LEARNING FUZZY MEASURE FUNCTIONS FOR INFORMATION RETRIEVAL

T.D. Gedeon [1,2], L.T. Kóczy [1]
A.H.H. Ngu [2], R.A. Bustos [2] and J. Shepherd [2]

[1] Department of Telecommunications and Telematics
The Technical University of Budapest
Budapest H-1111, Hungary
Fax: +36 1 204 3107

[2] School of Computer Science Engineering
The University of New South Wales
Sydney NSW 2052, Australia
Fax: +61 2 385 5995

Abstract

Words are significant to the documents in which they occur to varying degrees. We use fuzzy measures to assign values to each crisp set of words contained in queries for each document, signifying the degree of evidence or belief that that particular query belongs to the set of words representing each document. This assigns a value indicating the relevance of the query to particular documents. The retrieval of a ranked list of documents for a query becomes the relatively simple matter of the aggregation and subsequent ranking of the individual measure degrees. Creating the fuzzy measure functions for each document remains a difficult problem to be solved. We have used a neural network to learn these functions in a model domain with complete accuracy, and showed adequate generalisation performance to a wider domain.

Introduction

An information retrieval system allows users to efficiently retrieve documents that are relevant to their current interests from a very large collection of documents. A user typically specifies their interests via a set of words, for example a fragment of natural language text. The system then determines how closely each document in the system matches the specified interests and displays only those documents which match most closely. Ideally, information retrieval systems minimise the number of relevant documents not retrieved (*recall*), minimise the number of irrelevant documents retrieved (*precision*) and do all of this efficiently.

In practice it is not possible to achieve this goal perfectly because there are several areas of imprecision inherent in the process:

- the user may not know precisely what their interests are,
- the user may not be able to precisely specify their interests in words,
- the content of the document may not be able to be precisely specified (indexed), and
- the notion of relevance or matching is not precisely defined.

We describe first an idealised information retrieval process in which there is no imprecision. Thus, documents and queries are described using a crisp set C of *concepts* $\{C_1, C_2, \ldots, C_c\}$. The document database is made up of a set D of documents $\{D_1, D_2, \ldots, D_d\}$. Each document D_i is associated with a set of concepts $C_{D_i} \subseteq C = \{C_1, C_2, \ldots, C_{D_i}\}$. Since a document D_i comprises a sequence of words $\langle W_1, W_2, \ldots, W_{w_i} \rangle$, we require an *indexing* function which maps this sequence of words into a set of concepts which are relevant to the document:

$$\langle W_1, W_2, \ldots, W_{w_i} \rangle \rightarrow \{C_1, C_2, \ldots, C_{D_i}\}$$

Similarly a query Q is associated with set of concepts $C_Q \subseteq C = \{C_1, C_2, \ldots, C_q\}$.

If we regard the query as a conjunction of concepts, in that we want documents that contain all of the concepts, we can characterise the matching process as:

$$\begin{aligned} match_{\wedge}(Q, D) &= \{D_i \mid C_Q \subseteq C_{D_i}\} \\ &= \{D_i \mid C_Q \cap C_{D_i} = C_Q\} \end{aligned}$$

We can similarly characterise the other extreme where we regard the matching process as a disjunction of concepts, where we want documents that contain any of the concepts.

Both these extremes present difficulties, $match_{\wedge}$ has potentially poor recall, since a document may contain most of the specified concepts, but be rejected on the grounds that it is missing one or two of them. The $match_{\vee}$ has potentially poor precision, since documents will be accepted containing only one of the concepts and none of the others.

We wish to define fuzzy measures to assign values to each crisp set of concepts embodied by queries for each document, signifying the degree of evidence or belief that that particular query belongs to the set representing each document. That is, assigning a value indicating the relevance of the query to particular documents. A fuzzy measure is defined as a function:

$$g: \mathcal{P}(C) \rightarrow [0, 1]$$

and must satisfy the boundary conditions and monotonicity axioms of fuzzy measures:

boundary conditions: $g(\varnothing) = 0$ and $g(C) = 1$
monotonicity: $\forall A, B \in \mathcal{P}(C)$, if A $\subseteq$ B then $g(A) \leq g(B)$

In this paper we will subsequently use the vector retrieval notion as the fuzzy measure for the matching between query and document. That is, we will form a vector of the full set of concepts C, and the query and document representations produce vectors consisting of 0s and 1s. The query and document vectors are compared using the (first quadrant) angle they form in the c dimensional concept space. An angle of zero degrees produces a measure of 1 indicating a complete match, and ninety degrees produces a measure of 0, thus satisfying the boundary conditions. The abundant literature on the use of the vector method for information retrieval presupposes conditions equivalent to the monotonicity we require, hence we will take this as given and not attempt to demonstrate it here.

Note that the vector representation can readily be extended to include fuzzy presence of concepts, by the use of membership degrees other than 0 or 1 as components of the query. We will require this later.

The model using crisp sets of concepts with fuzzy measures has two remaining difficulties. The first is that discovering the concepts embodied in documents and queries is difficult. We have used legal documents in the first instance, because legal language has a more formally defined structure in terms of the relationship of words to concepts, nevertheless, the words in the documents are still only imprecise indicators of the concepts. Manual indexing of the documents is too time consuming and expensive, and suffers from the well known problems associated with the proliferation of labels, and the broadening of label meanings. Automatic indexing is an active area of research (Paice, 1990, Brookes, 1991, Gedeon and Ngu, 1993), and is beyond the scope of this paper.

Thus, we extend the vector representation to using words instead of concepts as components of the vector. This has two main consequences, due to the imprecise nature of words as denoting concepts. First, the number of words is much larger than the number of concepts in the same

document, and second, the frequency of occurrence of words in documents becomes important unlike concepts which are unitary notions. Thus, we will require larger dimensionality vectors with values in the interval $[0, 1]$.

The second difficulty is that of calculating the membership functions themselves. In this study we have used the inverse document keyword frequency measure (hereinafter referred to as the frequency-keyword measure for brevity) as an approximation of the membership function. Thus we have closed the circle, as this makes the vector retrieval testing model we have chosen most natural, with frequency-keyword vector components being natural for vector retrieval.

Description of the data and experimental aim

A collection of 352 short legal documents comprising of some 7,500 words was chosen. After removal of stop words and stemming, the number of words reduced to 5,400. All words which occurred only once or twice in the whole document collection were removed, further reducing the number of words to 1,470. 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.

A number of queries and the relevant documents to be retrieved for each query were available for a specific reader in the domain of this legal document collection. The words in the queries were reduced in number using the same process described above, removing the same stop words and any artefacts of the print to digital conversion process. From the remaining 830 words, a short list of 75 words was produced:

i) remove words with total word occurrence (wo) = 1 and query occurrences (qo) = 1,
ii) remove words with $qo > 35$ – these words appear in more that half the queries, and probably are not very useful for discriminating between queries and hence documents to be retrieved, and
iii) from the remaining 200 words, select the 75 words with $5 < wo < 9$, and $1 < qo < 7$.

The above choices are based on the notion that the ideal words to use to index documents are those that are neither too common, nor too rare (Blair and Marron, 1985). Using words which are too common provides limited resolution, while the use of words which are too rare is inefficient. An alternative method using the cumulative inverse document keyword frequency measure was also used to select the set of 75 words. The results were not markedly different, and are not reported here.

The 75 words chosen can be used to index the 352 documents in the overall collection. There are 46 documents which do not contain any of these 75 words, leaving a collection of 306 documents. 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 as the approximation to the fuzzy measure. The success of this approximation can then be tested using the queries, as well as on the overall document collection.

Neural network

The neural network topology we have chosen has three inputs for each of the 75 words. These are the measures for the title keyword, location and cue measures. These were calculated as follows.

The title lines of the documents were scanned to determine frequency of occurrence of the 75 words, and the resulting vector of fuzzy measure values normalised to 1. This accounts for multiple occurrences of words in titles (rare), and if there are a large number of words in the title, the significance of each of these is less.

The location measure was calculated using only the first and last 20% of each document, and a normalised frequency measure is produced. These values were chosen to reflect the increased significance of the beginning and ends of documents, and to reduce the number of words processed by over one half. Thus, only 40% of the words are processed in this measure.

Similarly, for the cue method, a window of 5 words on both sides of all cue words is examined for the presence of other words, the frequency is then normalised as above.

The network topology is shown in Figure 1. Note that not all connections are shown. The network was trained normally using the error back-propagation algorithm. The network after training has formed some internal representation of the combination of the different significance measures required to produce the output, which is the frequency-keyword as the approximation to the fuzzy measure.

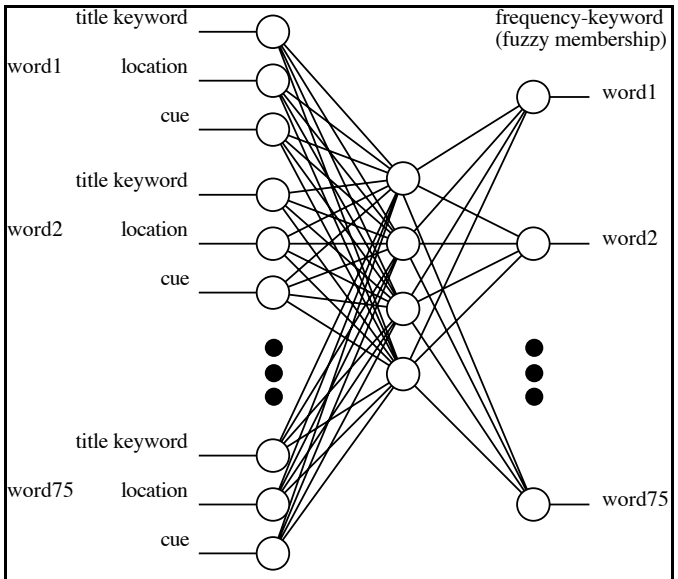

Figure 1. Neural network topology

The neural network output fuzzy measure can be tested in comparison to the calculated frequency-keyword values. Vectors formed using the calculated frequency-keyword are used to retrieve a list of documents, and then compared to another list retrieved using the neural network fuzzy measure outputs. The percentage overlap in the top 10 retrieved documents gives some estimate of the usefulness of the trained neural network, and which also has implications to the relevance of the input parameters used in training the network. In terms of the fuzzy measure functions, we have chosen an aggregation operation on these fuzzy measure values which is meaningful in the task domain, as an information retrieval task.

Note that the network is using a relatively small proportion of the document texts as its input. Further, all of the input values 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. Thus, we have used the frequency-keyword measure as a cheap and efficient approximation to the fuzzy measure functions we need, and have incidentally further reduced the computational cost.

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. This does not use our knowledge that the 75 words were chosen to be specific to a particular

collection of queries. We compare a document's index (or vector) in the 75 dimension space to all of the other documents to determine which of these are similar. The top ten documents using the calculated frequency-keyword values (as the components of the index) are compared to the top ten retrieved using the network's fuzzy measure value outputs (as the components of the index).

The average overlap is 60.9%. That is, six of the top ten documents in both lists are identical.

The comparison for which we have designed the experiment considers the original queries which were used to provide the short list of 75 chosen words. 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 the network output fuzzy measure values as before.

The average overlap in this case is 100%. That is, we have shown that the neural network can calculate the fuzzy measure values with an accuracy sufficient to provide completely correct behaviour on the retrieval task 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.

The computational complexity of this task is high, the network training took many hours on a fast Sun workstation. Replacing the crisp concepts in the vector representation by words increases it further. In the specific document domain it has been estimated that some twenty major concepts are present. We have used seventy-five words in the experiment! Creating this many fuzzy measures manually or by a sequential algorithm would be inefficient. Note that the benefits of the parallel processing of neural networks is still waiting on the widespread availability of cheap reliable neural network hardware implementations.

We have applied a simple analysis technique to the neural network to determine the relative importance of the three inputs (Wong, Gedeon and Taggart, 1994), to give some indication of the average relative contributions of the three different methods to the fuzzy measures produced by the neural network.

$$A_{title} = \frac{\sum_{i=1}^{75}\sum_{j=1}^{4}\left|w_{title_i j}\right|}{\sum_{h=title}^{cue}\sum_{i=1}^{75}\sum_{j=1}^{4}\left|w_{h_i j}\right|}.100\,\%$$

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

The following results are produced for the network before and after training:

	Title	Location	Cue
before training	32.4%	33.5%	34.2%
450 epochs training	14.1%	40.7%	45.3%

Before training the relative significance of the three inputs 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 indicates that the relative importance of various measures may differ across domains – in many domains it is commonly accepted that title keyword methods are very important.

Conclusion

Single words do not express crisp concepts which can be matched against the concepts embodied by documents. Instead, single words are significant to the documents in which they occur to varying degrees. This can be expressed as the membership of a fuzzy set denoting the *relevance* of terms to each document in the collection.

Our results demonstrates that some significant generalisation to the overall document domain was taking place. That is, the network has not just memorised the queries. Note that the queries were used to determine the 75 words to use, but these words comprised less than 10% of the words extracted from the queries, and the network was trained using the whole document collection, not just the queries. The 75 words comprise just 1% of the total number of words, which make the results on the overall collection more impressive.

We have used a neural network to learn the (implicit) fuzzy measure functions in a model domain with 100% accuracy, and showed adequate generalisation performance of 61% on the overall collection.

References

Blair, DC & Marron, ME "An Evaluation of Retrieval Effectiveness for a Full-text Document Retrieval System," *CACM*, vol. 28, no. 3, pp. 289-299, 1985.

Brookes, C, "grapeVINE: Concepts and Applications" Office Express Pty. Ltd., 1991.

Gedeon, TD and Ngu, AHH "Index Generation is better than Extraction," *Proceedings International Conference on Non-Linear Theory*, pp. 771-774, Hawaii, 1993.

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

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.