Discovering Indexing Parameters for Information Filtering

Anne H.H Ngu, Tom Gedeon and John Shepherd
School of Computer Science Engineering
University of New South Wales, P.O.Box 1, Kensington 2033, Australia
Tel.: +61 2 385 3970 Fax: +61 2 313 7987 E-mail: anne@cse.unsw.edu.au

Abstract

Widespread access to global networked information services is destined to become a reality over the next decade. Already, networks such as the Internet service user communities numbering in the millions. To minimise the amount of effort required by a non expert user to handle the large volume of information passing through the Internet, an automated and adaptive filtering tool is required. One of the major problems in designing automated and adaptive filter is to discover the right combination of indexing parameters to be used for computing document signatures. We propose to use a neural network technique to find the indexing parameters which are best for a particular reader and a particular information source.

Introduction

The next decade will witness an explosive growth in the quantity and availability of electronic information sources. The introduction of a large number of specialised television channels, already well underway in the United States, is one example of this phenomenon. The introduction of "personalised newspapers", derived by filtering the information from news agencies such as Reuters, Dow Jones and AAP is another example.

In the academic environment, information sources such as the Internet are already a valuable resource to researchers. For example, the "symbol grounding problem", an extremely important problem in artificial intelligence, was discussed extensively on the Internet in the AI/Natural-Language newsgroups before it appeared in the research literature. As another example, the first experimental results related to "cold fusion" were extensively reviewed on the Internet in the Physics newsgroups, and this provided an important focus for the debate on this controversial topic. The Internet news has copious other examples of newsgroups which discuss important research issues in their area on a daily basis. Another phenomenon, which highlights the importance of this new communication medium, is the recent advent of "electronic journals" which publish refereed research papers over the Internet.

There seems no reason to doubt that electronic information sources such as the Internet will play a vital role in the dissemination of scientific and technical results in the future. At present, however, these information sources are succeeding *despite* the difficulties encountered in using them effectively. The examples cited above of the success of the Internet arose primarily because a group of researchers managed to focus the discussion on a particular topic in a particular newsgroup. In general, however, the sheer volume of information flowing through the Internet makes it very difficult to locate relevant information. It must also be admitted that the Internet, along with its valuable information, carries a significant amount of "noise". The result of this is that users of the Internet news have to expend so much effort to locate relevant information that this effort becomes a barrier to using the news system at all. The development of information filtering tools is thus essential if we are to turn these high-volume information *sources* into useful information *resources*. This paper reports some initial work in building a Netnews filtering tool using an adaptive indexing scheme.

The Information Filtering Problem

Let us briefly define the general information filtering problem. An *information source* can be

viewed as a stream of (inter-related) documents. 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 (Ngu and Shepherd, 1993).

Despite the importance of the information filtering problem, both in the short-term for resources such as the Internet and in the longer-term for newly developing information sources, it is by no means yet completely solved, although some progress has been made. In December 1992, a special issue of the *Communications of the ACM* was dedicated to describing the latest research in this area. Several of the researchers cited (Fischer and Stevens, 1991, Foltz and Dumais, 1992, Goldberg et al., 1992) have developed systems to assist with information filtering in network news. Another, local example of a network-news filtering system is the grapeVINE system developed at the University of New South Wales (Brookes, 1991). The well-known SMART information retrieval system (Salton, 1980), has also been applied to the task of filtering network news. Commercial news filtering services, such as ClariNet also exist. All of these systems, however, require significant assistance from the user/news-provider in creating and/or maintaining the filters, specifying categories, building synonym lists, and so on.

We propose to develop automated and adaptive information filters: “automated” in the sense that they require minimal user intervention to perform all of the operations needed to extract relevant items; and “adaptive” in the sense that they keep track of the user's changing interest by observing the user's behaviour in dealing with news item. In particular, the system should mimic the complex cognitive tasks (understanding articles, retrieving the relevant ones and adapting to user changing interests and experiences) that occur in human information filtering. Monitoring and adapting the user's profile according to the user's current interests is the most obvious way of building an adaptive information filter. This is analogous to using the relevance feedback mechanism in a traditional information retrieval model. We believe that to be truly adaptive, the index generation for documents should be tailored to adapt to a specific user and a specific information source. Current filtering tools adopt a single way of indexing the news items regardless of the source of the news and the audiences that the news targets. This results in performing unnecessary indexing for a news reader who does not require such level of detail. In this paper, we focus on how to discover indexing parameters for Netnews using neural network techniques. This avoids wasting effort in generating unnecessary indexes.

Automatic Indexing of News Items

The aim of indexing news items is to 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 (Paice, 1990):

1. 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.

2. 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.

3. 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.

4. 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".

5. 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.

6. Structure of the news articles
   This includes header information and meta-language constructs used by the Internet community. Examples of these 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 system also assigns lower weights to keys in sections of news items which do not generally contribute to the content, such as author signatures and large sections of quoted text from other news items.

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 news reader 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 news item as judged by the news reader based on his/her 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).

Neural network discovery of indexing parameters

The six significance measures extracted from the news items need to be combined in some fashion to achieve a good overall measure. We wish to discover the relative significances of the various measures, and how they can be best combined.

The simplest means to discover a useful method of combination is to use the learning by example characteristics of neural networks. The advantage of neural network models over statistical techniques is that neural networks can implicitly learn the model the data should fit and fit it at the same time. This is in contrast to statistical methods which require some understanding of the model before data can be reliably fitted. The model we will use is the well known feed-forward neural network trained using the error back-propagation algorithm.

Description of initial study

Our intention in the overall study is use a neural network technique to find the indexing parameters which are best for a particular reader and a particular information source. Our aim in this initial study is to validate the usefulness of using neural networks in this fashion.

A suitable demonstration would be to train a neural network to perform a retrieval task, or to reproduce some component of a retrieval index.

A collection of 352 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 1,000. 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 relevant documents were available for a specific reader in the domain of this document collection. The words in the queries were reduced in number using the same process described above, removing the same words and so on. This produced a short list of 75 words only. These 75 words 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 would be 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. In this fashion we would reproduce a component of a retrieval index (the frequency-keyword measure used in vector retrieval methods), and at the same time we can derive some indication of the relative significance of the three input measures with respect to the output measure.

The network topology is shown in Figure 1. Note that not all connections are shown. The network was then trained normally using the error back-propagation algorithm. The network after training has formed some internal representation of the relative importances of the different significance measures.

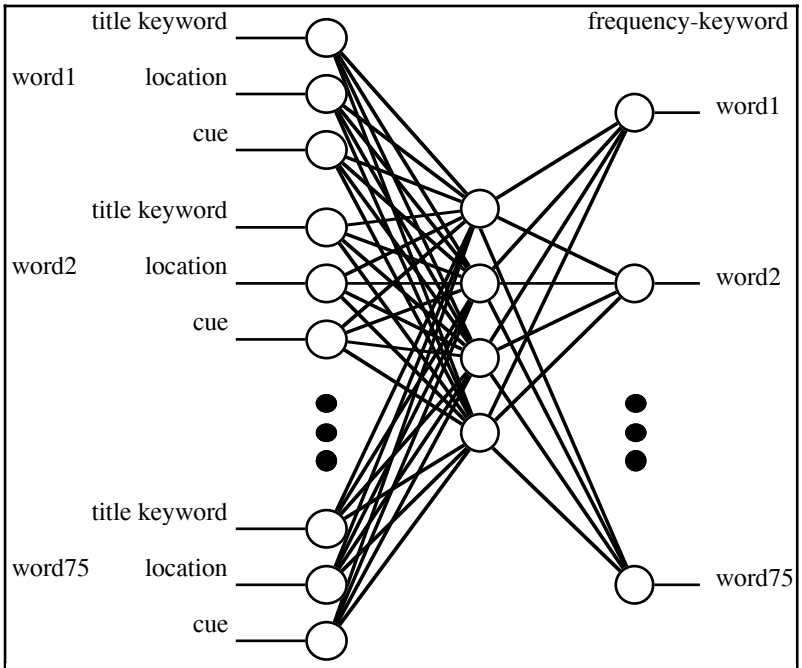

Figure 1. Neural network topology

The neural network can be tested in comparison to the calculated frequency-keyword measures. For simplicity, we have used vector retrieval using frequency-keyword measures. That is, a list of retrieved documents using the calculated frequency-keyword measures is compared to another list retrieved using the neural network outputs as approximations of the frequency-keyword measure. The percentage overlap in the top 10 retrieved documents gives some estimate of the usefulness of the trained neural network, which has implications to the relevance of the input

parameters used in training the 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 line of the documents were scanned to determine frequency of occurrence of the 75 words, and the resulting vector of 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. Similarly, for the cue method, a window of 5 words on both sides of all words is examined for the presence of cue words, the frequency of proximity to cue words is then normalised as above.

Thus, 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.

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 dimension space formed by the frequency-keyword values of the words to all of the other documents to determine which of these are similar. The top ten documents (using the traditional method) are compared to the top ten retrieved using the network's 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 other comparison that can be made is using the original queries which were used to provide 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 the 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 applied a simple analysis technique to the neural network to determine the relative importance of the three input measures (Wong, Gedeon and Taggart, 1994).

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

This is calculated by summing the absolute magnitude of the weights from a particular measure to the hidden units, and dividing by the sum of all the 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 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 commonly accepted that title keyword measures are very important.

Conclusion

In this paper we have shown how a feed-forward neural network can be used to discover the relative importance of indexing parameters which are best for a particular reader and a particular information source.

Results on all words show that 60% of the information contained in the frequency-keyword measure can be discovered from just the three other measures. This result demonstrates that some 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 the network was only trained using the whole document collection. We have also shown that the neural network can reproduce the full behaviour (100%) of the frequency-keyword vector retrieval using only the title, location and cue information, in a specialised subset of the document domain.

In the full scale experiment, we would of course use the frequency-keyword as one of the inputs and produce document indexes as outputs. In the final system, the neural network would be invisible to the user, and trained on observations of the user's behaviour, particularly using the saving of news items versus the 'killing' of news items as signifying the user's interests. The user should only notice a gradual decrease in boring items they see via their news program.

References

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

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

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

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.

Ngu, AHH and Shepherd, JA, "How to deal with 10,000 news articles per day: An Intelligent Assistant for Newsreading, *Proc. 1st Australia and New Zealand Internation Conf. on Intelligent Systems*, Perth, 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.