Comparing Neural and Probabilistic Relevance Feedback in an Interactive Information Retrieval System

Fabio Crestani*
Department of Computing Science
University of Glasgow
Scotland - UK

Abstract
This paper presents the results of an experimental investigation into the use of Neural Networks for implementing Relevance Feedback in an interactive Information Retrieval System. The most advance Relevance Feedback technique used in operative Interactive Information Retrieval systems, Probabilistic Relevance Feedback, is compared with a Neural Networks based technique. The latest uses the learning and generalisation capabilities of a 3-layer feedforward Neural Network with the Backpropagation learning procedure to distinguish between relevant and non-relevant documents. A comparative evaluation between the two techniques is performed using an advance Information Retrieval System, a Neural Network simulator, and an IR test document collection. The results are reported and explained from an Information Retrieval point of view.

1 Information Retrieval

Information Retrieval (IR) is a science that aims to store and allow fast access to a large amount of information. This information can be of any kind: textual, visual, or auditory. An Information Retrieval System (IRS) is a computing tool which stores this information to be retrieved for future use. Most actual IR systems store and enable the retrieval of only textual information or documents. To give a clue to the size of the task, it must be noticed that often the collections of documents an IRS has to deal with contain several thousands or even millions of documents.

A user accesses the IRS by submitting a query, the IRS then tries to retrieve all documents that “satisfy” the query. As opposed to database systems, and IRS does not provide an exact answer but produce a ranking of documents that appear to contain some relevant information. Queries and documents are usually expressed in natural language and to be processed by the IRS they are passed through a query and a document processors which breaks them into their constituents words. Non-content-bearing words (“the”, “but”, “and”, etc.) are discarded, and suffixes are removed, so that what remains to represent query and documents are lists of terms that can be compared using some similarity evaluation algorithms. Good IR systems typically rank the matched documents so that those most likely to be relevant (those with the higher similarity with the query) are presented to the user first. Some retrieved documents will be relevant (with varying degree of relevancy) and some will, unfortunately be irrelevant. The user appraises those ones that he considers relevant and feeds them through a process called Relevance Feedback (RF) which modifies the original query to produce a new improved query and as a consequence a new ranking of documents. If the IR process is interactive this will go on until the user is happy of the resulting list of documents. An example of the IR process is depicted in Fig.1.

In recent years big efforts have been devoted to the attempt to improve the performance of IR systems and research has explored many different directions trying to use with profits results achieved in other areas. In this paper we will investigate the possibility of using Neural Networks (NN) in IR and in particular we will concentrate on the RF process. RF is not always present in operative IR systems, but that has been widely recognised to improve considerably IR systems’ performance. The purpose of this work is therefore to investigate the possibility of using NN to implement a Neural RF. Previous research (see [1] for a survey) shows that, though giving encouraging results, NN cannot be effectively used in IR at the current state of the NN technology. The scale of real IR applications, were hundreds

*On leave from Dipartimento di Elettronica ed Informatica, Universita’ di Padova, Italy
of thousands of documents are at stake, makes it impossible to use NN in an effective way, unless we make use of very poor document and query representations. However, recent results [2] proves that it may be still possible to use NN in IR for very specific tasks were the numbers of patterns involved (and therefore the training) are reduced to a manageable size. In this paper we show that it is indeed possible to use NN to implement a RF device. However, the results here reported also show that a Neural RF device is not effective in real world’s IR applications, where it performs worse than traditional RF techniques.

2 Relevance feedback and query adaptation

Relevance Feedback is a technique that allows a user to express in a better way his information requirement by adapting his original query formulation with further information provided by indicating some relevant documents. When a document is marked as relevant the RF device analyses the document text, picking out terms that are statistically significant to the document, and adds these terms to the query. RF is a very good technique of specifying an information requirement, because it releases the user from the burden of having to think up lots of terms for the query. Instead the user deals with the ideas and concepts contained in the documents. It also fits in well with the known human trait of “I don’t know what I want, but I’ll know it when I see it”. Obviously the user cannot mark documents as relevant until some are retrieved, so the first search has to be initiated by a query. The IRS will return a list of ordered documents covering a range of topics, but probably at least one document in the list will cover, or come close to covering, the user’s interest. The user will mark the document(s) as relevant and starts the RF process performing another search. If RF performs well the next list should be closer to the user’s requirement. A schematic view of the RF process is depicted on the right side of Fig.1.

We may also think at a RF device as a filter that receives as input a query and a set of relevant documents (or considered so by the user) and that gives as output a modified or adapted query. This process of query adaptation is supposed to alter the original user formulated query to take into consideration the information provided by features of relevant documents. More formally the input of the RF device has the following form:

\[(q_i, d_1^i, d_2^i, d_3^i, \ldots d_k^i)\]

where \(q_i\) is the representation of the query \(i\), and \(d_j^i\) is a relevant document representation. The set is composed of only a subset (\(l\) documents) of all the documents (\(k\) documents, with \(k > l\)) that are relevant to that particular query. The aim of the RF is to enable the retrieval of the other (\(l - k\)) relevant documents.

In IR we can use different RF techniques, depending on the IR model being preferred. In the following Section we will illustrate a technique called Probabilistic RF, which is based on the Probabilistic IR model.

2.1 Probabilistic relevance feedback

Probabilistic Relevance Feedback (PRF) is one of the most advance technique for performing RF in operative IR systems. For a better and more complete explanation of the mathematics involved see [3]. Briefly, the technique consists in adding a few other terms to those already present in the original query. The terms added are chosen by
taking the first $m$ terms in a list where all the terms present in relevant documents are ranked according to the following weighting function:

$$w_i = \log \frac{r_i (N - n_i - R + r_i)}{(R - r_i) (n_i - r_i)}$$

where: $N$ is the number of documents in the collection, $n_i$ is the number of documents with an occurrence of term $i$, $R$ is the number of relevant documents pointed out by the user, and $r_i$ is the number of relevant documents pointed out by the user with an occurrence of term $i$. Essentially what this rather complex function does is compare the frequency of occurrence of a term in the documents the user marked as relevant with the term frequency of occurrence in the whole document collection. So if a term occurs much more frequently in the documents marked as relevant than in the whole document collection it will be assigned a high weight. In the experiments reported in this paper the number of terms added to the original query has been experimentally set to 10.

4 Experimental Settings and Evaluation

In order to perform an experimental analysis these tools are necessary: a document collection with relevance assessment, a probabilistic IRS with PRF, and a NN or a NN simulator on which implement a NRF.

The document collection chosen for the investigation is the ASLIB Cranfield test collection. This collection was built up with considerable effort in the 60s as the testbed for the ASLIB-Cranfield research project, aimed at studying the “factors determining the performance of indexing systems”. They produced two test collections of documents about...
aeronautics, comprehensive of documents, requests and relative relevance judgements. For a full description of the collections see [5]. In the investigation reported in this paper the 200 document collection, with 42 requests and relevance judgements, has been used. It is of course understood that this limits the generality of the results obtained. However, the main purpose of this investigation is to demonstrate the feasibility of the proposed approach. There are still many open issues in the application of NN to IR, and the problem of scaling the results is one of the major ones. Nevertheless, it must be noticed that the dimension of the data set used in the these experiments is one of largest used in any application of NN techniques to IR.

The choice of IR systems with PRF was quite limited. In fact, there are not many operative IR systems with such an advance feature. One of the best systems is News Retrieval Tool (NRT) [6], developed at the University of Glasgow for the Financial Times. NRT implements PRF as specified in Section 2.1. For the investigations reported in this paper a NN simulator, PlaNet 5.6 [7], running on a fast conventional computer has been used. The architecture and the algorithm’s details of the implementation are reported in [2].

For the evaluation, the main retrieval effectiveness measures used in IR are Recall and Precision. These two measures have been used in evaluating and comparing the performance of PRF and NRF. Recall is the proportion of all documents in the collections that are relevant to a query and that are actually retrieved. Precision is the proportion of the retrieved set of documents that is relevant to the query. In order to give a measure of the generalisation performance of NRF and PRF, Recall and Precision have been evaluated with different cardinality of the set of training examples. If some learning of the domain knowledge has taken place, and if the NN can generalise it, then an improvement in the retrieval performance has to be expected.

5 Experimental results

The results here reported refer to the PRF implemented in NRT and to a NRF implemented on PlaNet using a 3–layer feedforward NN with BP learning. The numbers of input nodes was set to 195, the number of hidden nodes to 100, and the number of output nodes to 1142. The number of input and output nodes correspond respectively to the number of terms used in all the queries and to the number of terms used in the Cranfield collection.

Fig.3 shows graphically how the performance of the NRF and PRF increases when the RF device is fed with increasing amount of relevant documents. This shows that the both PRF and NRF act like a pattern recognition device and the more information they receives the more they can discriminate between patterns of relevant and not relevant documents. The performance has been evaluated averaging over all the set of queries in the relevance assessment at different values of the number of relevant documents given as feedback (training examples). The graph shows that the NRF increases more rapidly in performance than the PRF. This is due to the better characteristics of non linear discrimination of NRF, that enables it to separate better the two patterns. However, the performance of PRF are better at any level of training and especially at lower levels. This makes PRF more useful in real world applications where the percentage of relevant documents used in the training over the total number of relevant document is usually very low.

Tab.1 reports performance figures for the case of 10% training, i.e. when 10% of the number of relevant documents totally present in the document collection are fed to the RF device for training. It shows numerically how PRF
Recall | Precision NRF | Precision PRF
--- | --- | ---
0.10 | 0.245 | 0.440
0.20 | 0.234 | 0.363
0.30 | 0.225 | 0.332
0.40 | 0.201 | 0.301
0.50 | 0.179 | 0.275
0.60 | 0.160 | 0.259
0.70 | 0.132 | 0.239
0.80 | 0.111 | 0.221
0.90 | 0.051 | 0.210
1.00 | 0.022 | 0.130

Table 1: Comparison of the performance of Neural and Probabilistic RF for 10% training

outperforms NRF in a realistic situation where the number of training examples, i.e. the number of relevant documents fed to the RF device is low compared to the total number of relevant documents present in the collection. This confirms what Fig.3 already shows, and gives a stronger argument for preferring PRF to NRF in IR.

6 Conclusions

The results of this investigation, briefly summarised in this paper, demonstrate that a RF device based on NN (NRF) acts in a similar way to classical IR RF techniques. However, a comparison with one of the most advance RF techniques presently in used in IR (PRF) shows that for low levels of training, which is the most common case in IR, NRF does not perform as well as PRF. Though it is necessary to proceed with further research using different NN architectures, and different learning algorithms, we think that NN are currently not suitable for use in IR applications where the number training examples is usually very little and the number of nodes and connections is extremely high.

References


