Application of Neural Network in Information Retrieval

Ajay Kumar Verma¹, Lokesh Kumar Shrivastav², Ramakant Prasad³, Narender Kumar³, Ankit Rajpal⁴, Surender Kumar Verma⁵

¹School of Computational & Integrative Sciences, JNU, New Delhi, India.
²Department of Computer Science, Atma Ram Sanatan Dharma College, University of Delhi, Delhi, India.
³Department of Mathematics, Gargi College, University of Delhi, New Delhi, India.
⁴Department of Computer Science, Deen Dayal Upadhyaya College, University of Delhi, Delhi, India.
⁵Department of Computer Science, Ramlal Anand College, University of Delhi, Delhi, India.

ajayverma81@gmail.com, lokesh.arsd@gmail.com, ramakant_iitd@yahoo.co.uk, nkumariitd@gmail.com, ankit30sep@gmail.com, er.surender08@gmail.com

ABSTRACT

This paper provides a survey of the application of neural network in information retrieval systems. Various soft computing techniques are being used in information retrieval (IR). Neural network, a soft computing technique, is studied to resolve the problems of extracting the keywords and retrieving the relevant from large database. Literature shows possible applications of variant of neural networks in IR to serve the purpose of retrieving the right documents. The theoretical implementation of neural network in vector space model of IR shows importance of neural network in IR.

Keywords

1. INTRODUCTION

For last few decades, the problem of information storage and retrieval has become a big and attractive but challenging issue. With developments in hardware and software technologies vast amounts of information can be stored, accurate and speedy access of information is also possible but retrieving relevant information, especially in textual databases is still very difficult. This has increased the need for automated information retrieval for extremely large document collections. In fact a search engine is also a form of an information retrieval system. However in spite of efficient search technologies being developed, the basic problem of getting relevant documents and that too in the top ranked documents is still unsolved. Different components of soft computing are being used in information retrieval. These techniques have changed the criteria of hard computing used for problems of information retrieval. Since the process of retrieving relevant information are full of uncertainty and imprecision. Soft computing is drastically used in every field attached with IR. The constituents of soft computing share common features and they are considered complementary instead of competitive [1, 2]. Artificial neural network is the machinery for learning and adaption [3]. It is used to modify stored information in response to new inputs from the user. Neural network provides capability to associatively recall information despite noise or missing pieces in the input. Information can be categorized by neural networks by their associative patterns. If there is no exact data against the query, neural network can be used to retrieve “nearest neighbor” data.

The paper consists of 6 sections. Section 2 discusses information retrieval system. Mathematical model of neural network is described in section 3. Section 4 presents a survey of the various applications of neural network in IR. In section 5, neural network in vector space model of IR presents the understanding of neural network in IR. Section 6 is conclusion.

2. INFORMATION RETRIEVAL SYSTEM

Information retrieval is devoted to finding relevant documents from storage in response of a query provided by a user [4]. Information Retrieval System (IRS) retrieves information in response to user queries. It stores a collection of information (generally called documents), accepts a user query, searches the document collection and returns ranked set of relevant documents. An IRS, shown in Fig 1, consists of three basic components: documentary database, query subsystem and matching mechanism [4].

The documentary database: This component stores documents, along with the representation of their information contents. It is
associated with the indexer module, which automatically generates a representation of each document by extracting the document contents. Textual document representation is typically based on index terms (single terms or phrases), which are content identifier of documents.

**The Query Subsystem:** It allows user to specify their information needs and presents the relevant documents retrieved by the system to them. To do this it requires a query language that allows user to formulate query and a procedure to select the relevant documents.

**The matching mechanism:** It evaluates the degree to which documents are relevant to user query giving a retrieval status value (RSV) for each document. The relevant documents are ranked on the basis of this value.

Most of the IRS’s use keywords to retrieve documents. The system first extracts keywords from documents and then assigns weights to the keywords by using different approaches. Efficiency of such systems is based on how efficiently they solve two major problems: one is how to extract keywords precisely and other is how to decide weight of each keyword.

To measure effectiveness of IRS, Precision and recall are used. Precision is the ratio of the number of relevant documents retrieved to the total number retrieved. Precision provides an indication of the quality of the retrieved set. However, this does not consider the total number of relevant documents. Recall considers the total number of relevant documents. It is the ratio of relevant documents retrieved to the total number of documents in the collection that are believed to be relevant.

### 3. NEURAL NETWORK

Neural network (NN) is a simplified model of the biological neuron system. It is a massively parallel distributed processing system consisting of highly interconnected neural computing elements which are intended to interact with the object of the real world in the same way as biological systems do [5].

Neural networks are one of the important components in artificial intelligence. Neural network is also called connectionist model, neural net, collectve model, parallel distributed processing model, and artificial neural network. Various definitions of neural network are given by researchers.

#### 3.1 Mathematical Model of Neural Network

“A biological neuron receives all inputs through the dendrites, sums them and produces an output if the sum is greater than a threshold value. The inputs are passed on to the cell body through the synapse which may accelerate or retard an arriving input” [5]. Based on this theory a mathematical model is constructed in Fig. 2.

Referring to Fig 2, let \( x_1, x_2, x_3, \ldots, x_n \) are the inputs to the artificial neuron and \( w_1, w_2, w_3, \ldots, w_n \) are weights attached to the inputs links. Here acceleration of the inputs is modeled by the weights. An effective synapse will have a larger weight if it transmits a stronger signal.

![Fig. 2: Neural Network Model](image)

Thus weights depends on input and hence total input received by the artificial neuron is given by

\[
I = \sum_{i=1}^{n} w_i x_i
\]  

(1)

The sum \( I \) is passed on to a non linear filter \( \lambda \) called activation function or transfer function, which release the output \( y \).

i.e.

\[
y = \lambda(I) \]  

(2)

There are many function used as activation functions. The thresholding function is commonly used. This function is defined by a threshold \( \theta \), the value of \( I \) is greater than \( \theta \), then the output is 1 else
it is 0. i.e. \( \lambda \) is called step function known as Heaviside function and is such that

and

\[
\lambda_{(I)} = \begin{cases} 
1 & \text{if } I \geq \theta \\
0 & \text{if } I < \theta 
\end{cases}
\] (3)

There are other activation functions like Signum function, Sigmoid function etc.

4. APPLICATION OF NEURAL NETWORK MODELS in IR: SURVEY

The application of connectionist models to Information Retrieval is not a recent phenomenon. A number of researches have adopted neural network as a technique to retrieve documents and various variant of neural network are being applied on IR. Neural networks seem to show good results with conventional retrieval models such as the vector space model [6] and the probabilistic model. As noted by Doszkocs et al. [7] several information processing approaches includes models of neural network like artificial neural networks, spreading activation models, associative networks, and parallel distributed processing. Connectionist models are “self-processing” in that no external program operates on the network: the network literally processes itself, with “intelligent behavior” emerging from the local interactions that occur concurrently between the numerous network nodes through their synaptic connections. Belew proposed connectionist model which is probably the one of earliest model adopted in IR [8]. In adaptive information retrieval (AIR), Belew developed a neural network of three layers: one for authors, another for Index terms, and last layer for documents. Belew’s model requires the relevant feedback from its users and thus changes its representation of authors, index terms, and documents. Kohonen represented high dimensional hierarchical relations to the two dimensional space by using the self organizing feature of neuronal network [9]. Belew and Rose proposed hybrid connectionist and symbolic system called SCALIR. SCALIR finds relevant documents by using analogical reasoning [10]. Wilkinson and Hingston [11] implemented vector space model by neural network for document retrieval. This network has three layers: queries, terms, and documents. Kohonen’s model was applied was applied to construct a self-organizing, visual representation of the semantic relationships between input documents [12]. MacLeod and Robertson exploited neural algorithm for document clustering [13]. Chen and et.al proposed model based on variant of Hopfield network [14] to develop a network of related keywords. An asymmetric similarity function is used to produce thesauri for different domain-specific databases. These thesauri help in concept exploration and query refinement. Similarly Hopfield network is implemented for concept clustering [15]. Hatano and et.al used Kohonen’s self-organizing map for clustering and retrieval of text and video data [16]. They applied vector space model and DCT (Discrete Cosine Transform) image coding to retrieve keywords. The Hopfield network has been adapted for concept analysis by chung [17]. The network is an asymmetric, continuous network in which the neurons are updated synchronously.

5. Vector Space Model with Neutral Networks

Neural network has been used to implement the vector space model [4]. The layered representation of NN using VSM model is shown in Fig 3. The input layer represents Query Layer and its nodes are set as queries. Hidden layer represents Term layer and its nodes denote terms, and output layer represents Document Layer and its nodes denote documents. The links between the nodes are defined as query-term links and document-term links. A link between a query and a term indicates the term appears in the query. The weight of the link is computed by \( tf-idf \) technique for the term [4]. Document-term links appears for each term that occurs in a given document. Again a \( tf-idf \) technique can be used for weighting the edges.

A feed-forward network works by activating a given node. A node is active when its output exceeds a given threshold. To begin, a query...
node is activated by setting its output value to one. All of its links are activated and subsequently new input weights for the TERM nodes are obtained. The input received by a term is \( \sum_{i=1}^{n} w_i x_i \) where, \( w_i \) is weight computed by tf-idf for the link, and \( x_i \) is the output which is taken 1 for respective query node. Since there is only one node activated in QUERY layer. So total input received by a term in TERM layer is equal to (tf-idf)(1) which becomes input for the next phase. i.e. \( x = tf-idf(1) \). In the next phase, the TERM nodes and all of the links connected with \( n \) terms are activated. The DOCUMENT node receives the sum of all the weights associated with each term in the document multiplied by respective output of term node. The input received by a document node is \( \sum_{i=1}^{n} w_i x_i \). Here, \( w_i \) is \( (tf_i)(idf)_j \), the weight of document-term weight and \( x_i \) is output of respective term node (i.e. the weight of query term weight). Therefore, input received by a document is product of query-term weight and document-term weight. This input is equivalent to a simple dot product.

6. CONCLUSION

We have presented a survey of the different aspects of neural network models in IR. These models are studied widely to resolve the problems of extracting relevant information from large database. It is interesting to note that the VSM model is implemented using neural network which makes the retrieval more efficient. As from the survey we conclude here that there is a wide scope of neural network models in information retrieval viz. keywords and documents.

REFERENCES