A Hybrid Neuro-Fuzzy System based Ranking Function

4.1 INTRODUCTION

Retrieval of reliable relevant information is the major concern in MIR. Among all components that influence the performance of retrieval system, ranking function is the most important component. IR system uses ranking function to compute the relevance score between the user query and the retrieved documents. Then, retrieved documents are arranged in the decreasing order of their relevance score (Gupta et al., 2015).

The previous work improves the performance of the retrieval system by providing related results, resulting in achieving the diversification of the user query in the medical domain. The performance can further be improved by using suitable ranking function to arrange the retrieved documents. A Hybrid neuro-fuzzy based Ranking Function (HRF) is proposed which can manage the uncertainty and vagueness present in the natural language text and it is also a highly robust uncertainty based model which is able to handle real life systems.

Hybrid neuro-fuzzy system is a neural network which is practically equal to FIS. Neuro-fuzzy system can incorporate domain knowledge. The proposed ranking function considers weight of document and query with respect to keyword as input features and gives relevance score between document and query as output. Analyses are performed on OHSUMED and PMC benchmark medical document corpus by utilizing 15 experimental queries.

The intuition behind using hybrid neuro-fuzzy system is that it consolidates the parallel calculation and learning capabilities of neural networks with human-like clarification and representation capacities of fuzzy system. The hybrid neuro-fuzzy based ranking function could become more transparent supplemented with learning and generalizing competency. It permits combining learning and logic based model with the VSM. Therefore, the resulting ranking model has learning capability, transparency and flexibility and performance of the VSM.
4.2 CONTRIBUTIONS

The contributions made in this work include the following:

- Designing a highly robust uncertainty based ranking function which has the capacity to capture the intrinsic features of natural language text.
- Joining the learning and generalization capacities of neural network with human-like information representation and clarification capacities of fuzzy system and thereby repaying the downsides of fuzzy system.
- Defining fuzzy rule base by using the existing domain knowledge of IR.

4.3 PROPOSED MODEL

In general, the structural representation of neuro-fuzzy model is same as the multi-layer neural network. It contains three hidden layers along with an input and output layer. The hidden layers represent the fuzzy membership functions and fuzzy rule base (Vieira et al., 2004).

Figure 4.1 depicts the proposed hybrid ranking function model. This hybrid neuro-fuzzy model is used to calculate relevance score between the user query and retrieved documents. As shown in Figure 4.1, proposed model takes the values of $Wt(D,K_i)$ and $Wt(Q,K_i)$ as input and outputs $RSCO$. $Wt(D,K_i)$ denotes the weight of document with respect to the keyword $K_i$, $Wt(Q,K_i)$ represents the weight of query with respect to the keyword $K_i$ and $RSCO$ represents the relevance score between document and query.

The five layers of proposed HRF model are:

1. Input
2. Fuzzification
3. Fuzzy rule
4. Output membership
5. Defuzzification
4.3.1 DESIGN PRINCIPLES

The hybrid neuro-fuzzy model which is used for designing the ranking function is constructed by using the following principles:

- The number of neurons in-
  a. Input layer = Number of input variables.
  b. Fuzzification layer = Number of fuzzy sets in the antecedent part of fuzzy rules in the rule base.
  c. Fuzzy rule layer = Number of rules in the fuzzy rule base.
  d. Output membership layer = Number of fuzzy sets in the consequent part of fuzzy rules in the rule base.
• The only neuron present in the defuzzification layer gives a single output of the neuro-fuzzy system which is the relevance score according to the proposed ranking function model.

4.3.2 REPRESENTATION OF DOCUMENTS

The retrieval process takes collection of medical documents as input. The input documents are pre-processed and annotated keywords are extracted from the input documents by adopting the procedure which has been explained in the previous chapter. As a result of annotated keyword generation, the set of related keywords \((K_p)\) for each individual input document is obtained. VSM is used to represent the input document collection into the form suitable for retrieval system. Documents are represented as a vector in \(x\)-dimensional space where, \(x\) is the number of unique keywords extracted from document corpus. Representation of documents as a vector in VSM is shown in Figure 4.2 in which \(n\) represents the number of documents in the corpus.

**Figure 4.2: VSM representation of documents**

Among the two input features, one is the weight of document with respect to the keyword which is represented as \(W_t(D,K_i)\). In order to compute this value, a table which consists of the information related to the occurrence of unique keywords in a document
has been maintained. Table 4.1 presents the sample keyword occurrence table in which the values 1 or 0 indicate the known value of the keyword and the value given in brackets indicate the frequency of the keyword in the document. Known value 1 indicates that this particular keyword is a related keyword for the document whereas value 0 indicates that the keyword is not related to the document. The related keyword may or may not occur in the document. For example, in Figure 3.3, for the given input document, the term 'kidney' is extracted as one of the related keyword. But it does not occur in the given document. This is because, in the given input document the term 'lupus nephritis' refers to swelling in kidney. Thus, the term 'kidney' is obtained as one of the related keyword for the given input document. For such keywords, the table entries will be 1(0) in which 0 indicates its frequency.

Table 4.1: Keyword occurrence in documents

<table>
<thead>
<tr>
<th>Doc Id</th>
<th>Unique Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K1</td>
</tr>
<tr>
<td>D1</td>
<td>1(15)</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>1(12)</td>
</tr>
<tr>
<td>D4</td>
<td>0</td>
</tr>
<tr>
<td>D5</td>
<td>1(0)</td>
</tr>
<tr>
<td>D6</td>
<td>0</td>
</tr>
<tr>
<td>D7</td>
<td>0</td>
</tr>
<tr>
<td>D8</td>
<td>1(25)</td>
</tr>
<tr>
<td>D9</td>
<td>1(28)</td>
</tr>
<tr>
<td>D10</td>
<td>0</td>
</tr>
</tbody>
</table>

4.3.3 INPUT FEATURE VALUE COMPUTATION METHOD

The proposed ranking function HRF, takes two inputs such as \( Wt(D,K_i) \) and \( Wt(Q,K_i) \). This section describes the strategy to calculate the worth for these input
variables. Weight of document \( D \) with respect to the keyword \( K_i \) is the product of Weighted Term Frequency (WTF) of the document with respect to keyword \( K_i \) and IDF of the keyword \( K_i \) as denoted in equation (4.1).

\[
Wt(D, K_i) = WTF(D, K_i) \times IDF(K_i) \tag{4.1}
\]

WTF measures the association between keyword \( K_i \) and document \( D \). It is computed by using equation (4.2).

\[
WTF(D, K_i) = \begin{cases} 
1 + \log(1 + \log(freq(D, K_i))) & \text{if } known = 1 \\
0 & \text{Otherwise} 
\end{cases} \tag{4.2}
\]

Where,

\( freq \) indicates the frequency of keyword \( K_i \) in document \( D \).

IDF is the scaling metric which measures the importance of keyword \( K_i \) in the document collection. It is calculated by using equation (4.3)

\[
IDF(K_i) = \log \frac{1 + |X|}{|Y|} \tag{4.3}
\]

Where,

\( X \) represents the total number of documents in the corpus and \( Y \) represents the total number of documents whose known value for keyword \( K_i \) is 1.

Likewise, weight of each document with respect to each keyword is calculated and maintained as a table.

The second input feature for the proposed HRF is \( Wt(Q,K_i) \). Here, \( Q \) indicates the user query. The retrieval system takes user query as input. The information need of the user transformed into the form suitable for retrieval system is called as user query. \( Wt(Q,K_i) \) represents the weight of query with respect to keyword \( K_i \) and it is expressed as the product of WTF of the query with respect to keyword \( K_i \) and IDF of the keyword \( K_i \). The formula to compute the value of \( Wt(Q,K_i) \) is shown in equation (4.4) and equation (4.5) is used to find the value of \( WTF(Q,K_i) \).

\[
Wt(Q, K_i) = WTF(Q,K_i) \times IDF(K_i) \tag{4.4}
\]
\[ W_{TF}(Q, K_i) = \begin{cases} 
1 + \log(1 + \log(freq(Q, K_i))) & \text{if known} = 1 \\
0 & \text{Otherwise} 
\end{cases} \] (4.5)

Where,

freq indicates the frequency of keyword \( K_i \) in query \( Q \).

All documents related to user query is retrieved and ranked using HRF. Then, ordered documents are presented to the user as answer list.

4.3.4 LAYERS OF PROPOSED HYBRID RANKING FUNCTION MODEL

Each individual layer in the neuro-fuzzy system is associated with particular step involved in the FIS.

4.3.4.1 INPUT LAYER

The neurons in this layer represent the input variables. Each neuron in this layer receives external crisp input values and sends directly to the next layer called as fuzzification layer as shown in equation (4.6). The proposed HRF model contains two neurons in this layer as there are two input variables such as \( Wt(D, K_i) \) and \( Wt(Q, K_i) \).

\[ M_i = N_i \] (4.6)

In equation (4.6), \( N_i \) represents the external crisp input value supplied to neuron \( i \) and \( M_i \) represents the output given by neuron \( i \).

4.3.4.2 FUZZIFICATION LAYER

The neurons in this layer denote the fuzzy sets in the antecedent part of the fuzzy rule. Every neuron in this layer gets outer crisp input value from the neuron of input layer and decides the level of its belongingness to the fuzzy set of the neuron.
The membership neurons in this layer are set to the activation function which specifies the fuzzy set of the neuron. As triangular sets have been utilized, the activation function of neuron is set to triangular membership function. Three parameters are utilized to indicate the fuzzy set \( X \) with triangular membership function. Let \( \{p, q, r\} \) be the three parameters. The level of membership of some value \( v \) to the triangular fuzzy set \( X \) is given by equation (4.7).

\[
\mu_X(v) = \begin{cases} 
0 & \text{if } v \leq p; \\
\frac{v - p}{q - p} & \text{if } p < v \leq q; \\
\frac{r - v}{r - q} & \text{if } q < v < r; \\
0 & \text{if } v \geq r;
\end{cases}
\]  

(4.7)

All input variables are represented by utilizing linguistic terms such as low, medium and high. These linguistic terms are denoted by triangular membership functions. Figure 4.3 (a) and Figure 4.3 (b) delineates the scope of input variables \( Wt(D,K_i) \) and \( Wt(Q,K_i) \) with respect to the linguistic terms used. The proposed HRF model contains six neurons in this layer which represents fuzzy sets of input variables.
4.3.4.3 FUZZY RULE LAYER

Each neuron in this layer represents a fuzzy rule. It receives input from the fuzzification layer neurons. For example in Figure 4.1, neuron R1 which corresponds to Rule-1 gets input from neurons L1 and L2 in fuzzification layer. In neuro-fuzzy model, product operator is used to implement intersection operation. In this layer, the output of neuron \( i \) is computed as shown in equation (4.8).

\[
M_i = N_{1i} \times N_{2i} \times N_{3i} \times \ldots \ldots \times N_{ni}
\]

(4.8)

Where,

\( M_i \) represents the output of neuron \( i \), \( N_{ji} \) represents the input values supplied to neuron \( i \) in which \( j \) ranges from \( l \) to \( n \) and \( n \) represents the number of input values.

Therefore in Figure 4.1, for Rule-1 the output is obtained as shown in equation (4.9).

\[
M_{R1} = \mu_{L1} \times \mu_{L2} = \mu_{R1}
\]

(4.9)
The proposed model contains nine neurons in this layer which represents the number of fuzzy rules in the rule base. The fuzzy rule base is shown in Table 4.2.

<table>
<thead>
<tr>
<th>Wt(D,K_i)</th>
<th>Wt(Q,K_i)</th>
<th>RSCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
</tbody>
</table>

The rule base contains the fuzzy rules of the form:

\[ If \ A1 \ AND \ A2 \ THEN \ C \]

Where,

\( A1 \) and \( A2 \) indicate the antecedents and \( C \) indicates the consequent of fuzzy rule.

The existing domain knowledge is transformed into fuzzy rules by using linguistic terms. Fuzzy rule base is created by using the domain knowledge of the IR.

4.3.4.4 OUTPUT MEMBERSHIP LAYER

Each neuron in this layer represents an output fuzzy set. The output of neurons in fuzzy rule layer is given as input to the neurons in this layer. All fuzzy rules must be combined in order to make decision. Therefore, fuzzy operation union is used to combine all input values of neurons in this layer. In neuro-fuzzy model, probabilistic OR operator
is used to implement union operation. The output of neuron $i$ in this layer is computed as shown in equation (4.10).

$$M_i = N_{1i} \oplus N_{2i} \oplus N_{3i} \oplus \ldots \ldots \oplus N_{ni} \quad (4.10)$$

Where,

$M_i$ represents the output of neuron $i$, $N_j$ represents the input values supplied to neuron $i$ in which $j$ ranges from 1 to $n$ and $n$ represents the number of input values.

Therefore in Figure 4.1, the output of neuron L, M and H are computed as shown in equation (4.11), equation (4.12) and equation (4.13) respectively.

$$M_L = \mu_{R_1} \oplus \mu_{R_3} \oplus \mu_{R_7} = \mu_L \quad (4.11)$$
$$M_M = \mu_{R_2} \oplus \mu_{R_4} \oplus \mu_{R_6} \oplus \mu_{R_8} \oplus \mu_{R_9} = \mu_M \quad (4.12)$$
$$M_H = \mu_{R_5} = \mu_H \quad (4.13)$$

The output variable is represented by using linguistic terms. These linguistic terms are denoted by triangular membership function as depicted in Figure 4.4. The range of output variable $RSCO$ is represented as low, medium and high. The proposed HRF model contains three neurons in this layer which corresponds to the number of fuzzy sets of output variable.

![Figure 4.4: Membership values for output variable RSCO](image-url)
4.3.4.5 DEFUZZIFICATION LAYER

The output fuzzy sets are given as input to the neuron in this layer. It combines them into single fuzzy set and gives a crisp value as output by using defuzzification method. In this proposed model, centroid defuzzification method is used and it is defined in equation (4.14)

\[
m^* = \frac{\int \sum_{i=1}^{n} \mu_{B_i}(m).m \, dm}{\int \sum_{i=1}^{n} \mu_{B_i}(m) \, dm}
\]  

(4.14)

Where, inputs for defuzzification function is fuzzy set \(\mu_{B_i}(m)\) and output is the crisp value \(m^*\).

4.4 RESULTS AND DISCUSSION

In medical literature, the most used and richest source of information is MEDLINE. The benchmark medical document corpus OHSUMED and PMC are used for experimental analysis (Hliaoutakis et al., 2009). PMC is a full document corpus of MEDLINE. In this experiment, the full document corpus with 894 medical documents is used. OHSUMED is a collection of MEDLINE document abstracts used for medical information systems evaluation. The document abstract corpus which contains 1246 medical document abstracts is used for experimental evaluation. The 15 experimental queries shown in Figure 3.5 are used to measure the performance of the retrieval system (Hliaoutakis et al., 2009). Document corpus contains 300 related documents for each query. The network is trained with the training data set that is generated from the rule base and parameters of the fuzzy membership functions.

The performance of the proposed ranking function is measured in terms of precision, recall and F-measure as characterized by mathematical statements (4.15), (4.16) and (4.17) respectively.
Where, \(\text{Num}_{RD}\) represents the total number of applicable documents retrieved based on user query, \(\text{Num}_{Ret}\) and \(\text{Num}_{Rele}\) indicates the total number of retrieved and applicable documents respectively.

The documents retrieved with respect to the given experimental query are arranged in the decreasing order of their relevance score. The values of precision, recall and F-measure are calculated for top 100, 200 and 300 retrieved documents. The obtained results are compared with conventional Cosine similarity based Ranking Function (CRF) and Fuzzy based Ranking Function (FRF) (Gupta et al., 2015).

4.4.1 OVERALL RETRIEVAL PERFORMANCE

Average precision, recall and F-measure are important measures used in the literature to verify the performance of the IR system. Therefore, Table 4.3 and Table 4.4 present the comparison of average precision, recall and F-measure values for OHSUMED and PMC corpus respectively. Results reveal that the proposed HRF attains better result when compared with other two ranking functions with respect to top 100, 200 and 300 retrieved documents. This is because, the proposed model is highly robust and it has the ability to capture the inherent features of the natural language text. Moreover, its learning, generalization and human-like information presentation capabilities improves its performance.
Table 4.3: Comparison of CRF, FRF and HRF for OHSUMED document corpus

<table>
<thead>
<tr>
<th>Ranking Function</th>
<th>Top 100 retrieved documents</th>
<th>Top 200 retrieved documents</th>
<th>Top 300 retrieved documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>0.94</td>
<td>0.19</td>
<td>0.32</td>
</tr>
<tr>
<td>FRF</td>
<td>0.96</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>HRF</td>
<td>0.97</td>
<td>0.29</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 4.4: Comparison of CRF, FRF and HRF for PMC document corpus

<table>
<thead>
<tr>
<th>Ranking Function</th>
<th>Top 100 retrieved documents</th>
<th>Top 200 retrieved documents</th>
<th>Top 300 retrieved documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>0.87</td>
<td>0.15</td>
<td>0.26</td>
</tr>
<tr>
<td>FRF</td>
<td>0.90</td>
<td>0.20</td>
<td>0.33</td>
</tr>
<tr>
<td>HRF</td>
<td>0.92</td>
<td>0.23</td>
<td>0.37</td>
</tr>
</tbody>
</table>

4.4.2 QUERY BASED RETRIEVAL PERFORMANCE

The precision, recall and F-measure results for top retrieved 100, 200 and 300 documents with regard to every query is shown below from Figure 4.5 to 4.13 for both OHSUMED and PMC document corpus. These figures clearly show that higher precision, recall and F-measure values are earned for proposed HRF for 15 queries in comparison to other two ranking functions, in case of OHSUMED and PMC document corpus.
Figure 4.5: Precision values of the top 100 retrieved documents with respect to 15 experimental queries
Figure 4.6: Precision values of the top 200 retrieved documents with respect to 15 experimental queries

(b) PMC

Figure 4.7: Precision values of the top 300 retrieved documents with respect to 15 experimental queries

(a) OHSUMED

(b) PMC
Figure 4.8: Recall values of the top 100 retrieved documents with respect to 15 experimental queries.
Figure 4.9: Recall values of the top 200 retrieved documents with respect to 15 experimental queries

(a) OHSUMED

(b) PMC

Figure 4.10: Recall values of the top 300 retrieved documents with respect to 15 experimental queries
Figure 4.11: F-measure values of the top 100 retrieved documents with respect to 15 experimental queries
Figure 4.12: F-measure values of the top 200 retrieved documents with respect to 15 experimental queries

(a) OHSUMED

(b) PMC

Figure 4.13: F-measure values of the top 300 retrieved documents with respect to 15 experimental queries
Thus, these experimental results reveal that the proposed HRF achieves better result when compared with other two ranking functions improving the retrieval performance.

4.5 CONCLUSION

A new hybrid neuro-fuzzy system based ranking function called HRF is proposed for improving the performance of the IR system. Conventional statistical ranking functions fail to capture the inherent features of documents and queries. FRF can address upon the vagueness and uncertainty present in natural language text. But they are not robust enough to the topological changes of the system. The main strength of the proposed HRF lies in combining the learning and generalization capabilities of neural network with human-like knowledge representation and explanation capability of fuzzy system. Thus, the drawbacks of fuzzy systems are paid by the capabilities of the neural networks. Benchmark medical document corpus OHSUMED and PMC are used to validate the proposed ranking function. The experimental results uncover that the proposed HRF attains better results when compared with FRF and CRF.