Clustering Significant Words using their Co-occurrence in Document Sub-Collections

Tamás D. Gedeon
School of Information Technology
Murdoch University, 6150 Australia
Phone: +61-8-9360-6430, Fax: +61-8-9360-2941
email: t.gedeon@murdoch.edu.au

ABSTRACT: Finding relevant documents in large document collections is a major information filtering problem, as individual words are not reliable indicators of concepts. Documents are related to specific user information needs by their conceptual content. Two experiments on a legal document subcollection are compared. The first experiment uses neural networks to learn clusters of words representing concepts not based on simple statistical co-occurrence. The second experiment uses clusters derived using fuzzy tolerance and similarity relations based on the counted or estimated values of (hierarchical) co-occurrence frequencies. These techniques can be used for high throughput information filtering to find documents likely to contain concepts relevant to a user's information need.

KEYWORDS: document processing; information filtering; neural networks; fuzzy logic; clustering; concepts.

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 was received from the Australian Research Council for 1996-1998 to Gedeon Greenleaf and Mowbray. The work reported in this paper was partially supported by this 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. From the end of August 1996, the AustLII site (http://www.AustLII.edu.au) average of 38,000 hits per work day has risen by July 1999 to 200,000 hits per work day.

INTRODUCTION

The problem domain is the provision of sophisticated access to legal information, which requires 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.

Web or hypertext systems (e.g. via AustLII) meet the first two criteria, our aim is to use neural network and fuzzy techniques to discover useful connections Gedeon et al. (1992) based on the document collections themselves, and to maintain and enhance the hypertext structure Gedeon and Mital (1991) 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 (1990) 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..

EXPERIMENTS

Two experiments on a legal document subcollection are compared. The first experiment uses neural networks to learn clusters of words representing concepts not based on simple statistical co-occurrence. The second experiment uses clusters derived using fuzzy tolerance and similarity relations based on the counted or estimated values of (hierarchical) co-occurrence frequencies. These techniques can be used for high throughput information filtering to find documents likely to contain concepts relevant to a user's information need. The neural network experiment is described first,

NEURAL NETWORK

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.

This experiment continues a preliminary experiment using INSPEC (computer science: neural networks) abstracts Gedeon and Bustos (1996). The aim is to use on-line legal data from AustLII, and also address the issue of sub-dividing large documents for retrieval Zobel et al (1995). A subcollection was formed by querying the AustLII web database at http://www.AustLII.edu.au with the following: "(bond* or deposit*) and not (no appearance)". As a result, 621 documents were retrieved.

Network topology

One hundred words were selected from the collection. To select the words we retained the cumulative (Bustos and Gedeon, 1995) inverse document term weight (Salton, 1971) method of the preliminary experiment. An input and output unit is created for each selected indexing word. Note that word stemming was done, so the use of the word "word" should be read as "word stem" throughout this paper. Varying numbers of hidden units were tested in the previous experiment over 700 epochs, and on the basis of performance a network with 10 hidden units was constructed. We have retained this topology for this experiment. It is illustrated in Figure 1, below.

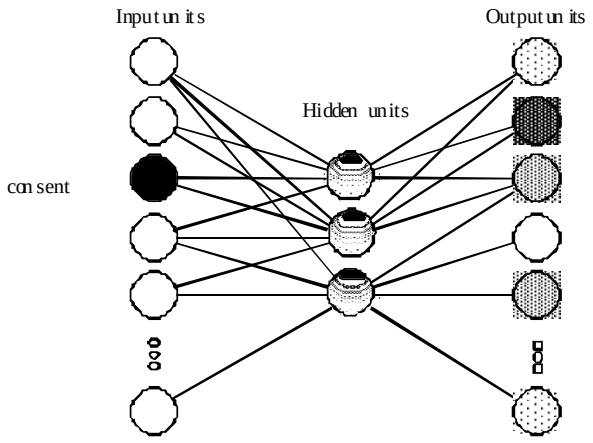

Figure 1: Network topology schematic.

FUZZY RELATION

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 621 documents 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.

FUZZY RELATIONS

Occurrence frequencies of the 100 words in the collection of documents were calculated. Based on the occurrence frequency – importance degree transformation sigmoid (Kóczy and Gedeon, 1998) defined in Figure 2 below, the frequencies are transformed into possibilistic importance degrees.

F	0	1	2	3	4	5	6	7	8	9
σ	0	0.1	0.2	0.4	0.7	0.9	0.95	0.98	0.99	1.00

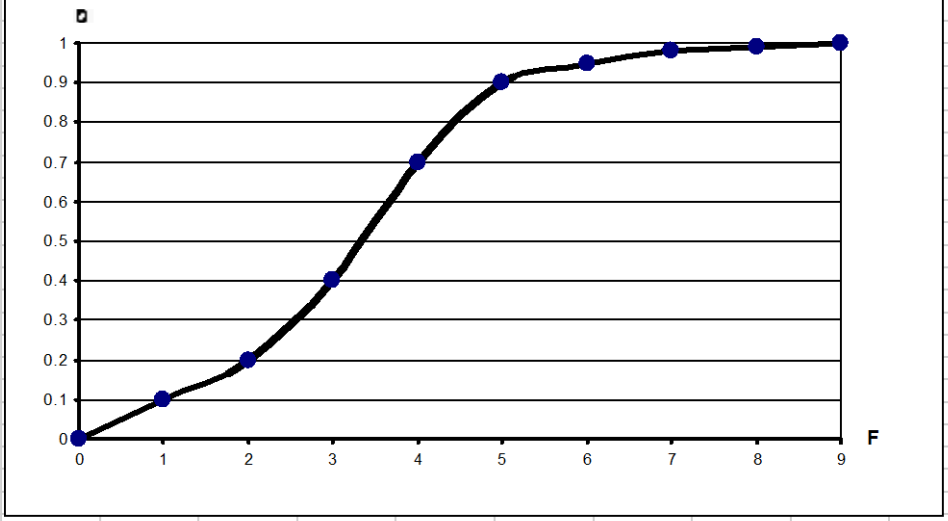

Figure 2: Example for sigmoid curve with typical occurrence frequencies.

Membership degrees or fuzzy measures range from 0 to 1, where 0 expresses the total lack of importance, and 1 stands for absolutely important. Words occurring very frequently are usually stop words (absolute or relative ones), and so they should be left out of consideration. For the remaining class of significant words it is generally true that higher occurrence frequencies indicate higher importance degrees as well. Although the connection between occurrence frequency (word count) and importance degree is strictly monotonic, it is certainly not proportional. The critical domain is somewhere what can be defined as "a few occurrences", depending on the type and size of the document, somewhere between 2 and 20 word counts. It does not matter much whether a word occurs in a document 20 or 22 times, it is highly likely that this document will be rather important for the querying person in both cases. On the other hand, one or two occurrences of a word might be coincidental or might indicate that the subject is touched upon only very superficially, while repeated mentioning (three, four, etc.) is an indicator that the word in question is an important word from the point of view of the document. With short documents these numbers might vary. It is quite different with keyword occurrences in titles, etc., where even a single occurrence usually indicates high importance.

The mapping from occurrence frequencies or counts to possibilistic membership degrees is thus a sigmoid function, with its steep part around the "critical" area of occurrences – the concrete values depending on the expected lengths and types of documents, and the category of environment (title, text, etc.).

Let us address now the problem of fuzzy co-occurrence graphs mapping the mutual relations of keywords into a set of fuzzy degrees. Here the fuzzy degrees are represented by the occurrence degrees . For each pair of words, a series of co-occurrence degrees can be calculated: one for each document in the collection. The average co-occurrence will be

calculated by applying the arithmetic means aggregation operation for each pair. It is interesting that self-equivalence is not 1, which can be explained by the axiomatic properties of fuzzy operations (cf. Klir and Folger, 1988). However, for practical purposes, reflexivity will be assumed in the establishing of fuzzy relational maps. Another observation is their symmetry, which results from the symmetric property of the relation described. We need to modify this calculation to take into account the similarity which follows from the fact that the some words occur with low counts, and many overlapping 0 counts increase the degree of equivalence. Because of this, in the rest we will modify the numbers by multiplying every degree by the average occurrence counts of the words in question.

RESULTS

To simplify the discussion, the number of words for the results described will be limited to the 16 most significant words (i.e. word stems) without loss of generality: agreement, bedroom, carpet, compensation, damag, evidenc, follow, liability, loss, material, occasion, premis, reasonable, replac, set, view.

FUZZY RESULTS

The top cut from the fuzzy tolerance relation obtained by weighting co-occurrence possibilities with the average occurrence counts calculated as above from occurrence counts and so as described above, produces just the following tolerance groupings:

{agreement, evidenc} and {agreement, follow, premis}.

The next cut produces the following groupings:

{agreement, evidenc, follow, premis} and {agreement, evidenc, follow, set} and
{compensation, evidenc, follow} and {follow, occasion} and {follow, replace} and {follow, view}.

NEURAL RESULTS

The equivalent of the top cut was produced by using only the top 5 words in terms of the neural network output.

{agreement, premis} and {carpet, damage} and {damage, occasion} and {occasion, reason} and
{evidence, material, follow} and {bedroom, liability}

The use of the top 10 words produces too many linkages and is not reproduced here.

DISCUSSION

There is some agreement between the techniques for the groupings of words. The words "agreement" and "premise" are linked by both, as is "evidence" and "follow". It is possible to conjecture that these represent the concepts relating to the nature of the legal contract, and the consequences of breaches, respectively. I.e., the legal contract is an agreement to rent the premises from the owner, while "evidence" is required or found for breaches and consequences "follow". We can note that the fuzzy technique links "agreement" and "evidence" also, which could relate to proof of extra conditions agreed to by the parties which do not form part of usual contracts.

CONCLUSIONS

In our preliminary experiments we demonstrated that the neural network finds clusters of words which are not due to simple co-occurrence. This was shown by comparison to both total and average co-occurrence measures, and a blind qualitative analysis of the clusters generated. This showed that the quality of the neural network clusters was best, and most similar to the average co-occurrence which is the more reasonable measure. Statistical analysis of the clusters showed that quantitatively there were substantially different from either statistical technique. The degree of similarity in the clusters produced by the neural and fuzzy techniques above clearly places it beyond statistical measures.

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 of words considering the source documents' origins. Some qualitative analysis using a comprehensive domain thesaurus is required to understand more fully the semantic value in these clusters.

It is apparent with the neural network technique 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. 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 fuzzy technique described here does not suffer from this problem to the same degree.

The techniques described here have possible practical application to off-line processing of retrieval collections, and with further development, automated generation of synonyms that are domain specific Bustos and Gedeon (1995). 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 can support the development of such thesauri.

REFERENCES

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

Bustos, R.A. and Gedeon, T.D., 1995, "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.

Gedeon, T.D., Johnson, L. and Mital, V., 1992, "Neural Networks for Information Retrieval," in Mital, V. and Johnson, L. Advanced Information Systems for Lawyers, pp. 268-277, Chapman & Hall.

Gedeon, T.D. and Mital, V. , 1991, "Information Retrieval in Law using a Neural Network Integrated with Hypertext," Proceedings International Joint Conference on Neural Networks, pp. 1819-1824, Singapore.

Gedeon, T.D. and Bustos, R.A. , 1996, "Word-Concept Clusters in Document Collections," Proceedings Australian Document Computing Conference, pp. 21-24, Melbourne.

Klir, G. and Folger, T., 1988, "Fuzzy Sets, Uncertainty and Information," Prentice Hall, Englewood Cliffs, NJ.

Kóczy, L.T., Gedeon, T.D. and Kóczy, J., 1998, "The construction of fuzzy relational maps in information retrieval," TR98-01, Dept. Information Engineering, University of New South Wales.

Kumar, V.R. and Lindley, C.A., 1994, "Improving Decision Support Through Hypermedia", Proceedings 3rd ACM Golden-West International Conference on Intelligent Systems, Kluwer Academic Publishers, Las Vegas.

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

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.