DESIGN OF INTELLIGENT NEURO-FUZZY BASED DOCUMENT RETREIVAL SYSTEM

[1]Ituma, C., [2]James, G. G., and [3]Onu, F. U.
Department of Computer Science
Ebonyi State University, Abakaliki.
Correspondence author: gabresearch@gmail.com 08107381867

Abstract
As information grows rapidly over the web, it becomes difficult for researcher to find the information they are looking for. The rapid growth of information on the web and the inadequacy of the conventional search engines to retrieve relevant information based on user's request have motivated this research. Tracking, classification and retrieval of documents required tools that could recognize patterns in data as well as process imprecise information. FL lacks the capability to learn from previous data and NN equally lacks the capability to handle imprecise and incomplete data. This makes Neuro-Fuzzy systems one of the best options for document tracking, as the weakness of Fuzzy Logic and Neural Network are complimented whilst the strength of the individual components are enhanced. This work extends earlier fuzzy IR models by adding more fuzzy linguitic values, fuzzy variables, using different membership functions, using rules that consider the document structure and optimizing the throughput using Nuero-fuzzy system. The document is quantified by describing it using features/linguistic variables that contribute most to its relevance to the query like the Lexical density, term weight, document similarity vector as well as word ratio. Linguistic values are assigned to each of these variables that associate them with membership degrees. Different membership functions were investigated for each of these linguistic variables and the best function was chosen for each variable which contributed the most to the IR precision. An ANFIS inference engine was built using the seugeno method that handle these variables to measure the degree of document relevance to the query. It was found that using ANFIS improved the performance slightly by some percentage over the FIS results.

Keywords: Neuro-fuzzy geno Method, ANFIS model, Intelligence.

1. Introduction

The search for information is always a major issue for researchers. Often time, people travel over a distance to track the most needed data for their research work, thereby making the task of research difficult. This has been the tradition until researchers in the field of information technology came up with the techniques of intelligent search engine, which used the web to facilitate the search for information. This work proposes a hybrid intelligent search system based on neuro-fuzzy paradigm for document tracking and retreival. Existing search engines such as Google, Yahoo, and Bing often return a long list of results which forces users to sift through it to find relevant documents; thereby making search for information difficult.The dissertation proposed a neuro-fuzzy based model for classification of search results based on the strength of words in the query. The neuro-fuzzy clustering technique will be employed to classify the documents into groups of similar topics for specific knowledge.

Fuzzy Logic (FL) is a logic that its ultimate goal is to provide foundations for approximate reasoning using imprecise propositions based on fuzzy set theory. It serves mainly as apparatus for fuzzy control, analysis of vagueness in natural language and several other application domains. FL has the capability of handling imprecise, incomplete and vague information, as well as ability to represent partial truth. FL is limited by its inability to learn from previous data.

Neural-Network (NN) (or Artificial Neural Network (ANN)) are considered as family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) which are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. NN is limited by the inability to handle imprecise and incomplete data. Similarly, neural network can approximate a function, but it is impossible to interpret the result in terms of natural language. Hence, the need for the fusion of neural networks and fuzzy logic in neuro-fuzzy models provide learning as well as readability.

Neuro-Fuzzy systems are derived from the combination of Neural Network (NN) and Fuzzy Logic (FL) techniques which result in hybrid intelligent system. According to Yuayuan et al (2009), Neuro-Fuzzy hybrid systems offer solutions to real life problems by synergizing the human-like reasoning capabilities of FL with the learning and connectionist structure of NN. In the hybridization, the weakness of Fuzzy Logic and Neural Network are complimented whilst the strength of the individual components are enhanced.The essence of applying the neuro-fuzzy techniques is to build an adaptive intelligent information retrieval system which will cluster internet documents into similar topic using an unsupervised machine learning techniques to reduce the percentage of irrelevant documents that are retrieved and presented to users.

2. Literature Review

Neuro-fuzzy techniques have been shown to be effective in the area of detection, control, processing, approximation, and building of adaptive based systems. In Ejiofor, (2014) an intelligent search system for topic tracking and classification of Documents was presented. The work was motivated by the need to develop an automated intelligent search tracking and classification system to collaborate with the internet search engines using web services to: cluster the result of user query into groups of related topics, find the information they are looking for more easily, realize faster that a query is poorly formulated and to reformulate it, and reduces the fraction of the quires on which the user give up before reaching the desired information.

In Cartos, (2015) an Event-based multi-document summarization was presented. The work was motivated by the desire to explore three different ways to integrate event information, achieving state-of-the art results in both single and multi-document summarization using filtering and event – based features. The approach used was based on a two-stage single-document method that extracts a collection of key phrases, which are then used in a centrally-as-relevance passage retrieval model. The event detection method was based on fuzzy fingerprint, which is a supervised method trained on documents with annotated events tags.

In Tina, (2001) a Building Intelligent Agent that learns to retrieve and extract information was presented. The work was motivated by the need to develop extractor components that creates specialized agents that accurately extract pieces of information from document in the domain of interest.

In Kermaldeep*et al.*, (2012), a survey of topic tracking techniques was presented. According to Kermaldeep*et al.*, (2012) the field of text missing has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available documents; hence the need to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The main purpose is to develop a topic tracking system for text missing process to identify and follow events presented in multiple news sources; and to collects dispersed information together and makes it easy for user to get a general understanding. The techniques employed usually do not involve deep linguistic analysis or parsing, but rely on simple 'bag-of-words' text representation based on vector space.

4. System Analysis and Design

4.1Analysis of the Proposed System

The proposed system is a neuro-fuzzy based model for searching and classification of document using library of e-books as a case study. FL lacks the capability to learn from previous data and NN equally lacks the capability to handle imprecise and incomplete data. This makes Neuro-Fuzzy systems one of the best option for document tracking and classification as the weakness of Fuzzy Logic and Neural Network are complimented whilst the strength of the individual components are enhanced.

The essence of applying the neuro-fuzzy techniques is to build an adaptive intelligent information retrieval system which will cluster electronics library documents into similar topic using an unsupervised machine learning techniques to reduce the percentage of irrelevant documents that are retrieved and presented to users.The document that shall be used as case study is e-library files. The system shall have a four (4) input variables and two (2) outputs. The input variables to be considered are: File size, Search Words (keywords), search index and multiplicity. The system shall be made up of two sections; that is, the fuzzy Logic section and then the neural network section. The fuzzy section shall handle the tracking while the neural network will do the classification and retrieval and as well optimized fuzzy retrieval system.

4.2 Benefits of the Propose system

The propose system have the following benefits:

- ➢ The hybridization will help, the weakness of Fuzzy Logic and Neural Network are complimented whilst the strength of the individual components are enhanced to enhanced an effective and efficient tracking and classification of web document.
- ➢ The system has an inbuilt hybrid intelligent search model based on neuro-fuzzy paradigm that has the capability to cluster internet documents into similar topic using an unsupervised machine learning techniques to reduce the percentage of irrelevant documents that are retrieved and presented to users.
- ➢ There is available ANFIS architecture and well-defined neuro-fuzzy based mathematical models to enhance the intelligent searching, tracking, clustering, ranking and classification of the relevant web documents.
- ➢ Availability of a vector model that enhances feature selection, reduction and weighting.
- ➢ It has the ability to classify documents into similar group for domain knowledge.
- ➢ It meet advanced search needs. And ability to automatically expand users query without intervention

5.System Architecture

Figure 1 below show the architecture diagram of the proposed NFDTRS:

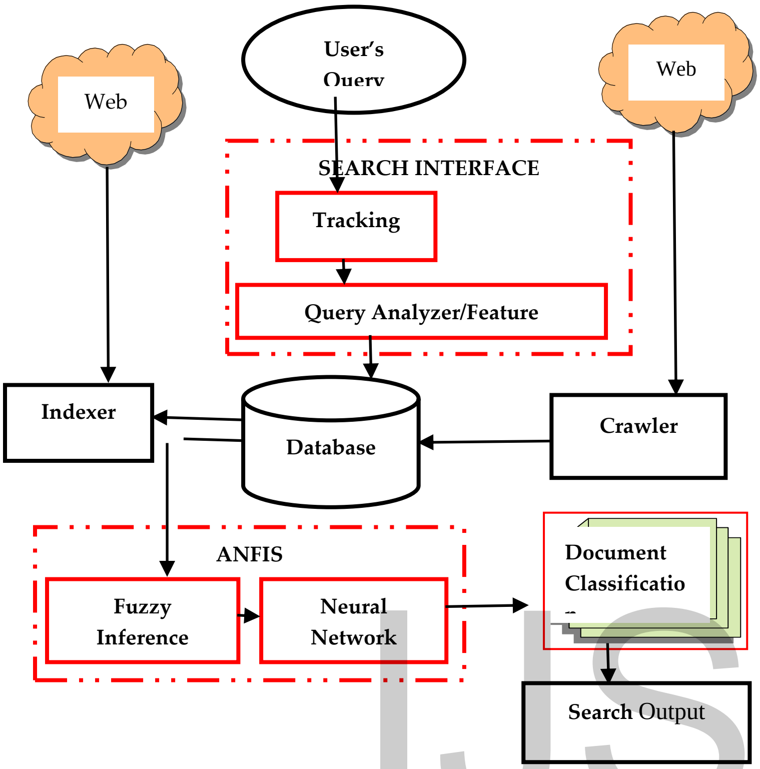


Fig. 1: The Architecture of the Proposed Neuro-Fuzzy Based Document Tracking and Classification System (NFDTRS).

6Class Diagram

The class diagram of a NFDTRS depicted in figure 41 shows the classes, the method and the operation on the data.

7. Program Algorithm

Step 1: Generate Fuzzy System
Step 2: Optimize the FIS using ANFIS Model
Step 3: Test ANFIS performance
Step 4: Choose best Model
Step 5: Evaluate ANFIS

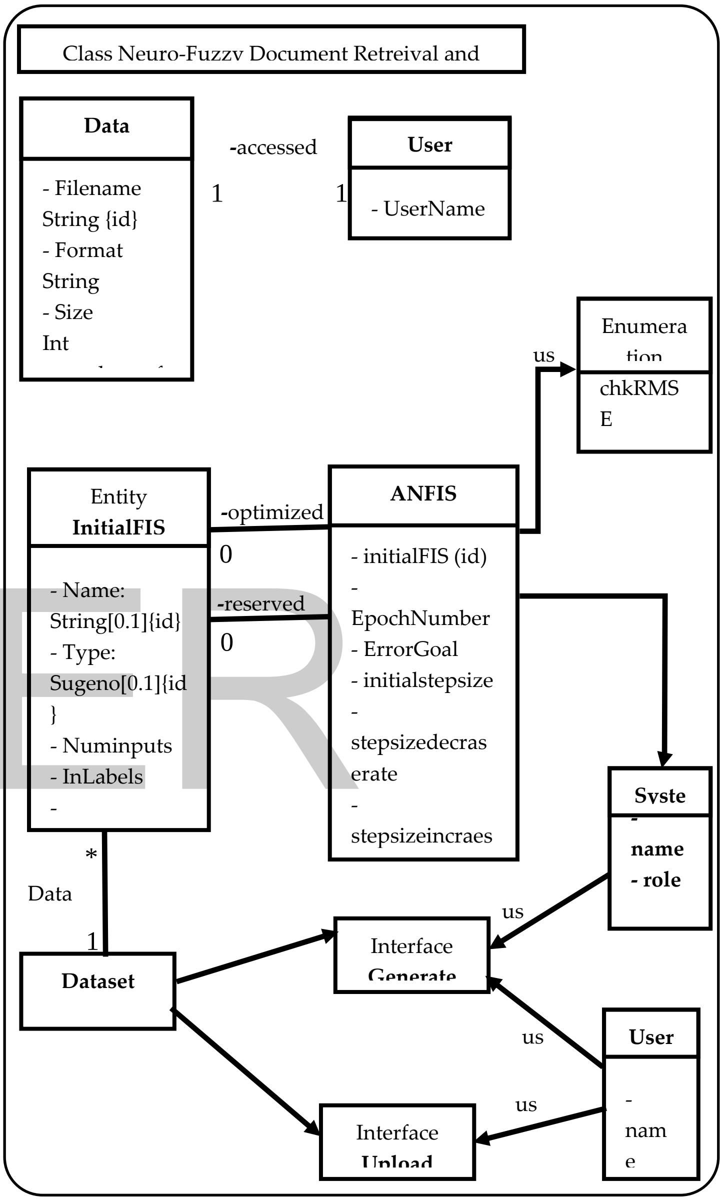


Fig. 2: Class Diagram for NFDTRS

8 Learning Procedure for the Proposed System

The Learning procedure is based on ANFIS as shown in figure 3

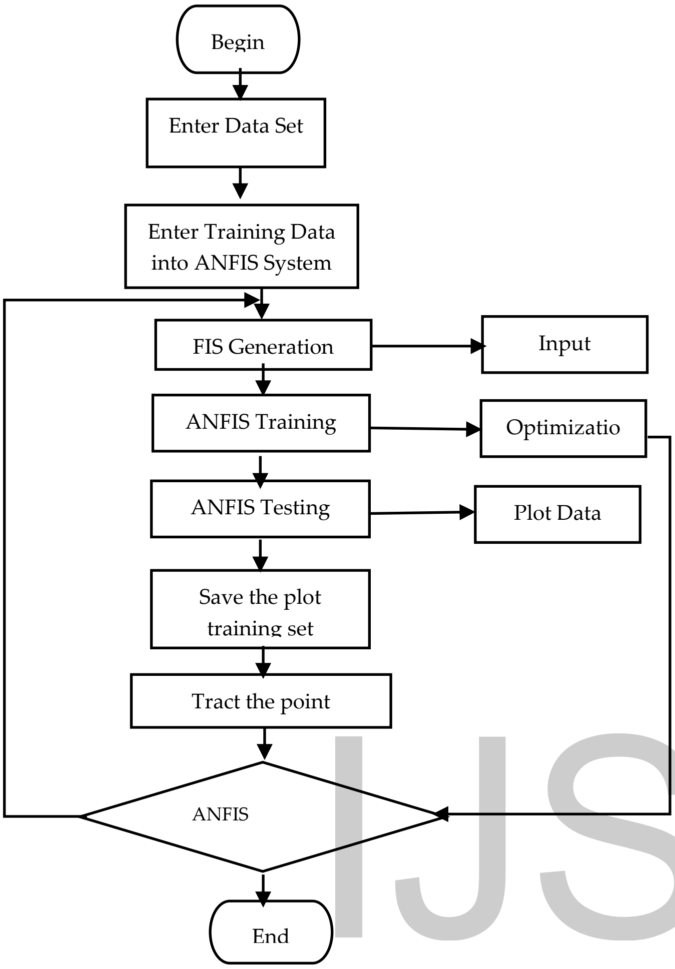
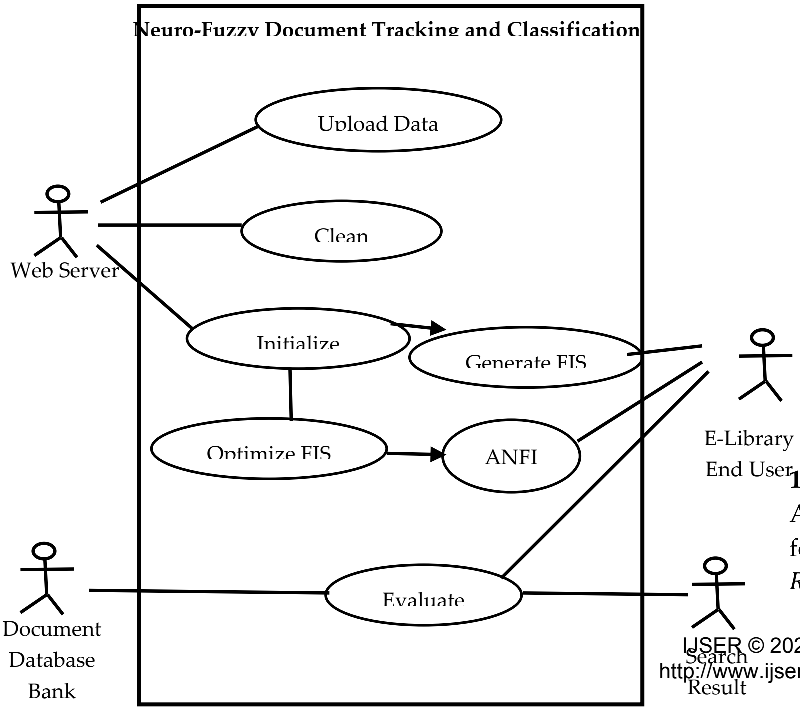


Fig. 3: A Learning Procedure of a NFDTRS

9. Use Case Diagram

The Use Case Diagram depicted in figure 4 shows the actors, the relationships and interaction within the system



10 Fuzzy Logic Model

In this work, a type-1 fuzzy logic model is used. This model is based on a triangular membership function that defines a degree of membership of all crisp values within the specified universe of discourse. The conceptual architecture of the fuzzy logic model used is presented in Figure 5:

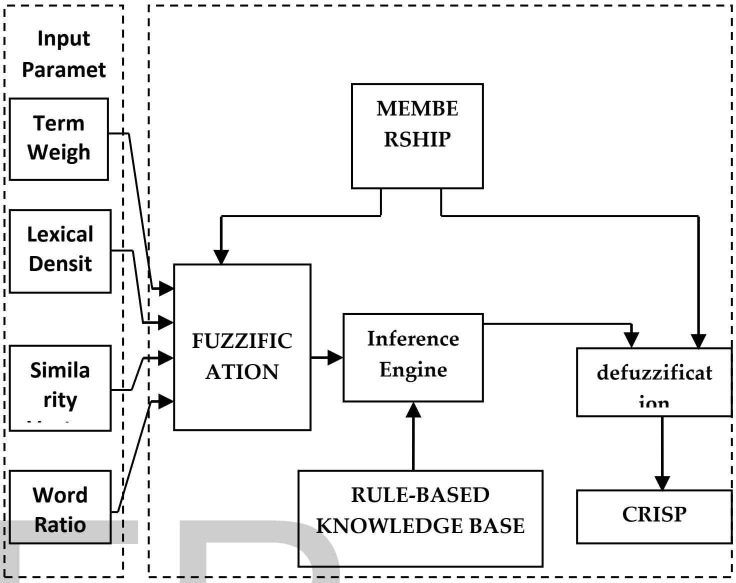


Fig. 5: Conceptual Architecture of Fuzzy Logic Model

10.1 Components of the Fuzzy logic Model

The following components constitute the fuzzy logic model used in this work;

i. Fuzzification Module: this module maps the crisp input to a type-1 fuzzy set using the triangular membership functions defined for this work.

ii. Inference Engine: this module evaluates the rules in a rule base against fuzzy set gotten from Fuzzification to produce another fuzzy set.

iii. Defuzzification Module: it maps the fuzzy set from inference engine to a crisp output using center of gravity defuzzification method (or Centroid).

iv. Knowledge Base: This is a database of rules (rules are generated from experts' knowledge) to be used by the inference engine.

v. Membership Function: This is a mathematical equation that maps a crisp input value to a degree of membership between 0 and 1, called the fuzzy set.

12Fuzzy Rule Base

A fuzzy rule is defined as a conditional statement in the form:

$R^l: IF\ x_1\ is\ \tilde{F}_1^l\ and\ \ldots\ x_p\ is\ \tilde{F}_p^l\ THEN\ y\ is\ \tilde{G}_1^l$

Where:
$l = 1, \ldots, M$, is rule number
R - is the current rule
p - is the number of linguistic variable
x_p - is the p's linguistic variable
$\tilde{F}_p^l$ - is the p's linguistic term of rule l
$\tilde{G}_1^l$ - is the output linguistic variable of rule l

13ANFIS Initialization and Generation

FIS parameter can either be initialize to one's preference, ANFIS can initialize the parameters automatically, it has the following parameters: 4 inputs, 5 MFs and one linear output are showed in figure 6:

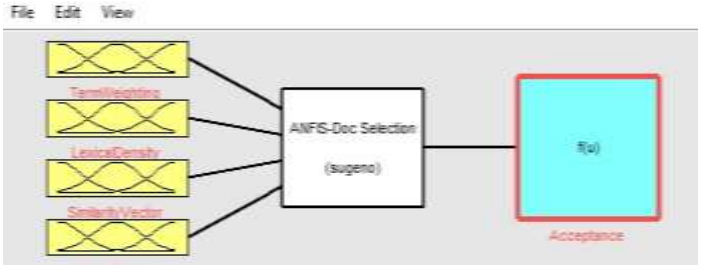


Fig. 6: ANFIS Inference Editor

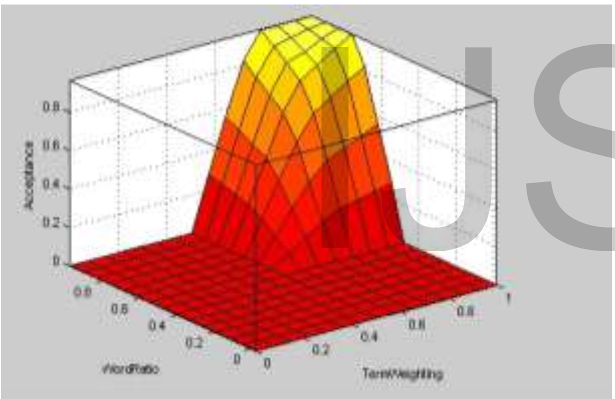


Fig. 7: ANFIS Surface Viewer (Term Weighting to Word Ratio)

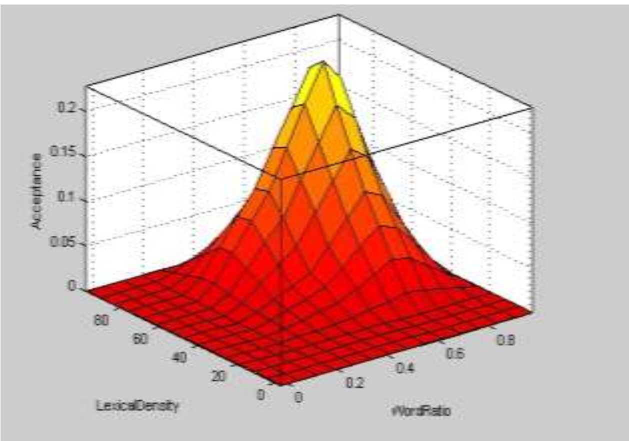


Fig. 8: ANFIS Surface Viewer (Lexical Density to Word Ratio

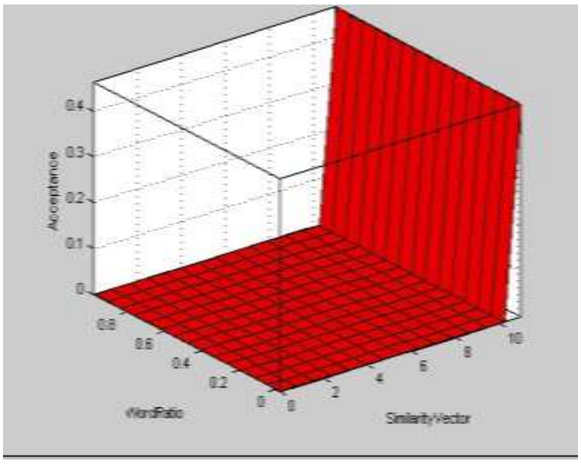


Fig. 9: ANFIS Surface Viewer (Similarity Vector to Word Ratio)

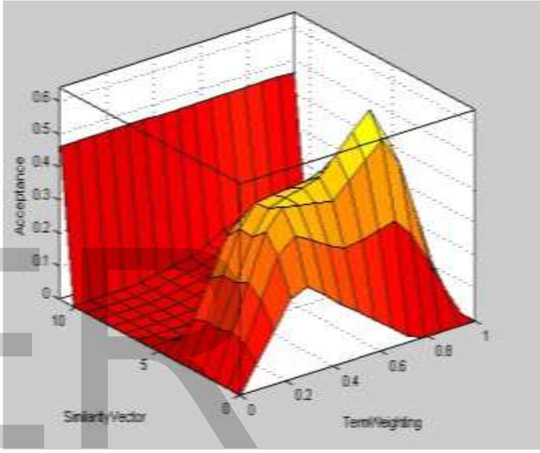


Fig. 10: ANFIS Surface Viewer (Term Weighting to Similarity Vector)

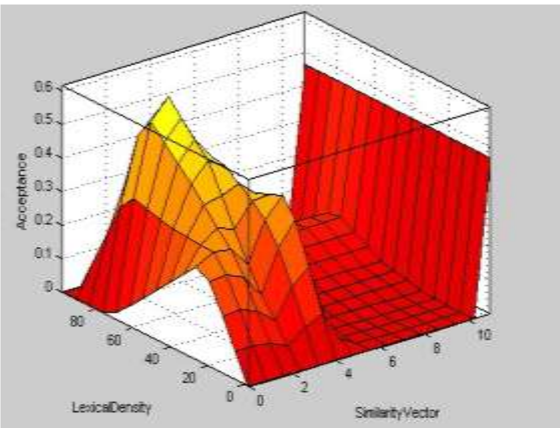


Fig. 11: ANFIS Surface Viewer (Similarity Vector to Lexical Density)

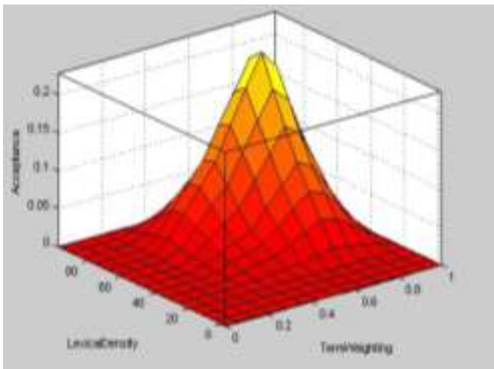


Fig.12: ANFIS Surface Viewer (Term Weighting to Lexical Density)

14 ANFIS Implementation

The adaptive-network-based fuzzy inference system is capable of constructing input-output. Mapping accurately based on both human knowledge and stipulated input-output data pairs. However, once a fuzzy model is developed, in most cases its needs to undergo an optimization process. The aim optimizing and refining in two folds: the model structures and parameters.

14.1 ANFIS Training Procedure

Figure 13 shows the ANFIS training editor which is made up of six major parts namely: load data, generate FIS, Test FIS Output and ANFIS information. Figure 14 shows the training error, figure 15 shows the checking window, and figure 35 shows the training output at 300 epochs.

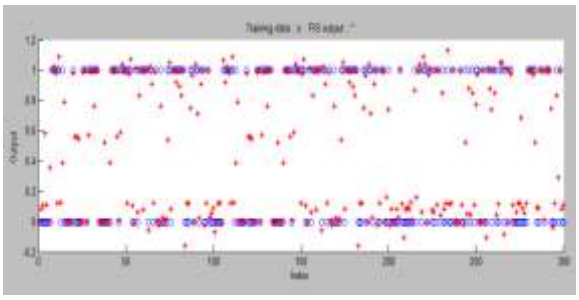


Fig. 13: ANFIS Training Window

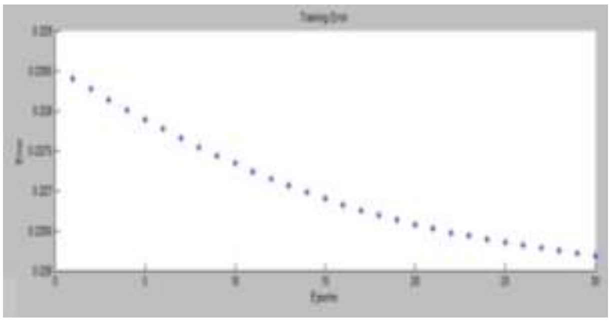


Fig. 14: ANFIS Training Error window

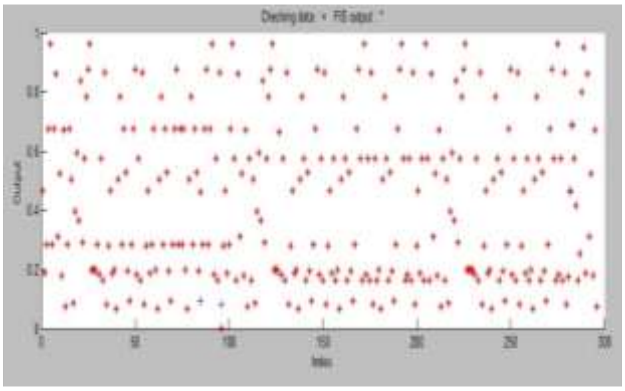


Fig. 15: ANFIS checking Window

The checking data are represented by the (+) sign the number of checking data pairs are 300.

15. Conclusion

This paper on adaptive neuro-fuzzy document retrieval systems is divided into six chapters. Chapter one discussed the aim, statement of problems, motivation of the study and the scope of the study. Chapter two explored the review of related literature in the following headings: Information Retrieval Systems, Information Retrieval Engine Classification, Distributed System, Classification Models, Information Retrieval Models, History of Intelligent Search Agents, Neural Networks and Learning System and its applications in IR, Document Clustering, Fuzzy Logic (FL) and its application in IR, Clustering and Clustering Algorithm. Chapter three discussed the analysis and methodology of the NFDTRS. This included the Design Methodology, Analysis of the Existing System, Analysis of the Proposed System and Feature Analysis of Programming Tools. Chapter four discussed on the design and implementation of the new system, with emphasis on: System Architecture, Class Diagram, Learning Procedure for the Proposed System, Use Case Diagram, FIS Generation for the Proposed NFDTRS, FIS Implementation of the of the new system, ANFIS Implementation of the of the new system and Java implementation of the ANFIS model. The fuzzy model was designed in three main stages to include fuzzification, inference system and defuzzification. The triangular membership (Tmf) function was used to map the input parameters to the output parameter. ANFIS model was designed using Takagi Sugeno inference mechanism. Java programming language was used on a windows 10 platform to integrate the Matlab ANFIS output for better optimization and MySQL database 5.7.14 from WAMP server 3.0.6 was used as the back-end engine. To validate this work, data was collected by retrieving test data from Text Retrieval Conference (TREC) of 2011.

We used the hybrid supervised learning approach in training our network. The training, testing and checking

KMSI values of 0.026347, 0.026073, 0.025819 respectively were observed in the hybrid learning process at 150 epochs. The ANFIS processes faster and have a minimal error of 0.026344 at 300 epochs.

Average error of 0.047283 was observed in the hybrid algorithm against the average error of 0.024642 with hybrid learning at 300 epochs. Therefore, ANFIS evaluation of NFDTRS with hybrid learning performed better than the FIS evaluation.

16. RESEARCH CONTRIBUTION

The main contribution of my study is that my IR-ANFIS system succeeded in enhancing the results achieved by previous IR systems which used fuzzy logic. As one can see, system proved to outperform IR-FIS and other industry standard search engines.

Secondly, my contribution is the discovery that the standard use of Neuro-Fuzzy techniques as it is used in other fields slightly improved the performance in the information retrieval field.

References

Iwok, S O (2018) A Model of Intelligent Packet Switching in Wireless Communication Networks. PhD Thesis, Department of Computer Science, Ebonyi State University Abakaliki.

Udoh, Samuel Sunday (2016) Adaptive Neuro-Fuzzy Discrete event System Specification for Monitoring Petrol Product Pipeline. PhD Dessertation of the Department of Computer Science, Federal University of Akure.

Yuanyam, C., Limin J., Zundong Z., (2009) Mamdani Model Based Adaptive Neural Fuzzy Inference System and its Application. International Journal of Information and Mathematical Sciences 5(1), 2229-2235.

Chu, H., “Information representation and retrieval in the digital age”, American Society for Information Science and Technology, ISBN 1-57387-172-9, Vol. 9, Pp. 111-112, 2003.

Chen, H., Shank, G., Iyer, A., & She, L., "A machine learning approach to inductive query by examples: An experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing", Journal of the American Society for Information Science, Vol. 49, Pp. 693-705, 1998 .

Ejiofor, C. I. (2014), An Intelligent search system for topic tracking and classification of documents, Ph.D dissertation of University of Port Harcourt, Nigeria, June 2014.

Tina Eliassi-Rad(2001), Building intelligent Agents that learn to retrieve and extract information, Ph.D dissertation, 2001.

Kamadeep Kaur, Vishal Gupta (2012), A survey of topic tracking techniques, international journal of Advanced Research in computer science and software engineering, volume 2, Issue 5, pp. 384-385.

Wang, S., Socher, R.; Perelygin, A.; Wu, J. Y.; Chuang, J and Manning, C. D. 2017. Baselines and bigrams:Simple, good sentiment and topic classification. In *ACL*, 90–94.

Wu, B-F., C-C. Chiu, and Y-L. Chen. "Algorithms for compressing compound document images with large text/background overlap."*IEE Proceedings-Vision, Image and Signal Processing* 151.6 (2004): 453-459.