing," *14th International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 46-56, 1991.

[4] 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.

[5] 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.

[6] Jacobs, P.S. *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

[7] 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.

[8] Salton, G. *The SMART Retrieval System - Experiment in Automatic Document Processing*, Englewood Cliffs, Prentice-Hall, 1971.

[9] Wilks, Y., Guthrie, L., Guthrie, J. and Cowrie, J. "Combining Weak Methods in Large-Scale Text Processing" in Jacobs [6] at pp. 35, 1992.

[10] 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.