Learning Synonyms and Related Concepts in Document Collections

R.A. Bustos and T.D. Gedeon

School of Computer Science Engineering
The University of New South Wales
Sydney NSW 2052, Australia
{tom | robertb}@cse.unsw.edu.au

Abstract

For very large document collections or high volume streams of documents, 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. We use a neural network approach to learn synonyms and related clusters of words defining similar concepts from a sample document set. The task is then to filter the document collection to find more of the same.

1 The Information Filtering Problem

A *high-volume information source* is one where the rate of document arrival makes it infeasible for an individual to examine and assess the importance of every document. Each *document* is a (possibly structured) piece of text which is concerned with a set of topics. Users of the information source are interested in a particular set of *topics*. The *information filtering problem* is to select, from the incoming document stream, *all* documents which are closely related to the user's interests, and to select *only* those documents [1].

A number of researchers [2-4] have developed systems to assist with information filtering. A local example of an information filtering system is the grapeVINE system developed at the University of New South Wales [5]. The

well-known SMART information retrieval system [6], has also been applied to the task of information filtering on Internet news. All of these systems, however, require significant assistance from the user/information-provider in creating and/or maintaining the filters, specifying categories, building synonym lists, and so on. Hence we are interested in automatic indexing as applied to information filtering [7].

2 Automatic Indexing

The aim of indexing text items is to (implicitly) summarise their content. The possible approaches to this problem can be categorised according to whether they are syntactically-based or semantically-based. One extreme, the semantic, natural language understanding approach would construct a deep representation of the semantic content of documents. This is an extremely difficult, complex and as yet unresolved problem. At the other extreme, the information retrieval approach extracts a list of key terms from the document via simple syntactic processing, and devises a document signature based on the following significant measures [8]:

Frequency-keyword approach
: Take the complete text, remove the "stop" words, then sort all the remaining distinct words (keywords). Count the frequency of each keyword. Assign significance to keywords according to their frequency.

Title-keyword Approach
: Compile a list of keywords from the title, subtitle and headings of the document on the basis that the main concepts of the document are likely to be mentioned there. Higher significance can be assigned to the keywords from the main title, and so on.

The location method
: A keyword occurring at the introduction and/or conclusion of a paragraph is likely to be the most central to the theme of the text.

The Cue method
: Is based on the notion that certain of the words, which are not keywords, nonetheless increase or decrease the score of certain keywords, for example by the use of "significant" versus "impossible".

The Indicator-Phrase method
: Phrases about the topic of the text, for example "the purpose of this work," "the main aim of this paper" lend extra significance to following keywords.

Structure of the document

> This includes header information such as the title of this section, markup languages [9] and meta-language constructs used by the source community. Examples of these deriving from the Internet, for example, are: ":-)" to indicate something is being said in jest, use of repeated "!!!" and ALL CAPITALS for emphasis, and so on. Such information affects the weight attributed to key phrases in its vicinity.

The above significance measures form the indexing parameters for an automatic indexer. Current automatic indexing mechanisms assume that there is a best way for combining the significant measures to arrive at a particular signature for a document. Yet, the criteria for arriving at a particular combination is based on intuition rather than active observation of user behaviour. For example, the linear combination of such diverse indications of significance would require the assumption that each of them is providing independent evidence, which can not be readily justified to be correct. The problem of determining the most appropriate way of combining these indexing parameters appears to be amenable to solution using a neural network approach which can learn an appropriate composition function. This involves initial training of the neural network with the inputs which are the individual indicators of significance, and the desired output is the level of relevance of a document as judged by the user reader based on current interests. The hidden neurons in the network could learn the composition function required to best match the inputs (the significance measures) and the outputs (the relevance of the examples).

The first significance measure listed is a global measure, but has been used most extensively, and successfully in the past [10]. We use this method to find consistent indices in full text as follows.

The set of documents with keywords already known is used as a training set in supervised (Hebbian) learning, and their *L* link weights updated. We retain from prior work the concept of *Textual-Associative* (*T-A*) extra links between words and documents to indicate the statistical relationship between each document and the words it contains. This involves the size of the document, frequency of word occurence, and the overall rarity or commonness of the word. These factors are aggregated to produce the linking *T-A* weight.

Weight correlations in the multi-dimensional space of input patterns have been shown to be a good indicator of the difference of functionality of neurons [11]. The most consistent index generation will take place when the weights linking *word* and *document* neurons are most different, as the words then can best distinguish between different documents, via the generalisation of the keywords it has learnt. To discover the keywords for a new document, we dynamically add an extra neuron to the trained network. The extra *document* neuron has weights derived from its similarity (correlation) to the training documents (see Figure 1).

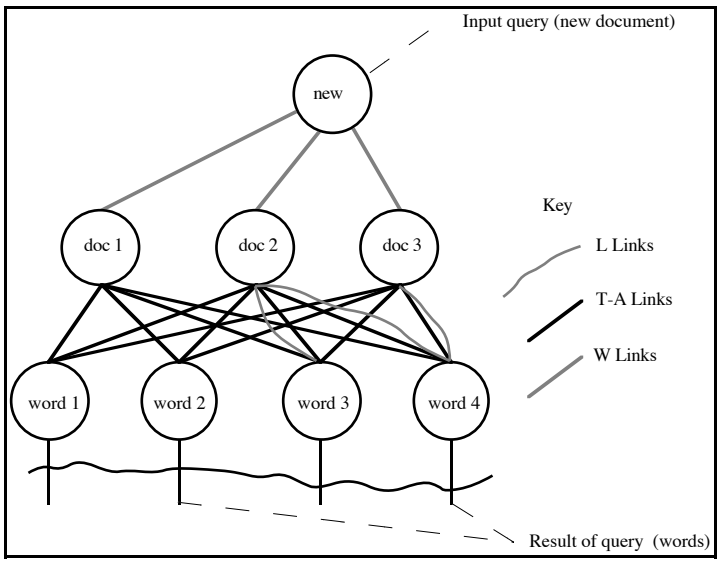

Figure 2. Dynamic (temporary) addition of a new document

This extended neural network is iterated (cycled) until convergence, when the *word* neurons which are active indicate the keywords generated for the new document. These keywords for the new document are used to determine whether it is to be filtered or retained.

3 Description of study

We wish to train a neural network to reproduce the word frequency measure component of a retrieval index.

A collection of 306 documents (being sections of a legal textbook) comprising of some 1,901 different words was chosen. By removal of stop words and stemming, the number of words was reduced. All words which occurred only once or twice in the whole document collection were removed, further reducing the number of words to some 831 words. This number of words was still too large to input into a supervised training neural network, so some other means of reducing the number of words needed to be found. All words occurring only once or in more than half the documents were removed leaving 200 words. From these, 75 words were selected manually looking for words with a moderate frequency and moderate document occurrence.

The network topology is shown in Figure 2. Note that not all connections are shown. The network was then trained normally using error back-propagation.

When being used for retrieval, it is necessary to create a representation of each document in the collection for comparison to the query vector. For these experiments we have used the [12] technique of inverse document-term frequency weighting (IDTW) to produce these vectors.

This technique (for reducing the size of vectors to be used in the comparison) seeks to identify the most important words in the collection by examining the cumulative total of word weights for each document. This was done by totalling the word-document significance values for each word. This resulted in a second short list of 75 words which were likely to be useful for distinguishing documents in the given collection.

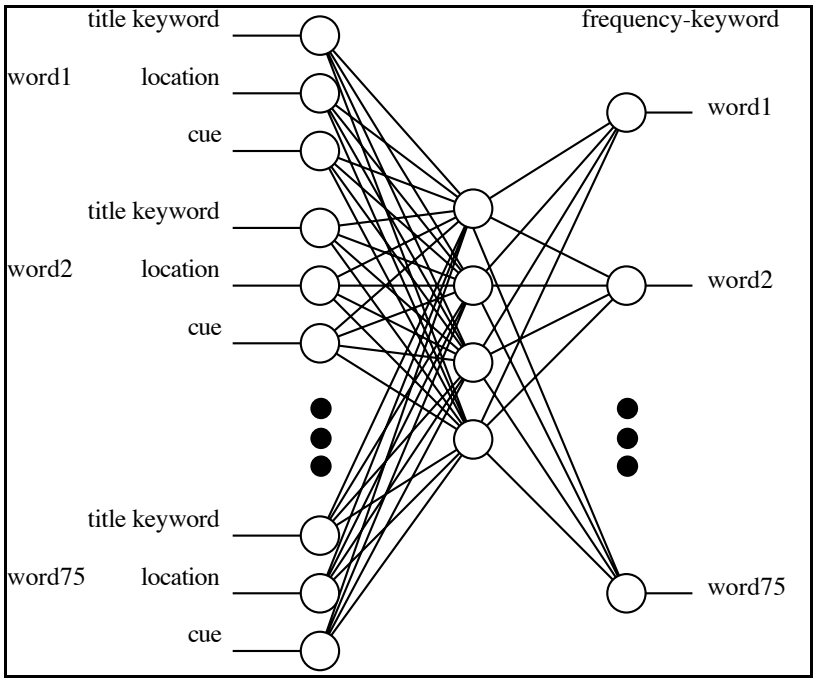

Figure 3. Neural network topology

The experiment is to use the title-keyword measure, the location measure and the cue measure as training input to a network, and attempt to predict the frequency-keyword measure. The network will learn the clusters of words forming related concepts, and will produce significant output activations for words which do not occur in documents if enough of the rest of the cluster is present. In effect, the network will be learning to perform a word cluster completion task.

The neural network topology we have chosen has three inputs for each of the

75 words. The network is using a relatively small proportion of the document texts as its input. Further, all of these measures can be calculated locally and do not require global document collection information as is required for the frequency-keyword calculations. This may provide efficiency gains in the future in highly parallel implementations.

4 Results

There are two ways we can test the network produced using comparing lists of documents retrieved.

The more difficult test is to use each document as a query to find similar documents. We compare its index (or vector) in the 75 dimensional space formed by the frequency-keyword values of the words to the vector of all of the other documents to determine which of these are similar. The top ten similar documents using the actual frequencies to form the vector (the traditional method) are compared to the top ten retrieved using the network's outputs as the components of the index vector. The average overlap is 60.9%. That is, six of the top ten documents in both lists are identical.

The other comparison that can be made is using the original queries which are relevant to the short list of 75 words that were used. The queries are converted into a 75 component vector, and compared to the vectors for the 306 documents, using both the frequency-keyword values and network output values as before.

The average overlap in this case is 100%. That is, we have shown that the neural network can reproduce the behaviour of the frequency-keyword vector retrieval using only the title, location and cue information, in a specialised subset of the document domain. The previous result of 61% demonstrates that some generalisation to the overall document domain was also taking place.

We have also applied a simple analysis technique to the weigth matrix of the neural network to determine the relative importance [13] of the three input measures, as shown in equation (1):

$$C_{title} = \frac{\sum_{i=1}^{75}\sum_{j=1}^{4}\left|w_{title_{i-j}}\right|}{\sum_{h=cue}^{title}\sum_{i=1}^{75}\sum_{j=1}^{4}\left|w_{h_{i-j}}\right|}.100\,\% \qquad (1)$$

This is calculated by summing the absolute magnitude of the weights from a particular measure to the hidden units, and dividing by the sum of all the weights connecting to the hidden units.

Table 1 shows the results for the relative importance of the three kinds of syntactic processing derived significance information represented by the inputs, as produced for the network before and after normal back-propagation training.

Table 1: Relative importance of inputs

	Title	Location	Cue
before training	32.4%	33.5%	34.2%
trained, simple choice	14.0%	40.7%	45.3%
trained, IDTW choice	22.8%	37.4%	39.8%

Before training the relative significance of the three measures is the same, as expected due to the random initialisation of network weights.

After training, the Cue method is shown to be the most important, closely followed by the location method. The title method in this domain is relatively unimportant. This accords with observations regarding the quality of the titles of the documents used, which all seem to have fairly similar words in their titles. This clearly demonstrates our contention that the relative importance of various measures will differ across domains. In many domains it is accepted that title keyword measures are important.

The 'better' choice of words using the cumulative inverse document term weighting scheme produced slightly higher contribution for the Title method, however the overall greater significance of the Location and Cue methods remains, as well as the slightly greater relevance of the Cue method.

5 Conclusion

For very large document collections or high volume streams of documents, finding relevant documents is a major information filtering problem. We use a neural network approach to learn synonyms and related clusters of words defining similar concepts from a sample document set. The task is then to filter the document collection to find more of the same.

The network will learn the clusters of words forming related concepts, and will produce significant output activations for words which do not occur in documents if enough of the rest of the cluster is present.

We contend that the difference in the predicted and actual frequency keyword values are the most significant measure of the usefulness of any potential cluster member words found.

References

[1] Ngu, AHH, Gedeon, TD and Shepherd, J "Discovering Indexing Parameters for Information Filtering,"*Proceedings 2nd International Conference on Intelligent Systems*, 6 pages, Singapore, 1994.

[2] Fischer G and Stevens C, "Information access in complex, poorly structured information spaces", *CHI'91 Conference Proceedings*, 1991.

[3] Foltz, PW and Dumais, ST, "Personalized information delivery: an analysis of information filtering methods", *CACM*, vol. 35, no. 12, pp.51-60, 1992.

[4] Goldberg, D, Nichols, D, Oki, BM and Terry, D, "Using collaborative filtering to weave an information tapestry", *CACM*, vol. 35, no. 12, pp.61-70, 1992.

[5] Brookes, C, *grapeVINE: Concepts and Applications*, Office Express Pty. Ltd., 1991.

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

[7] Gedeon, TD and Ngu, AHH "Index Generation is better than Extraction," *Proc. International Conf. on Non-Linear Theory*, pp. 771-774, Hawaii, 1993.

[8] Paice, CD "Constructing Literature Abstracts by Computer: Techniques and Prospects," *Info. Proc. and Management*, vol. 26, no. 1, pp. 171-186, 1990.

[9] Botham, A, Fuller, M, Mackie, E, Sacks-Davis, R, Wilkinson, R "An SGML-based Hypertext Information Retrieval System," *Australian Computer Science Communications*, vol. 15, no. 1, pp. 111-123, 1992.

[10] Gedeon, TD 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.

[11] Good, RP and Gedeon, TD "Network Analysis Techniques as Visualisation Tools," *Proceedings International Conference on Non-Linear Theory (NOLTA)*, pp. 945-952, Hawaii, 1993.

[12] Rose, D and Belew, "A Connectionist and Symbolic Hybrid for Improving Legal Research," *International Journal of Man Machine Studies*, vol. 35, 1991.

[13] Wong, PW, Gedeon, TD and Taggart, IJ "An Improved Technique in Porosity Prediction: A Neural Network Approach," *IEEE Transactions on Geoscience and Remote Sensing*, (in press) 1994.