An Online Information Retrieval Systems by Means of Artificial Neural Networks

Marta E. Zorrilla, José Luis Crespo, Eduardo Mora

Department of Applied Mathematics and Computer Sciences, University of Cantabria.
Avda. de los Castros s/n 39005 Santander. SPAIN
{zorrillm, crespoj, morae}@unican.es

Abstract. The aim of this paper is to present a new alternative to the existing Information Retrieval System (IRS) techniques, which are briefly summarized and classified. An IRS prototype has been developed with a technique based on Artificial Neural Networks which are different from those normally used for this type of applications, that is, the self-organising networks (SOM). Two types of network (radial response and multilayer perceptron) are analyzed and tested. It is concluded that, in the case of a limited number of documents and terms, the most suitable solution seems to be the Multilayer Perceptron network. The results obtained with this prototype have been positive, making the possibility of applying this technique in real size cases a cause for a certain degree of optimism.

1 Introduction

At present, thanks to the technological advances of telecommunications and computer technology, information is becoming more and more accessible to the user. Internet for example, is one of the greatest sources of information ever known. With sources of information on such diverse topics, introduced by users with such different search criteria, new requirements arise in the areas of storage, searching and visualisation of information.

Such needs, however, have not arisen now – they have been felt since the sixties, Gerard Salton [10],[11] and his disciples took their first steps in this area with a view to improving the management of library catalogues.

Traditionally, this information is stored in Relational Data Base Systems, in which a document is represented by means of a series of structured fields, such as author, title, ISBN... and the searches are carried out through Boolean techniques. At present, technology enables library catalogues to be amplified and to incorporate summaries and even complete electronic versions of articles, books, etc in their own data bases. However, relational data base systems cannot support search techniques which enable texts to be found by the words used in them – this is called full-text search.

These characteristics are supported by the so-called Information Retrieval Systems (IRS), the case under study in this paper. In the following section, the various

R. Moreno-Díaz et al. (Eds.): EUROCAST 2001, LNCS 2178, pp. 343-351, 2001.

Information Retrieval Techniques will be briefly outlined; next, the neural network architectures for classification applications are presented; and finally, the characteristics of the proposed IRS, some results obtained and future lines of research will be specified.

2 Information Retrieval Techniques

Before entering into further detail, it should be borne in mind that, whatever technique is used, a preliminary stage is practically always required, consisting in obtaining what could be called the "pure text", independently of the original format of the document. This process is usually called "Filtering".

The techniques most widely used in Information Retrieval Systems are now classified and described. Figure 1 shows a basic classification.

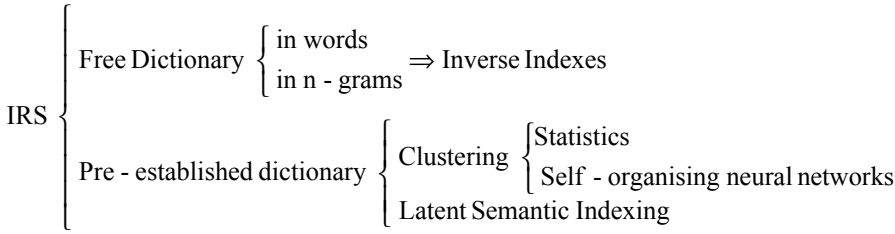

Fig. 1. Classification of Information Retrieval Systems

The most important aspect for the classification of this type of techniques is related to the set of words with which they can work. There are then two main blocks, depending on whether they use any word (free dictionary) or only a certain set of words (pre-established dictionary).

In the case of the free dictionary, there are also two possibilities: to act on words from the language being used (by words), which implies that they maintain, to a certain extent, their meaning; or to act on blocks of characters of a fixed length n, (by n-grams), with the result that their meaning in the text is, in some way, lost. In the latter case, the number of indexes may be quite small and, moreover, this system is independent of the language.

In both cases, the indexing technique most widely implanted commercially is that of "Inverse indexes". With respect to the "by words" system, this technique consists in considering the texts to be divided into documents, paragraphs and sentences. As texts are entered, a DICTIONARY of indexed (usable in searches) words is generated dynamically. New words are incorporated in it, updating the number of documents in which each word is found and the total number of appearances. At the same time, a LIST OF WORDS is formed, in which, for each appearance of a word, its number of appearance in the order of the sentence, of the sentence in the paragraph, of the paragraph in the document and the document code is stored. The process is similar in the case of n-grams, the function of the words being substituted by the n-grams.

The other possible form of grouping consists in limiting the set of words to be used (pre-established dictionary) to a specific set of a relatively small volume. There are basically two alternatives using this approach: Document Clustering and Latent Semantic Indexing.

In both methods, each document is represented by a vector in which each component assumes a numerical value indicating the relevance of a specific word in it. Hence, the dimension of these vectors will be equal to the number of words in the dictionary of terms (key words), which must be defined before the process begins.

The document search is carried out using a query text and consists in encoding it as a vector of all those used to represent documents, making a subsequent comparison between them.

In the case of Document Clustering, the next step consists in obtaining the types of documents closest to this vector and, of these, the documents most alike the query text. The agility of this method is based on comparing the search vector with types of documents rather than doing this with all the documents individually. This system requires a previous classification of documents. There are basically two alternatives for making this classification:

1. Clustering through statistics [4]:

This is based on the use of statistics for evaluating the degree of similarity between vectors; the most widely used are Euclidian Distance and the cosine similarity function. The calculation of the similarity between all of the vector pairs leads to the similarity matrix. The required classification is obtained from this. Traditionally the most efficient algorithms in cluster generation, hierarchic agglomerative methods, are the least efficient from a computational point of view, since they need to calculate the matrix of similarity between all the documents in the collection. The main advantage of these methods is that they are clearly defined and there is ample literature available about them both from a theoretical and the practical viewpoint.

2. Clustering through self-organising neural networks:

One alternative to clustering by means of statistics is the use of artificial competitive learning neural networks, as outlined below.

Latent Semantic Indexing [1] is a method which tries to solve the problem of lexical pairing between the terms of a search and the available documents, using conceptual indexes rather than individual words. The system is based on the reduction of the dimensions of the problem, by means of changes of base obtained using a truncated Singular Value Decomposition of the document-word frequency matrix.

3 Artificial Neural Networks for Classification

There are a large number of reference works which offer the results obtained using artificial self-organising neural networks applied to IR systems. In [12], Scholtes assesses the performance of the self-organising network method according to Kohonen's algorithm (SOM) applied to a small collection of documents. In [15], the use of the Kohonen method is compared with Fritzke's Growing Cell Structures method and in [6] Teuvo Kohonen, describes the WEBSOM system, developed by

him and his colleagues, based on the use of his method and applied to large collections of documents (over one million).

Neural networks are computational structures which process information in a distributed and parallel way, using a number of units called processors or neurons. These communicate with each other by sending signals through the connections joining them. Depending on their topology, learning mechanism, the type of association between the input and output information and the way in which this information is represented, the neural network will have one application or the other.

There are several proposed architectures oriented to classification tasks such as multilayer perceptrons, competitive networks (self-organising, for instance) and radial basis function networks.

In the competitive networks, each processor in the output layer corresponds to a class (in this case, a document). For each input point, there is only one processor in the output layer that has a non-null response, which indicates the class to which it belongs.

The radial response networks, unlike the competitive networks, offer a continuous output. All the processors may have a response, some higher than others. In order to find out the classification of each point, the categories are assigned according to the responses of the corresponding processors. For each input, the most probable category is that of the processor with the highest response.

These networks are quite similar to the perceptrons; the main difference lies in the activation function (radial basis function) and in the operations made at the connections.

A radial basis network with local activation function may require more neurons than a feedforward multilayer perceptron network with a tagsig or logsig activation function. This is because sigmoid neurons can have a non-zero output in a more extensive region of the input space, while radial neurals only respond to a small region.

4 Proposed Example

Most indexing techniques related to neural networks are based on the use of a pre-established dictionary (between 100 and 300 terms) and on the representation of documents as a vector of terms whose components respond, to a certain degree, to the importance of this term in the document and with respect to the collection to be indexed, obtaining a classification of these.

The indexing technique proposed here consists of an artificial neural network with suitable classification characteristics. So the aim is not to create clusters of documents but rather to identify each output neuron with a document from the collection. Along these lines, some prototypes have been developed using general purpose tools such as MATLAB, NODELIB and PDP++ [8],[9],[13] and the results are discussed and analysed in the context of the usual techniques.

In short, the model proposed consists in a neural network which has as its input layer the word from the dictionary in binary format, and at its output has as many

processors as the collection has documents (each of these processors represents one document).

From among the different types of neural networks proposed for classification, our tests focus on the use of neural networks with a radial basis functions (RBFs) and on multilayer perceptrons (MLPs).

4.1 Radial Basis Networks

Radial response networks are normally networks with one single layer, although it is possible to build arbitrary networks. The radial function is of the Gaussian type.
These networks share a certain likeness with perceptrons, the main difference being in the activation function (radial base function) and in the operations performed at the connections (distance instead of scalar product).

4.2 Multilayer Perceptrons

Basically, in this type of network, processors are grouped together in layers and these are generally ordered in such a way that the processors of each layer take the signal from the previous one (forward feeding). Each processor calculates a function from the scalar product of its weights and its inputs. This function is normally the logistic or the hyperbolic tangent function.

The training in these network models is supervised. The adjustment is an important point in the implementation of the neural networks. Originally, a gradient descent in the error function was proposed as the adjustment technique.

The learning methods must be computationally efficient, and must converge, that is, approach a minimum reliably, and this must be the overall minimum, avoiding any local minimum which the objective function may have. Algorithms based on derivatives are generally used as they are the fastest. These include the gradient descent (based on the fact that the variation in each step taken to reduce the error is more effective if it is in the direction of the gradient); the conjugate gradient descent, which attempts to solve the limited efficiency of the gradient descent method by advancing in orthogonal directions; and, the Newton method, based on an approximation up to the second derivative of the error function [14]. Given that the calculation of second derivatives is more difficult to obtain, there are variants of this algorithm in which an approximation of the Hessian matrix is performed. These are: the Gauss-Newton algorithm, the Levenberg-Marquardt algorithm and the Quasi-Newton algorithm, by means of the approximation of the inverse matrix to the Hessian matrix. In these methods, a one-dimensional minimization is required at each step all along the direction to be taken. Several algorithms can be applied for this minimization, such as those of Newton, Brent or Charalambous [2],[14].

4.3 Data Used

The collection of documents used for this work corresponds to the articles of Spanish Civil Law and, in particular, to the articles under Section Headings I, II and III. For the sake of agility, and to solve the problem of the search and storage of documents, each article was considered a document in ASCII format. In total, 140 documents were considered.

The dictionary of the system is made up of a subset of words, taken from the COES[1] dictionary, verifying that between 60-70% were words found in the collection of documents. Given that there is no preference for the coding, each item of vocabulary is encoded as the binary number corresponding to its alphabetical position in the dictinary.

4.4 Discussion of Results

Firstly, the results obtained in the tests performed with RBF type networks will be outlined and next the results achieved with MLP type networks will be presented.

Results with Radial Basis Function Networks

The solution offered by hidden radial layer and linear output layer networks [5] is ineffective, since it creates as many neurons in the hidden layer as there are different words in the dictionary. Therefore the number of parameters required with respect to the inverse index technique is higher.

In the case of a multilayer perceptron with a hidden layer and radial response, it was observed that the geometry of the error function (the entropy and square functions were used) tends too readily towards degenerate minimums with the data handled.

Results with Multilayer Perceptron

The results presented here correspond to an example in which a dictionary of 14 words and a collection of 10 documents were used. The tests employed an architecture of 10 input neurons, 5 neurons in the hidden layer and 10 neurons in the output layer; the square and entropy error functions and various learning methods were used such as stochastic gradient with sample permutation, scaled conjugate gradient, Quasi-Newton with one-dimensional linear minimization using the Brent algorithm and with fixed rate minimization.

The most significant outcomes of the tests performed are as follows:

- A network with 5 processors in the hidden layer is capable of learning the sample without error.
- The method used in the optimization may be of crucial importance.

[1] **COES**. Spanish dictionary over 53.000 terms developed by Santiago Rodríguez and Jesús Carretero who work in the Architecture and Computer Systems Department of the Polytechnique University of Madrid.

- The same method of learning used in different programs [8],[9],[13] does not offer the same results.
- The error function does not seem to be of great importance.

In order to access and act directly on the parameters of the optimization process, the multilayer perceptron was programmed with the architecture shown in Figure 2. Two learning methods were used, the conjugate gradient method and the Quasi-Newton method with one-dimensional linear minimization using the Golden method (minimum by intervals) and the Brent method (parabolic interpolation). It was tested with two error functions, the square error and the entropy functions.

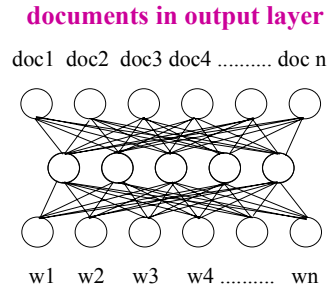

Fig. 2. Graph Representation of the programmed network

From the tests performed, it can be concluded that:

- The final results may often be considered as local minima.
- In a 10 –5 –10 architecture network, the mean error is still high (> 0,2) although it approximates perfectly several of the patterns.
- The higher the number of hidden processors, the more satisfactory the results, optimal results being obtained with 10 processors (see Graph 1).
- The adjustment is most efficient with the Quasi-Newton method and one-dimensional minimization using the parabolic interpolation method.
- Learning using the conjugate gradient method is slower and offers worse results (mean error 0.5)
- The use of the entropy error function does not offer better results than the mean square error function.
- The results offered by the network are not sensitive to variations in the coding of words (order by frequency rather than by alphabetical order).

Figure 3 presents in graph form the evolution of the mean square error with respect to the number of neurons present in the hidden layer.

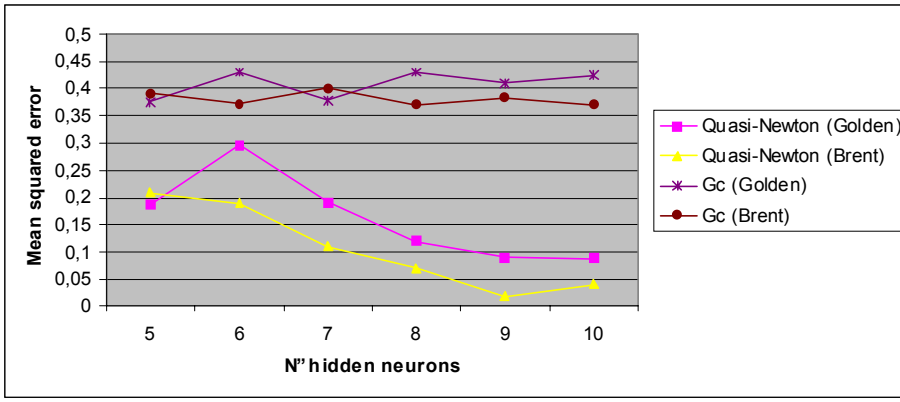

Fig. 3. Behavior of error with respect to number of hidden neurons

5 Conclusions

This paper presents a classification of the main indexing techniques in use today in the field of IRS and describes them briefly. With a view to incorporating a new option into this classification, an IRS prototype has been developed with a technique which the authors consider to be unused at present. It is based on non-self-organising Artificial Neural Networks.

Various architectures have been analyzed for use, concluding that the most suitable one would seem to be the MLP. In this area, the results obtained show that the solution proposed is valid in practical cases of limited dimensions; that is, few documents as opposed to few terms. It has also been observed that the learning method used may be vitally important to the successful operating of the network.

Finally, the tests presented here lead to some optimism as to the possible use of the prototype in real size cases.

References

1. Berry, M.W; Dumais, S.T.; O'Brien, G.W. *Using Linear Algebra for Intelligent Information Retrieval*. Computer Science Department. University of Tennessee. (1994).
2. Crespo, J.L. Redes Neuronales Artificiales para Ingenieros. (1996). ISBN: 84-605-5020-6.
3. Croft, W.B. *A model of cluster searching based on classification*. Information Systems, Vol. 5, pp. 189-195.(1980) .
4. Cutting, D.R.; Karger, D. R.; Pedersen, J. O.; Tukey, J.W. *Scatter/Gather: a Cluster-based Apporach to Browsing Large Document Collections.* ACM SIGIR Conference, pp 318-329. (1992).
5. Demuth, H.; Beale, M. Neural Network Toolboox for Use with MATLAB. User's Guide. The Math Works Inc.

6. Kohonen, T. Self-Organization of Very Large Document Collections: State of Art. In Proccedings of ICNN'98, pp. 65-74.
7. Lagus, K. *Generalizability of the WEBSOM Method to Document Collections of Various Types*. European Congress Intelligent Techniques and Soft. Computing (EUFIT'98), Vol. 1, pp. 210-214.
8. NODELIB. Gary Willian Flakes
9. PDP++ Software, version 2.0. Chadley K. Dawson, Randall C. O'Reilly, J. L. McClelland
10. Salton, G.; McGill, M. Introduction to Modern Information Retrieval. McGraw-Hill. (1983)
11. Salton, G.; Yang, C.S.; Wong, A.. A Vector Space Model for Automatic Indexing. Departament of Computer Science, Cornell University. TR 14-218. (1974).
12. Scholtes, J.C. Neural Networks in Natural Language Processing and Information Retrieval. Ph.D. Thesis, Universiteit van Amsterdam. (1993).
13. SNNS, version 4.1, A. Zell et al., University of Stuttgart
14. Vettering, W.; Press, W.;Flannery, B.;Teukolsky, S. Numerical Recipes in C. Cambridge University Press. (1988)
15. Zavrel, J. An Experimental Study of Clustering and Browsing of Document Collections with Neural Networks. Ph.D. Thesis, Universiteit van Amsterdam. (1995).