FUZZY RELEVANCE VALUES FOR INFORMATION RETRIEVAL AND HYPERTEXT LINK GENERATION

T.D. Gedeon [1], S.Singh [1,2], L.T. Kóczy [3] and R.A. Bustos [1]

[1] School of Computer Science Engineering
The University of New South Wales
Sydney NSW 2052, Australia.
Phone #: +61 2 385 3965 .
Fax #: +61 2 385 5995 .
tom@cse.unsw.edu.au .
.

[2] School of Computing .
University of Plymouth
Drake Cr PL4 8AA UK
Fax #: +44 1752 232540

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

ABSTRACT: Concepts are significant to the documents in which they occur, unfortunately finding the concepts is a difficult task. The words in documents are only imprecise indicators of concepts. We report here on an experiment to evaluate a fuzzy importance measure applied to a retrieval problem, and is the continuation of previous work using learning techniques to find fuzzy measures. Our measure is used to assign values for the relevance of a query to particular documents. The retrieval of a ranked list of documents for a query then becomes a simple matter of the aggregation and ranking of the values. Our overall goal is to find concepts automatically for index generation and hypertext link creation.

1 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 primarily maximise user satisfaction. This is very difficult to measure reliably and in a comparative fashion, hence some property or properties which can be measured is required. The goal becomes to minimise the number of relevant documents not retrieved (*recall*), minimise the number of irrelevant documents (*precision*), and do this efficiently. Note that it is not known in general how user satisfaction correlates with measures of recall and precision (Turtle, 1995).

In practice it is not possible to achieve this goal perfectly due to 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.

2 CRISP INFORMATION RETRIEVAL

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 below (left):

$$match_{\wedge}(Q, D) = \{D_i \mid C_Q \subseteq C_{D_i}\} = \{D_i \mid C_Q \cap C_{D_i} = C_Q\} \quad \cdots \quad match_{\vee}(Q, D) = \{D_i \mid C_Q \subseteq C_{D_i}\} = \{D_i \mid \exists C_q : C_q \in C_{D_i}\}$$

We can similarly characterise (above, right) 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 only one 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.

3 VECTOR RETRIEVAL

In this paper we will use the vector retrieval notion 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. Thus, we can extend the 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. Firstly, the number of distinct words is much larger than the number of distinct concepts in the same document. Secondly, the frequency of occurrence of words in documents becomes important. This is unlike concepts which are unitary notions. Thus, we will require larger dimensionality vectors with values in the interval $[0, 1]$.

The words in documents are 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, and is beyond the scope of this paper, but is the overall goal of our research.

4 INVERSE DOCUMENT TERM WEIGHT (IDTW)

For each indexed word j in each document i the IDTW values are calculated by the following method:

$$IDTW_{ij} = \frac{F_{ij} \times \log(N / DF_j)}{\sqrt{\sum_{k=1}^{T} (F_{ik} \times \log(N / DF_k))^2}}$$

where:

N = Number of documents
T = Number of terms
F_{ij} = Frequency of term j in document i.
DF_j = Document frequency of term j.

The use of IDTW has been evaluated in detail by Salton (1989).

5 COLLECTION

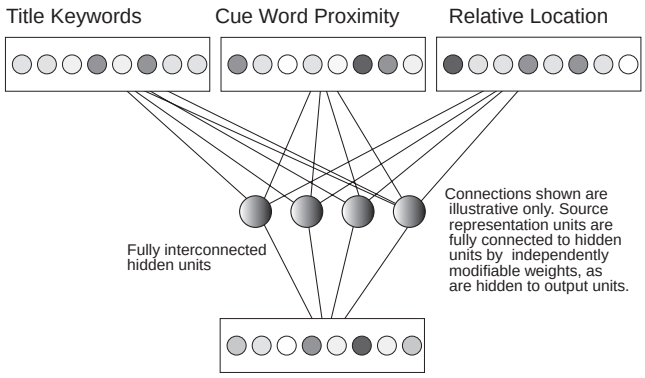
The initial collection used was a set of 2,191 abstracts of papers in the neural networks literature, and is the version this paper reports on. We are continuing the work using a collection of legal documents from the Australasian Legal Information Institute (http://www.AustLII.edu.au), as we expect the greater precision in the meaning of words in legal language may substantially improve our results. The first n abstracts were chosen from each separate INSPEC database.

FULL TEXT QUERY	SOURCE: INSPEC	RETRIEVED / DOWNLOADED
neural network OR connectionist OR backpropagation	Jan 1994 - Jun 1994 Jan 1993 - Dec 1993	1786 retrieved, 700 downloaded 3599 retrieved, 700 downloaded
boltzman OR simulated annealing	Jan 1994 - Jun 1994 Jan 1993 - Dec 1993 1992	237 retrieved, 200 downloaded 370 retrieved, 300 downloaded 426 retrieved, 291 downloaded

6 NEURAL NETWORK EXPERIMENT

A neural network was trained using 720 documents from the total documents collected. During training the network was tested with a small number of vectors (47) selected randomly from this smaller set, to determine an appropriate time to stop training. This test set of patterns was not used to update the network weights during training. The remaining two thirds of the 2,191 documents were retained to evaluate the generalisation performance.

One hundred words were selected by the cumulative IDTW values method (Bustos and Gedeon, 1995), and a three input and one desired output vector component created for each. The task the network attempts to solve is that of learning the relationship between title cue and proximity measures to reproduce the IDTW measure. The effectiveness of the choice of words is 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. A schematic of the network topology is below.



The neural network has three inputs for each of the 100 words. These are the measures for the title keyword, location and cue measures, calculated as follows:

Title: The frequency of occurrence of the target words in the title lines only of the documents is counted, and the resulting vector of values normalised to the range [0, 1]. If there are a large number of words in the title the significance of each of these is reduced.

Location: The location measure was calculated using the first and last 20% of each document to double the frequency of any words encountered, and a normalised frequency measure is produced.

Cue: A window of 5 words on both sides of cue words is examined for target words, the frequency count is again doubled, and then the frequency is normalised as above. Examples of cue words are 'overview, 'therefore', 'significant'.

After training the network for many epochs, a network output after 600 epochs was selected as the best network approximation of the calculated IDTW vectors, using the test set as described above. These network produced output vectors were then used as document representations for vector retrieval. Calculated IDTW vectors for each document were used as the query in each case (very much like query by example). To evaluate this retrieval some calculated IDTW vectors were used as document representations for vector retrieval and also as queries for each document.

7 FUZZY IMPORTANCE MEASURE

The measure is calculated as follows:

Importance ij = w1 * LocationKey + w2 * cueWordsFreq + w3 * TitleKeyFreq

where:

Importance ij is the importance of query word i in doc j,

w1, w2 , w3 are the respective weights to be given to each representation in making up the importance measure derived from an analysis of the network input contributions (Wong et al, 1995), and

LocationKey is the document representation as described above and used in the neural network experiment.

In addition to finding appropriate weights for each component representation, we also considered desirability of dividing each representation by the global frequency of each word. Based on sample queries and a conceptual analysis of the results to use the IDTW weights, since: this allows an explicit link with the neural network work; these weights are also domain specific and take into account the peculiarities of the doc collection; and to not divide by global frequency as the effect was only to de-emphasise the contribution of high frequency words.

The importance measure was extended to allow multiple query words by weighting each query word by its 'share' of the overall query as supplied by the user and combining the importance values in a simple weighted sum:

ie for query = data, recognition; weighted 2,1

Importance (Doc A) = Importance (data) * 2 / (2+1) + Importance (recognition) * 1 / (2+1)

8 QUERIES

In the neural network experiment we used the entire document vector as a query vector. This has some plausibility in practice as users will occasionally like to retrieve documents similar in content to a known document. Nevertheless, for practical use, we require queries which consist of a relatively small number of words. We had no queries available, and have chosen to derive synthetic queries from the document representations. Thus, the representation is compressed to the desired query length.

This choice again allows us to make explicit comparisons to the neural network experiment and the control vector retrieval result. Further, the vector components remaining after compression to the desired query length are used to provide the weighting for the query words in our importance measure.

9 RESULTS

Our importance measure and the neural network output 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 *R2*, and then compared to another list retrieved using the neural network measure *R1*, or the documents retrieved using the importance measure *R3*.

The percentage overlap in the top 10 retrieved documents gives some estimate of the usefulness of the methods. In terms of fuzzy measure functions, we have chosen an aggregation operation (vector retrieval and ranking) on these 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. 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 reduced the effort required. The computational cost of this task is still high, the network training took many hours on a fast Sun workstation. Creating this many fuzzy measures manually or by a sequential algorithm would be inefficient. The importance measure we introduce here is clearly more efficient, as the neural network was not really necessary to provide the weightings for the three measures, in that we could have relied on our available domain knowledge.

This experiment was done with multiple length queries to produce the following retrieved sets:

Retrieval set	Document Representation	Query Source Representation
R3(3)	Fuzzy Importance measure	Length = 3, IDTW
R3(5)	Fuzzy Importance measure	Length = 5, IDTW
R3(20)	Fuzzy Importance measure	Length = 20, IDTW
R2	Calculated IDTW	Calculated IDTW
R1	Network Approximated IDTW	Calculated IDTW

Note that all of the queries are based on the calculated IDTW document representations, with the full vector used in the last two entries in the above table, while the first three use the compressed queries synthesised from the IDTW representation.
The following comparisons were then made between the retrieved sets as follows:

Set Name vs	Set Name	% of shared retrieved documents (st. dev.)
R3(3)	*R2*	45.5 (17.9)
R3(5)	*R2*	48.3 (17.1)
R3(20)	*R2*	52.5 (15.1)
R1	*R2*	60.0 (28.4)
R1 $\cup$ *R3(3)*	*R2*	72.1 (21.0)
R1 $\cup$ *R3(5)*	*R2*	73.3 (20.4)
R1 $\cup$ *R3(20)*	*R2*	76.6 (18.6)

The importance measure using only three word queries produced surprisingly good results, considering that only three of 100 vector components were used. With increasing lengths of queries, the results are tending to match the results of the neural network experiment. Queries of 20 words are unlikely by users of an information retrieval system in normal use.

The co-ranking of the importance measure derived retrieved documents with the neural network retrieved set demonstrates that document representations are complementary, and used thus there is little more benefit from the very long queries.

10 CONCLUSION

Our fuzzy importance measure was able to reproduce 46% of the information retrieval behaviour of full inverse document

term weight vectors of 100 words, using just three query words. Thsi demonstrates the usefulness of the two major components of our approach. The first is the importance measure itself, and the second is the method of synthesising queries based on our notion of using the cumulative inverse document term weight.

Words do not express crisp concepts which can be matched against the concepts embodied by documents. Instead, 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.

This work has shown that we can use this notion to produce document representations which capture the contents of the documents in a collection in the context of the queries being used. We have also shown that the results from our importance measure can augment the complementary document representation produced by a different (learning) technique. We are continuing this work on a larger document collection in the legal domain.

11 ACKNOWLEDGEMENTS

This work was supported by the Australian Research Council. This paper continues previous work (Gedeon et al, 1995).

REFERENCES

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

Bustos, RA and Gedeon, TD, 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, TD, Kóczy, LT, Ngu, AHH, Bustos, RA and Shepherd, JA, 1995 "Learning Fuzzy Measure Functions for Information Retrieval," *Proc. Int. Joint Conf. of the 4th IEEE Int. Conf. Fuzzy Systems and 2nd Int. Fuzzy Engineering Symp. (FUZZ-IEEE/IFES'95) Workshop on Fuzzy Database Systems*, pp. 43-48, Yokohama.

Salton, G, 1989 *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley.

Turtle, H, 1995 "Text Retrieval in the Legal World," *Artificial Intelligence and the Law*, vol. 3, pp. 97-142.

Wong, PW, Gedeon, TD and Taggart, IJ, 1995 "An Improved Technique in Porosity Prediction: A Neural Network Approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, n. 4, pp. 971-980.