I.J. Modern Education and Computer Science, **2015, 8, 69-84**
Published Online August 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.08.08

Web Pages Retrieval with Adaptive Neuro Fuzzy System based on Content and Structure

Mohammad Saber Iraji
Faculty Member of Department of Computer Engineering and Information Technology, Payame Noor University, I.R. of Iran
Email: iraji.ms@gmail.com

Hakimeh Maghamnia
Department of Computer Engineering and Information Technology, Payame Noor University, I.R. of Iran
Email: h.maghamnia@gmail.com

Marzieh Iraji
Department of Computer Engineering and Information Technology, University College of Rouzbahan, Sari, Iran
Email: marziehiraji@gmail.com

Abstract—Volume of web pages and information on the web is constantly increasing. In this paper, we presented a system to retrieve pages relevant to a query, that can be used by the search engines. The design of our proposed system, content, Page content of neighbors, Connectivity (link analysis) features were used and the methods of fuzzy Sugeno and adaptive fuzzy neural network methods considered .Results showed that the neural method, the error is less than other methods, in the retrieval of web pages tailored to the users search query on the Web, can increase the efficiency of search engines.

Index Terms—Web pages retrieval, adaptive neuro fuzzy, search engines.

I. Introduction

Applications of computer and Internet is searching large volumes of pages, and information retrieval researchers and computer users. Most people use searching through a query search engines like Google use. The volumes of information available on the Internet are increasing every moment. Members looking for useful information on the mass are important for search engines to provide useful information to users, often. In Search engines the main challenge is determine relevant documents and irrelevant to the query.

Search engine benefit from spiders such as Web robots, using different algorithm to retrieve Web pages relevant to a specific domain. Filtering methods are divided into four categories:

1. Determine the relevance of a Web page to a subject manually by experts [1].
2. Suitability of a web page to a specific topic, the number of occurrences of keywords [2].
3. TFIDF (term frequency inverse document frequency) is computed based on a lexicon [3].
4. Text classification methods that applied to web pages[4].

Web page filtering can be pragmatic in search engines and Web applications such as Web content management. Event spamming circumfuse on web pages, after email. Result of web spamming is decrease quality for search engine. Thus, it wasteful pages indexed in the search engines and query processing cost increases[5].This is a challenge for servers, provide the appropriate information to Internet users based on the content and links of web pages. This article is motivated by designing a neural fuzzy system in order to accurately retrieve Web pages, according to Internet users' queries.

The aim of this study is to examine and discuss about the web pages retrieval system, this paper attempts to optimize the web pages retrieval algorithm. The paper is organized in five sections. After the introduction in Section I, Section II which also introduces the related works of web pages filtering. Section II continues with Adaptive neuro fuzzy models for proposed system and examples in section III. Section IV and V presents the results, conclusions of the research. The paper ends with a list of references.

II. Work History

Google scholar is a search engine that use for researcher. Google Scholar Citation is the highest factor in the retrieval process and the incidence of a search word in an article's title to have a potent impact on the article's ranking[6].Rongmei Li be evolved clicked pages from clicked domains In order to improve the efficiency of Web information retrieval[9].

Hema Dubey, B. N. Roy offer a new page rank algorithm base on mean page ranks and reduces algorithm complexity[10] .Bhamidipati, et all introduce the score fusion technique and apply when two pages

have same ranking[11] .Sharma, et all were compared Different methods for ranking web pages with different algorithms[12].

Minnie, et all have implemented Links Algorithm and other algorithms[13]. the result is retrieved If an exact match occurs, otherwise not. Qiu, Hemmje, et all offer page filtering system based on page links, information's page links for improve search query algorithms[14].

In [15] report a machine-learning-based method that mix Web content and structure analysis. They display each Web page by a set of content-based and link-based features. They were used type of neural networks Namely support vector machine and compare their proposed method with two existing web page filtering methods — a keyword-based method and a lexicon-based method and results perform better.

Scarselli, et all have introduced a machine learning type for web spam discovery based on Graph Neural Networks PM-GraphSOMs. they use Link-based(Degree-related measures, PageRank, TrustRank, Truncated PageRank, Estimation of supporters) and Content-based features(Fraction of anchor text, Fraction of visible text, Compression rate, Corpus precision and corpus recall, Query precision and query recall, Independent trigram likelihood, Entropy of trigrams) in your proposed system[5]. They were appraise their system into a training data (8339 pages) and a test data (1851 pages) from WEBSPAM-UK2006 dataset. The results show that the optimization of their method.

Khokale, et all offered a Web information retrieval with Fuzzy logic. In the proposed system, page Rank Score, Normalized Hub Score, Normalized Authority Score as inputs, and outputs related Document and Non-related Document, had been considered. In these system, would improve imprecise query users using fuzzy inference systems and increase the efficiency of search engines[16].

III. Method and Model

Soft computing are included, different types of neural networks, fuzzy systems, genetic algorithms, etc that in information retrieval applications. Fuzzy theory was developed by Zadeh[17], a new intelligent method stated To solve different problems than the old calculations. Chau, & Chen have been identified 14 features be used as the input values for svm neural network[8].

We have considered page content, Page content of neighbors, Connectivity (link analysis) features to filter out Web pages as input to the web page filter proposed system. Then these features were fuzzy, In order to determine the exact number of categories correct, relevant pages with queries.

A. Fuzzy Page content

Page content feature is dependent on the title, TFIDF factors of web page. Title(p) is determined by count of words in the title of page p discover in the domain lexicon and TFIDF(p) Meaning summarize of TFIDF of the words in page p discover in the domain lexicon[3].mamdani Fuzzy systems to calculate the exact value for page content is shown(Fig. 1).

Input variables Title(p), TFIDF(p) are given in the range 0 to 1 after the change of scale. we consider Three linguistic variables are low, average, high, with triangular membership functions for the variables (Fig. 2). Fuzzy rules is presented in Fig. 3, to determine the exact amount of fuzzy page content value by experts working.

B. Fuzzy Page Content of Neighbors

Due to the hierarchical tree structure in architecture of web pages, neighbors Web pages for a Web page have a very important role in retrieving a web page. Page content of neighbors feature is dependent on the InTitle(p), InTFIDF(p), OutTitle(p), OutTFIDF(p), SiblingTitle(p), SiblingTFIDF(p) factors of web page p[19] .

- InTitle(p) Is determined by medium count of words i.
- N title of page z discover in the domain lexicon .
- InTFIDF(p) Is determined by medium summarize of TFIDF of the words in Page r discover in the domain lexicon) for all incoming Pages r of p.
- OutTFIDF(p) Is determined by medium summarize of TFIDF of the words in Page j discover in the domain lexicon) for all outcoming Pages j of p.
- OutTitle(p) Is determined by medium count of words in title of page t discover in the domain lexicon for all outgoing pages t of p
- SiblingTitle(p) Is determined by medium count of words in title of page w discover in the domain lexicon for all sibling pages w of p.
- Sibling TFIDF(p)) Is determined by medium summarize of TFIDF of the words in Page u discover in the domain lexicon) for all sibling Pages u of p.

The mamdani Fuzzy systems to compute the exact score for Fuzzy Page content of neighbors is shown(Fig. 4).Input variables are given in the range 0 to 1 after the change of scale. we consider Three linguistic variables are low, average, high, with triangular membership functions for the six input variables (Fig. 5). Fuzzy rules is introduced in Fig. 6, in order to the exact amount of fuzzy Page content of neighbors score by experts working. As an example rule : if InTitle(p) is average, InTFIDF(p) is low, OutTitle(p) is average, OutTFIDF(p) is low, SiblingTitle(p) is low, SiblingTFIDF(p) is low then Fuzzy Page content of neighbours is low.

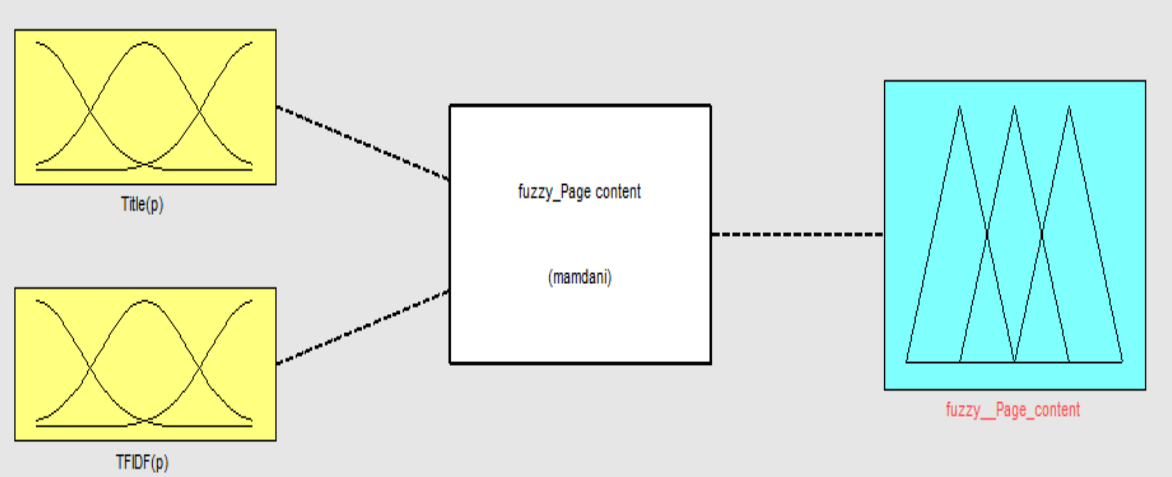

Fig.1. Fuzzy Page Content Score

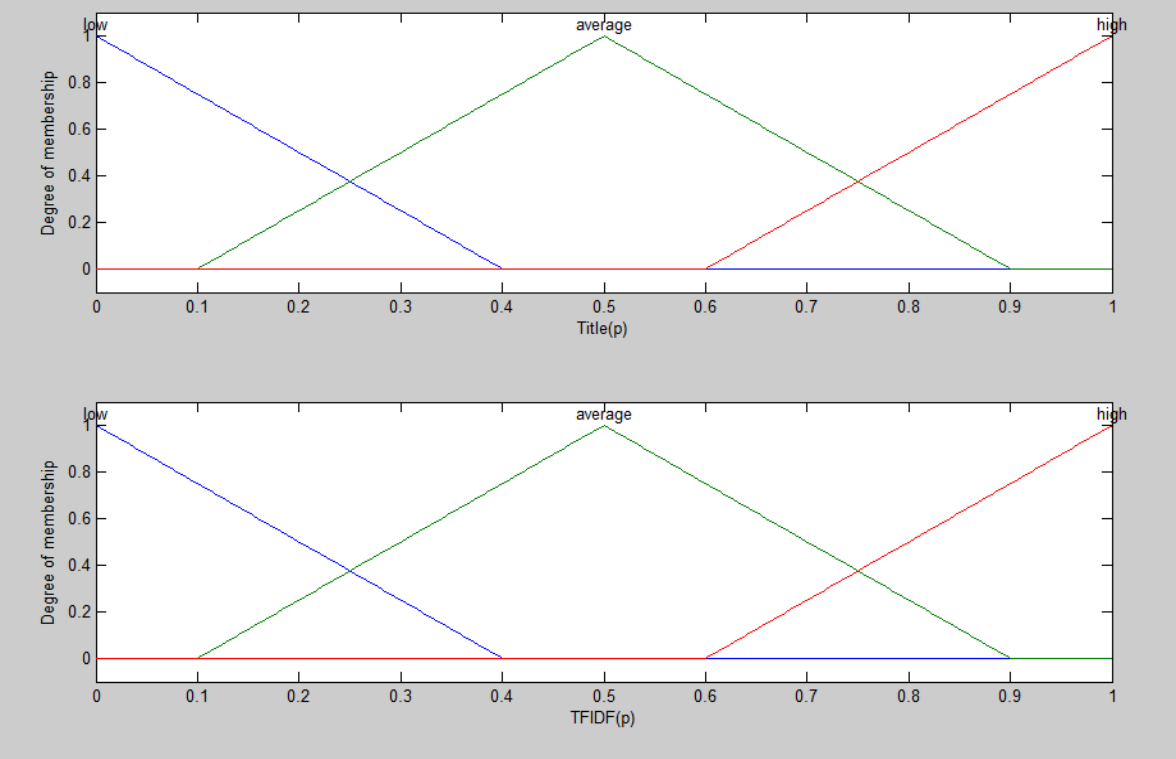

Fig.2. Fuzzy Input Variable for Page Content

1. If (Title(p) is low) and (TFIDF(p) is low) then (fuzzy__Page_content is low) (1)
2. If (Title(p) is low) and (TFIDF(p) is average) then (fuzzy__Page_content is average) (1)
3. If (Title(p) is low) and (TFIDF(p) is high) then (fuzzy__Page_content is high) (1)
4. If (Title(p) is average) and (TFIDF(p) is low) then (fuzzy__Page_content is low) (1)
5. If (Title(p) is average) and (TFIDF(p) is average) then (fuzzy__Page_content is average) (1)
6. If (Title(p) is average) and (TFIDF(p) is high) then (fuzzy__Page_content is high) (1)
7. If (Title(p) is high) and (TFIDF(p) is low) then (fuzzy__Page_content is average) (1)
8. If (Title(p) is high) and (TFIDF(p) is average) then (fuzzy__Page_content is high) (1)
9. If (Title(p) is high) and (TFIDF(p) is high) then (fuzzy__Page_content is high) (1)

Fig.3. Fuzzy Rules for Fuzzy Page Content Value

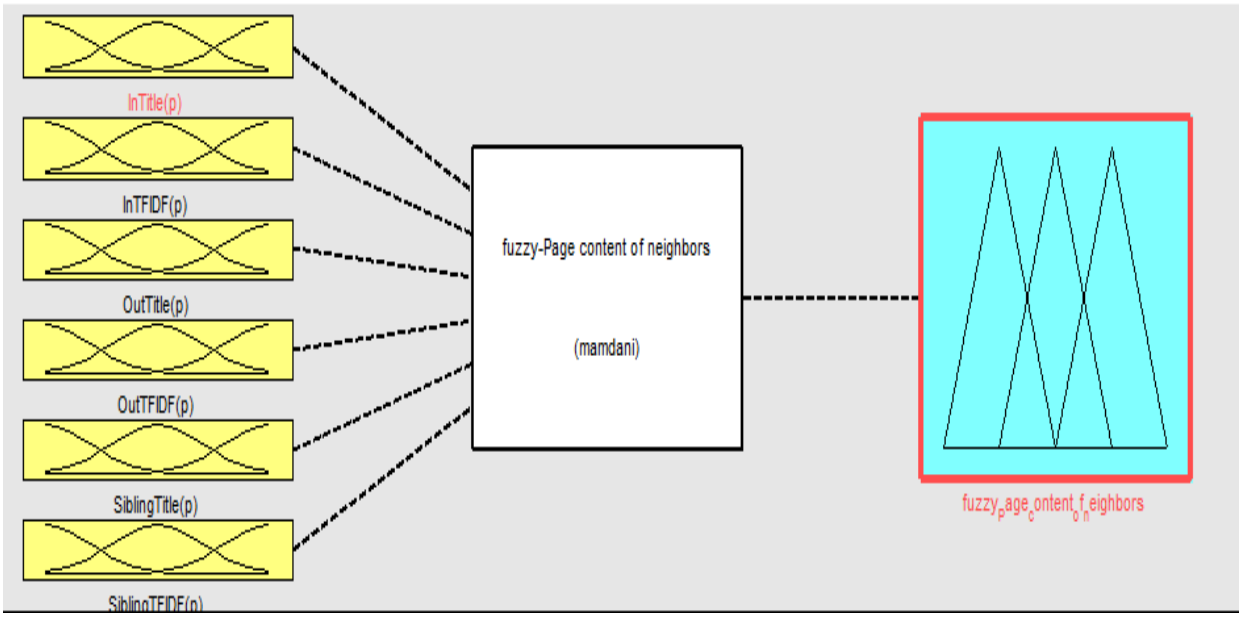

Fig.4. Fuzzy Page Content of Neighbor's Value

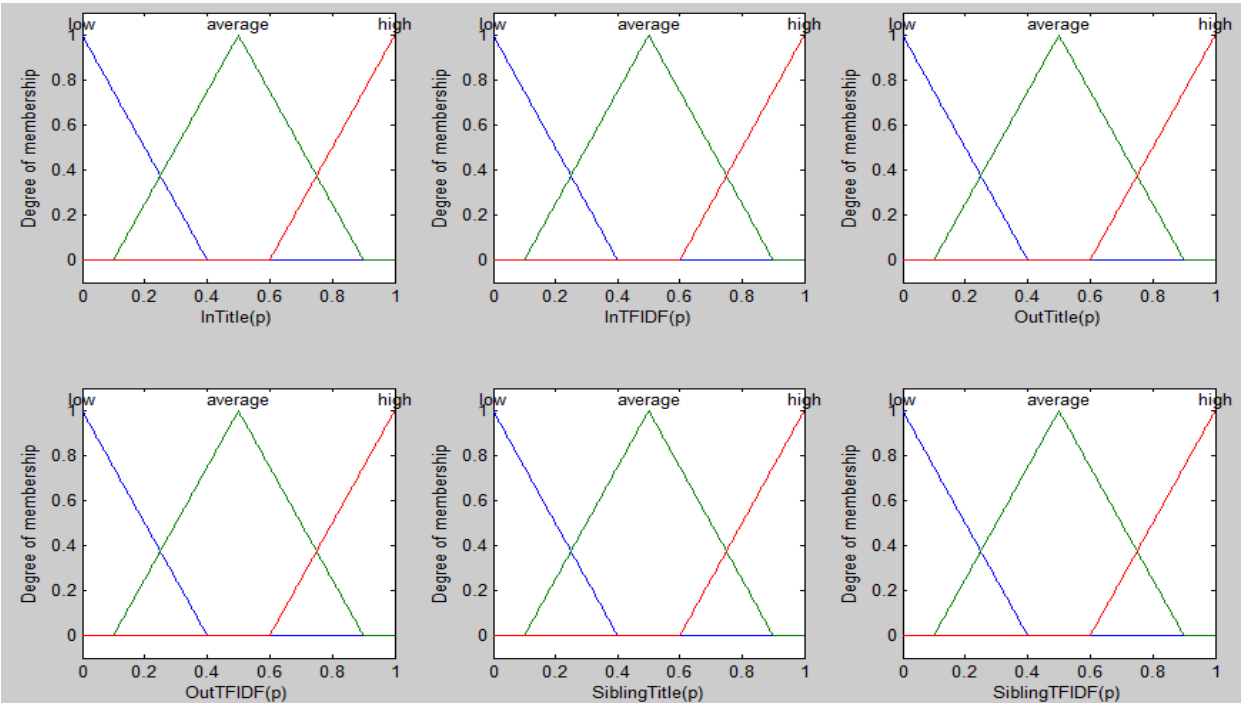

Fig.5. Fuzzy Input Variable for Fuzzy Page Content of Neighbors Value

1. If (InTitle(p) is low) and (InTFIDF(p) is low) and (OutTitle(p) is low) and (OutTFIDF(p) is low) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is low) (1)
2. If (InTitle(p) is average) and (InTFIDF(p) is average) and (OutTitle(p) is low) and (OutTFIDF(p) is low) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is low) (1)
3. If (InTitle(p) is average) and (InTFIDF(p) is average) and (OutTitle(p) is average) and (OutTFIDF(p) is low) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is low) (1)
4. If (InTitle(p) is average) and (InTFIDF(p) is average) and (OutTitle(p) is average) and (OutTFIDF(p) is average) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is average) (1)
5. If (InTitle(p) is average) and (InTFIDF(p) is average) and (OutTitle(p) is average) and (OutTFIDF(p) is average) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is average) (1)
6. If (InTitle(p) is average) and (InTFIDF(p) is average) and (OutTitle(p) is average) and (OutTFIDF(p) is average) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is average) (1)
7. If (InTitle(p) is high) and (InTFIDF(p) is high) and (OutTitle(p) is high) and (OutTFIDF(p) is average) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is high) (1)
8. If (InTitle(p) is high) and (InTFIDF(p) is high) and (OutTitle(p) is high) and (OutTFIDF(p) is high) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is high) (1)
9. If (InTitle(p) is high) and (InTFIDF(p) is high) and (OutTitle(p) is high) and (OutTFIDF(p) is high) and (SiblingTitle(p) is high) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is high) (1)
10. If (InTitle(p) is average) and (InTFIDF(p) is low) and (OutTitle(p) is average) and (OutTFIDF(p) is low) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is low) (1)
11. If (InTitle(p) is average) and (InTFIDF(p) is low) and (OutTitle(p) is low) and (OutTFIDF(p) is average) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is low) (1)
12. If (InTitle(p) is average) and (InTFIDF(p) is low) and (OutTitle(p) is low) and (OutTFIDF(p) is low) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is low) then (fuzzy_Page_content_of_neighbors is low) (1)
13. If (InTitle(p) is average) and (InTFIDF(p) is low) and (OutTitle(p) is low) and (OutTFIDF(p) is low) and (SiblingTitle(p) is low) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is low) (1)
14. If (InTitle(p) is average) and (InTFIDF(p) is low) and (OutTitle(p) is low) and (OutTFIDF(p) is average) and (SiblingTitle(p) is high) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is average) (1)
15. If (InTitle(p) is average) and (InTFIDF(p) is low) and (OutTitle(p) is high) and (OutTFIDF(p) is high) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is average) (1)
16. If (InTitle(p) is average) and (InTFIDF(p) is high) and (OutTitle(p) is high) and (OutTFIDF(p) is high) and (SiblingTitle(p) is average) and (SiblingTFIDF(p) is average) then (fuzzy_Page_content_of_neighbors is high) (1)

Fig.6. Fuzzy Rules for Fuzzy Page Content of Neighbors Value

C. Fuzzy Connectivity (link analysis)

Connectivity feature make a huge impact on category pages are appropriate for a search query. The connectivity feature influenced from Hub(p), Authority(p), PageRank(p), Inlinks(p), Outlinks(p), Anchor(p) factors. How to calculate these factors is given below[18].

- Hub(p) Is determined by Hub score for page p computed by the HITS algorithm
- Authority(p) Is determined by Authority score for page p computed by the HITS algorithm
- PageRank(p) Is determined by PageRank score for page p
- Inlinks(p) Is determined by count of incoming links pointing to p
- Outlinks(p) Is determined by count of outgoing links from p
- Anchor(p) Is determined by count of works in the anchor texts explaining page p discover in the domain lexicon

The mamdani Fuzzy systems to estimate the exact score for Fuzzy Connectivity is shown(Fig. 7).After calculating the above-mentioned factors Input variables are given in the range 0 to 1 after the change of scale. Again we consider Three linguistic variables are low, average, high, with triangular membership functions for the six input variables (Fig. 8). Fuzzy rules is introduced in Fig. 9, in order to the exact amount of fuzzy connectivity score by experts working. As an example rule if Hub(p) is low, Authority(p) is low, PageRank(p) is low, Inlinks(p) is low, Outlinks(p) is low, Anchor(p) is low then Fuzzy connectivity is low.

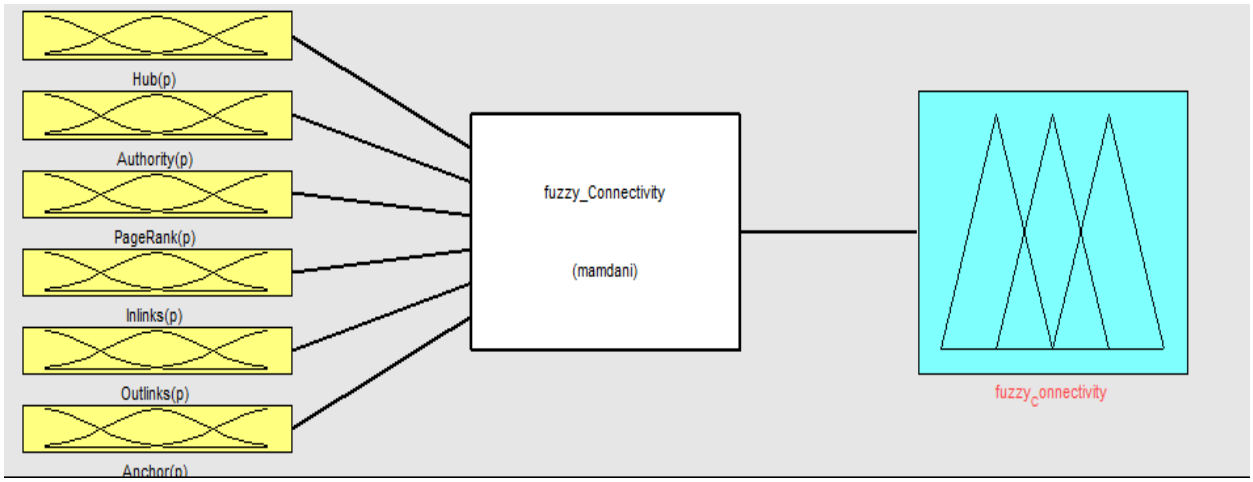

Fig.7. Fuzzy Connectivity Value

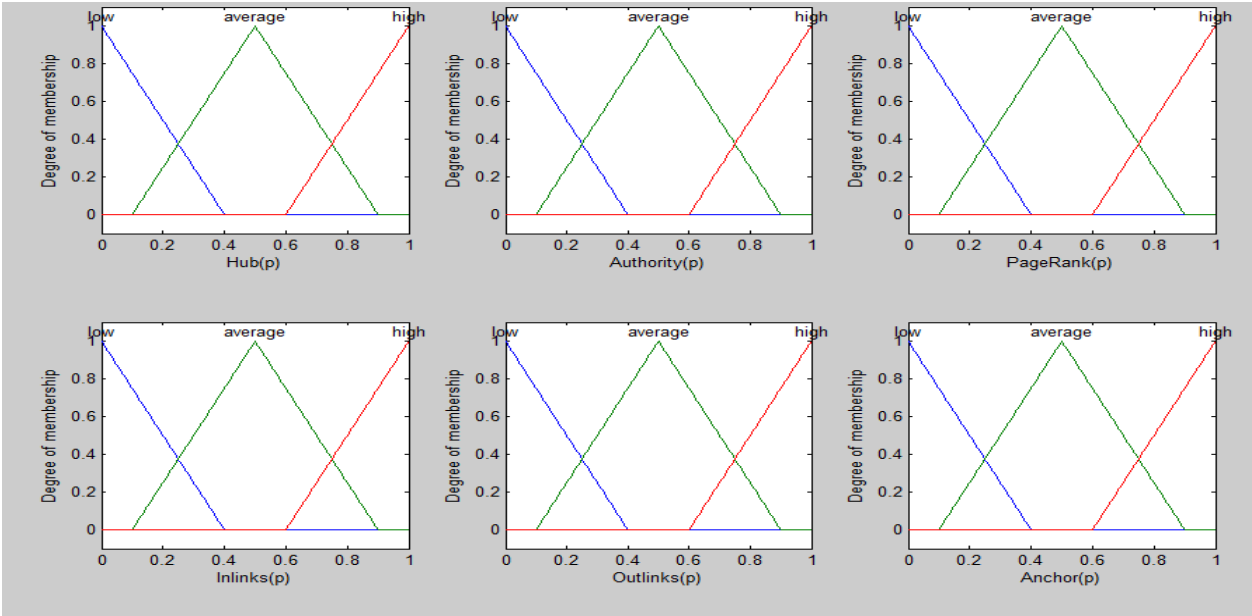

Fig.8. Fuzzy Input Variable for Fuzzy Connectivity Value

1. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is low) (1)
2. If (Hub(p) is average) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is low) (1)
3. If (Hub(p) is high) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
4. If (Hub(p) is low) and (Authority(p) is average) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is low) (1)
5. If (Hub(p) is low) and (Authority(p) is high) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
6. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is average) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is low) (1)
7. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is high) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
8. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is average) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is low) (1)
9. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is high) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
10. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is average) and (Anchor(p) is low) then (fuzzy_Connectivity is low) (1)
11. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is high) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
12. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is average) then (fuzzy_Connectivity is low) (1)
13. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is high) then (fuzzy_Connectivity is average) (1)
14. If (Hub(p) is average) and (Authority(p) is average) and (PageRank(p) is low) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
15. If (Hub(p) is low) and (Authority(p) is average) and (PageRank(p) is average) and (Inlinks(p) is low) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
16. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is average) and (Inlinks(p) is average) and (Outlinks(p) is low) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)
17. If (Hub(p) is low) and (Authority(p) is low) and (PageRank(p) is average) and (Inlinks(p) is low) and (Outlinks(p) is average) and (Anchor(p) is low) then (fuzzy_Connectivity is average) (1)

Fig.9. Fuzzy Rules for Fuzzy Connectivity Value

D. Proposed system with fuzzy sugeno

After the calculation Fuzzy content page, fuzzy Page content of neighbours, fuzzy Connectivity (link analysis) of the fuzzy method, we applied them Sugeno fuzzy systems(Fig. 10).We considered Three linguistic variables are low, average, high, with triangular membership functions For each input three variables (Fig. 11). Fuzzy rules is introduced in Fig. 12, in order to the exact value of fuzzy web pages filter based on sugeno by experts working. As an example rule if Fuzzy content page is average and fuzzy Page content of neighbours is average and fuzzy Connectivity (link analysis) is average then fuzzy relevant web pages is yes.

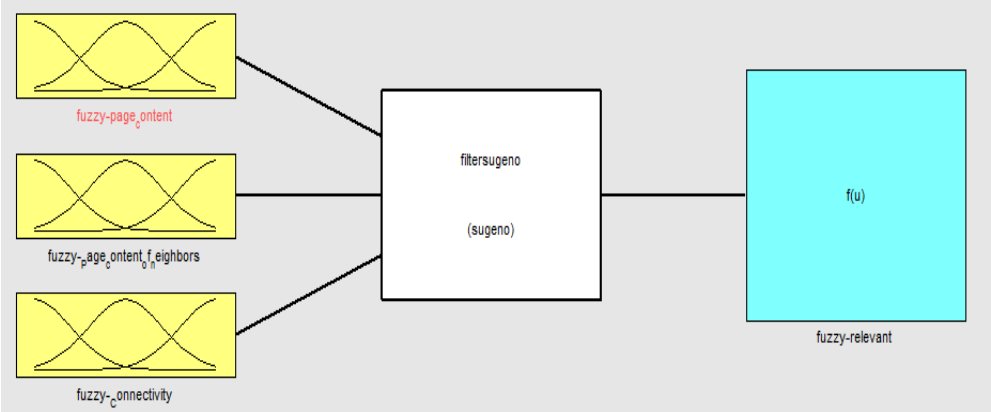

Fig.10. Fuzzy Filter Sugeno System

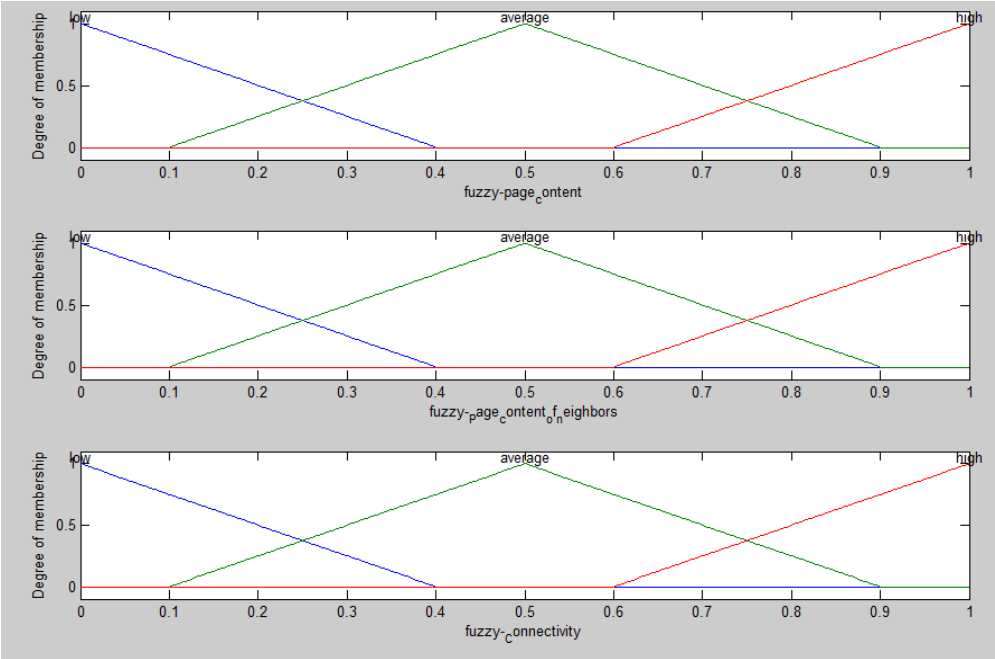

Fig.11. Fuzzy Input Variable for Fuzzy Sugeno Filter

1. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is low) and (fuzzy-_Connectivity is low) then (fuzzy-relevant is no) (1)
2. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is low) and (fuzzy-_Connectivity is average) then (fuzzy-relevant is no) (1)
3. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is low) and (fuzzy-_Connectivity is high) then (fuzzy-relevant is yes) (1)
4. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is average) and (fuzzy-_Connectivity is low) then (fuzzy-relevant is no) (1)
5. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is average) and (fuzzy-_Connectivity is average) then (fuzzy-relevant is yes) (1)
6. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is average) and (fuzzy-_Connectivity is high) then (fuzzy-relevant is yes) (1)
7. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is high) and (fuzzy-_Connectivity is low) then (fuzzy-relevant is yes) (1)
8. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is high) and (fuzzy-_Connectivity is average) then (fuzzy-relevant is yes) (1)
9. If (fuzzy-page_content is low) and (fuzzy-_Page_content_of_neighbors is high) and (fuzzy-_Connectivity is high) then (fuzzy-relevant is yes) (1)
10. If (fuzzy-page_content is average) and (fuzzy-_Page_content_of_neighbors is average) and (fuzzy-_Connectivity is high) then (fuzzy-relevant is yes) (1)
11. If (fuzzy-page_content is average) and (fuzzy-_Page_content_of_neighbors is average) and (fuzzy-_Connectivity is average) then (fuzzy-relevant is yes) (1)
12. If (fuzzy-page_content is average) and (fuzzy-_Page_content_of_neighbors is high) and (fuzzy-_Connectivity is average) then (fuzzy-relevant is yes) (1)
13. If (fuzzy-page_content is high) and (fuzzy-_Page_content_of_neighbors is high) and (fuzzy-_Connectivity is average) then (fuzzy-relevant is yes) (1)
14. If (fuzzy-page_content is high) and (fuzzy-_Page_content_of_neighbors is high) and (fuzzy-_Connectivity is high) then (fuzzy-relevant is yes) (1)
15. If (fuzzy-page_content is average) and (fuzzy-_Page_content_of_neighbors is low) and (fuzzy-_Connectivity is low) then (fuzzy-relevant is no) (1)
16. If (fuzzy-page_content is average) and (fuzzy-_Page_content_of_neighbors is average) and (fuzzy-_Connectivity is low) then (fuzzy-relevant is yes) (1)

Fig.12. Fuzzy Rules for Fuzzy Filter Sugeno System

E. Proposed system with anfis

ANFIS (Adaptive Neuro Fuzzy Inference System) is based sugeno (Jang, Sun & Mizutani, 1997; Jang & Sun,1995). A generic rule in a Sugeno fuzzy pattern has the form If Input 1 = x and Input 2 = y, then Output is z = ax + by + c. Fig. 13 explain the anfis neural network [20,21]. In Fig. 13 first layer are the degree of membership of linguistic variables, The second layer is 'rules layer' .after the linear composition of rules at third layer then specify the degree of belonging to a special class by Sigmund function in layer 4.ANFIS is a type of fuzzy neural network with a learning algorithm based on a set of training data for tuning an available rule base that permit the rule base to reconcile The training data[7].

We have applied the Fuzzy content page, fuzzy Page content of neighbors, fuzzy Connectivity (link analysis) entries to ANFIS the given training data, the related rules is set, and obtain more accurate output (Fig. 14).

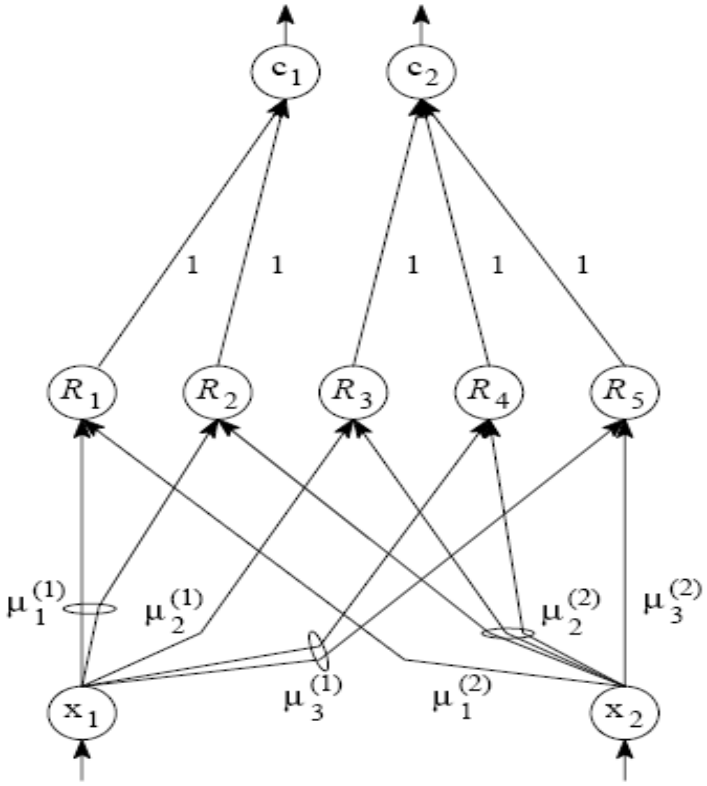

Fig.13. Adaptive Neuro fuzzy Network (anfis)

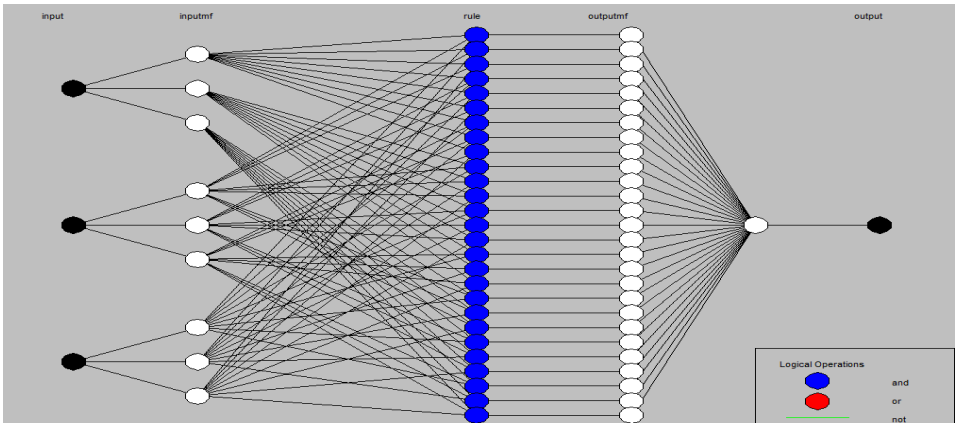

Fig.14. Adaptive Neuro fuzzy Network (anfis) for Web Page Relevant

IV. Experimental Results

We implement and test our proposed system in MATLAB version 7.12 on 1320 web page documents. As shown in Fig. 15 when the title (p) =0.323, TFIDF (p) =0.283, the Fuzzy page content, 0.439 respectively (Fig. 15). Surface View plot for fuzzy page content, title (p), TFIDF (p) is shown in Fig. 16.

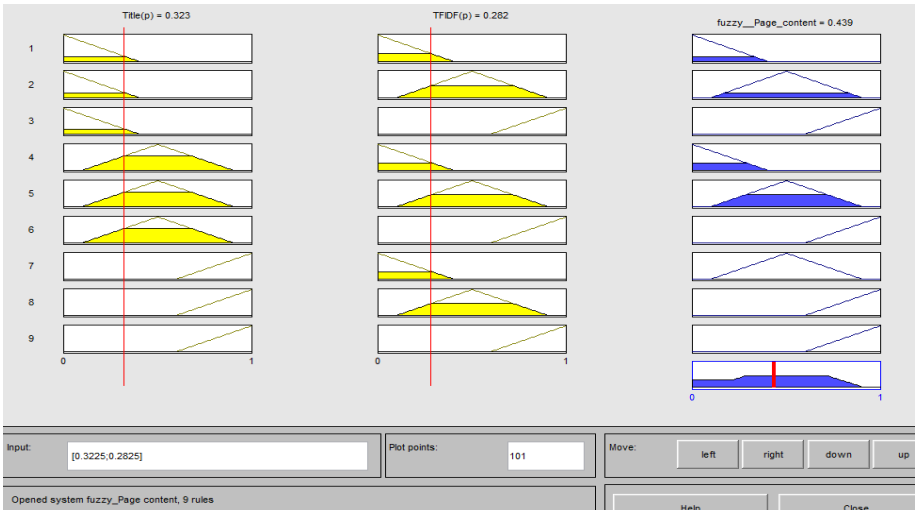

Fig.15. The Exact Amount of Proposed System For Fuzzy Page Content

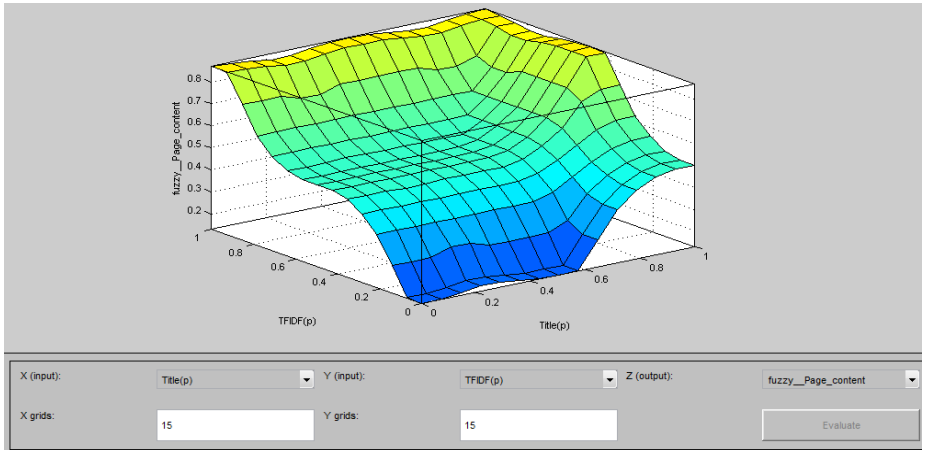

Fig.16. Surface View Plot for Fuzzy Page Content

In proposed system for Fuzzy Page content of neighbours when the InTitle(p) =0.506, InTFIDF(p) =0.529, OutTitle(p)=0.238, OutTFIDF(p) =0.378, SiblingTitle(p)=0.343, SiblingTFIDF(p) =0.335 then Fuzzy Page content of neighbours =0.488(Fig. 17). Surface View plot Is shown in Fig. 18 for fuzzy page content of neighbours, InTitle(p), OutTitle(p).

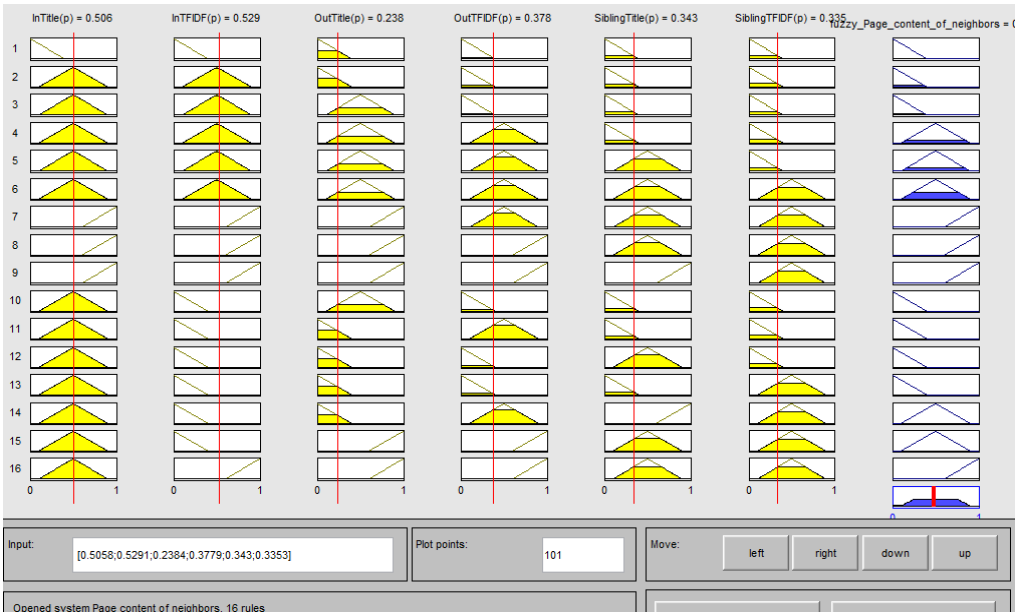

Fig.17. The Exact Amount of Proposed System for Fuzzy Page Content of Neighbors

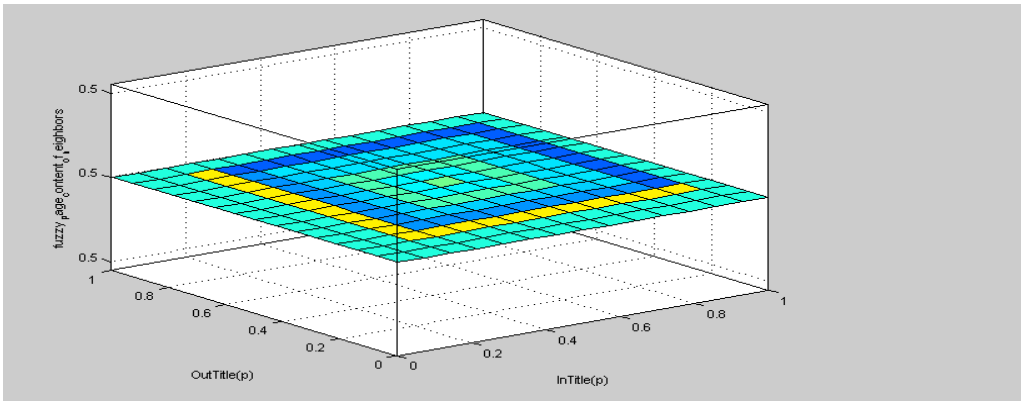

Fig.18. Surface View Plot for Fuzzy Page Content of Neighbors

Fig.19. The Exact Amount Of Proposed System For Fuzzy Page Connectivity

In proposed system for Fuzzy connectivity the Hub(p) =0.192, Authority(p) 0.773, PageRank(p)=0.203, Inlinks(p)=0.297, Outlinks(p) =0.238, Anchor(p)=0.253 then Fuzzy connectivity is 0.491(Fig. 19). Surface View plot Is shown in Fig. 20 for fuzzy connectivity, Hub(p), Inlinks(p).

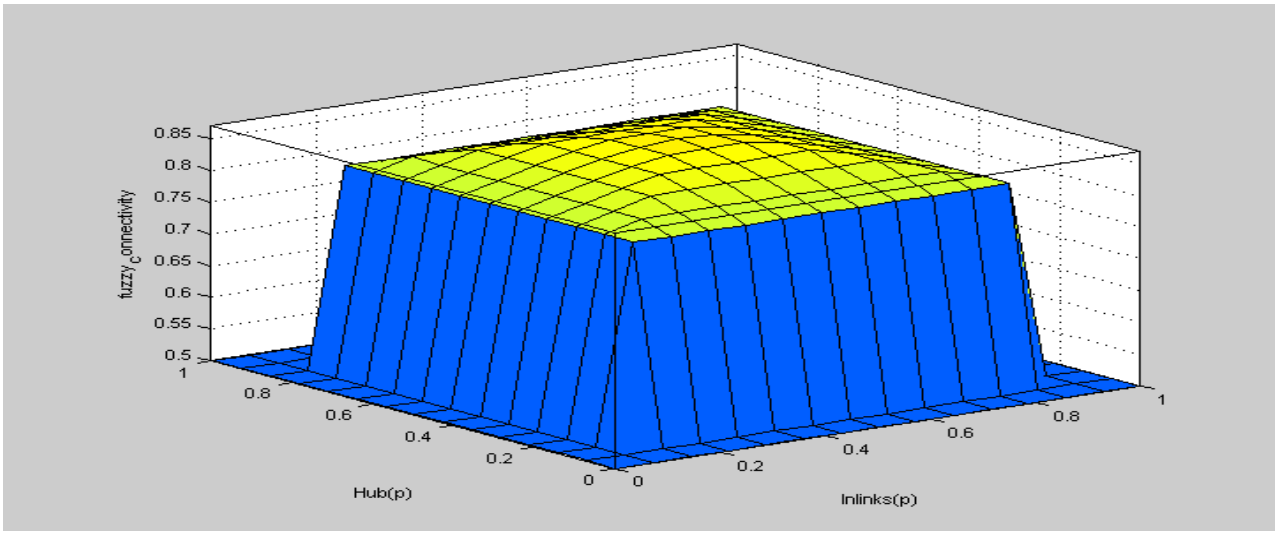

Fig.20. Surface View Plot for Fuzzy Connectivity

In Anfis proposed system was considered a database of 1320 web page documents, In order to train and test the fuzzy neural network. After calculate 3 Features described above for 1320 web page documents, 1100 web page Was considered for train Anfis(Fig.21) and 220 web page Was Allocated to Test system. After setting network parameters to generate fis =grid partition, train fis epochs=30 Rmse(Root mean squared errors) for training data Obtained 0.088(Fig. 22).

After completing the process the training adaptive fuzzy neural network, fuzzy input variables were calibrated (Fig.s 23,24,25), and the number 220 web page was to test the system Rmse error value of 0.079 was obtained for test data (Fig.26).In the Fig.s 27,28 As shown Predicted, actual outputs and errors for test data with propose anfis system.

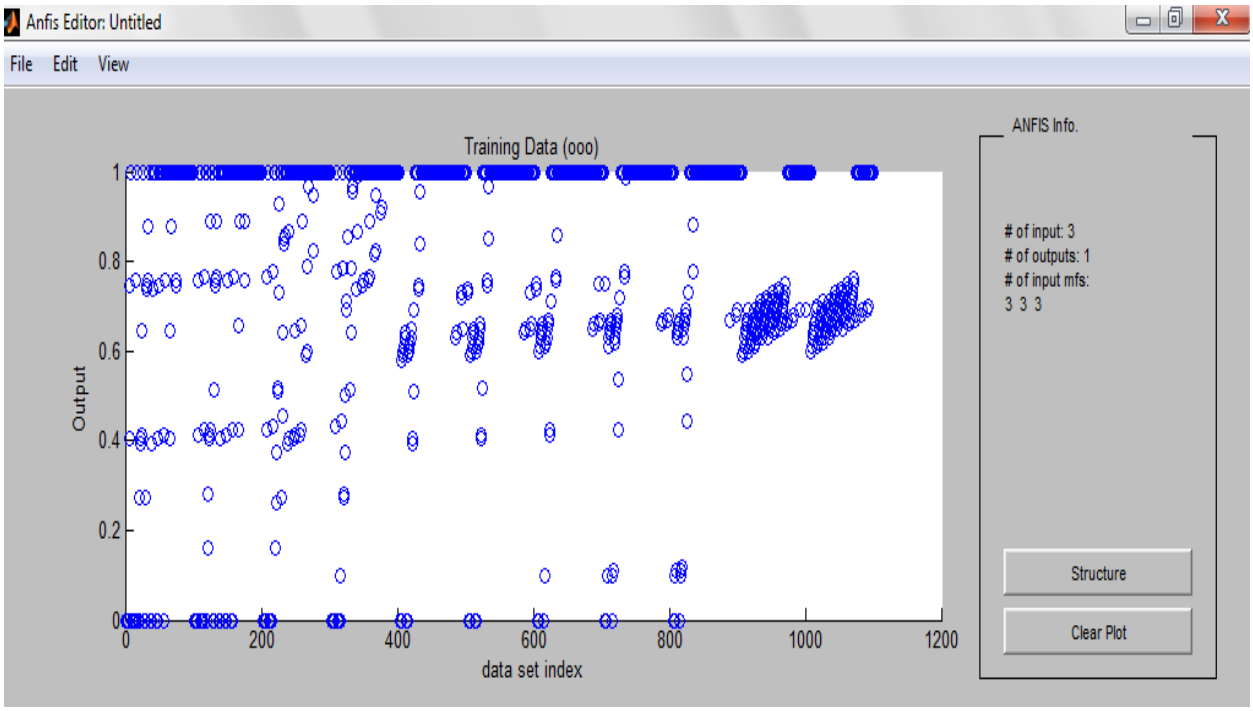

Fig.21. Train Data for Proposed Anfis System

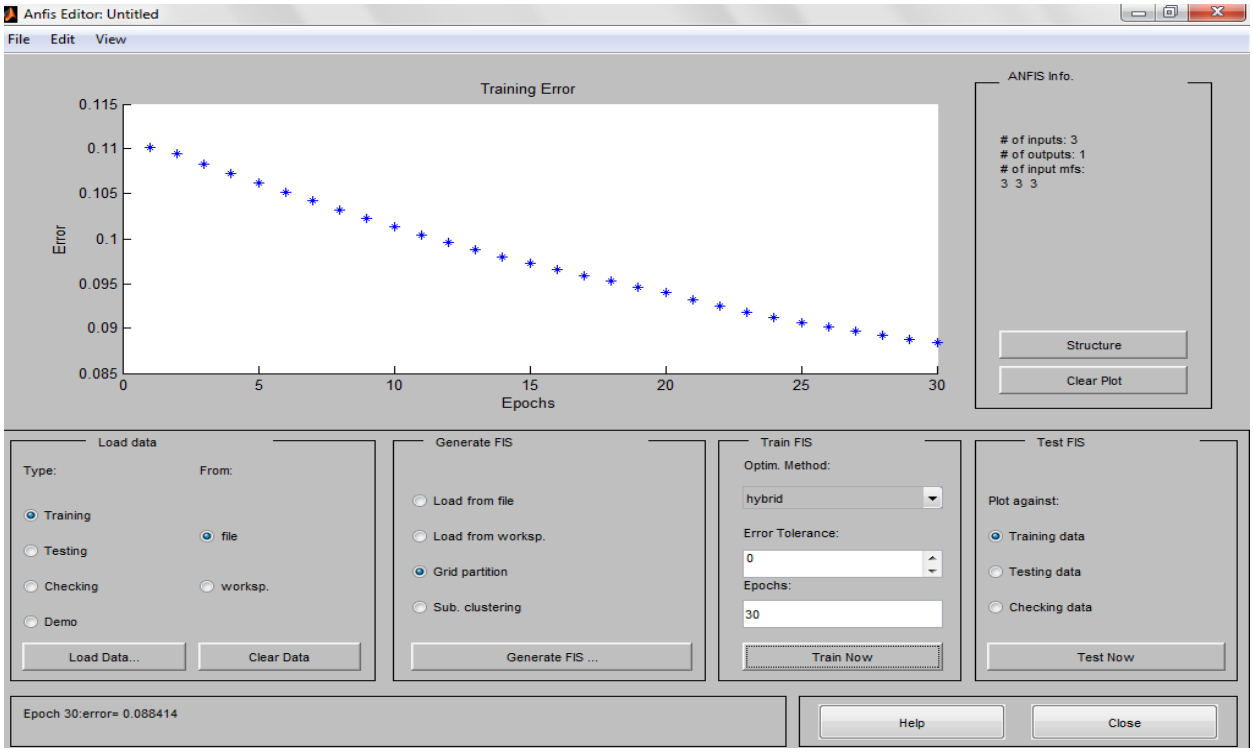

Fig.22. Calculate RMSE for Train Data with Proposed Anfis System

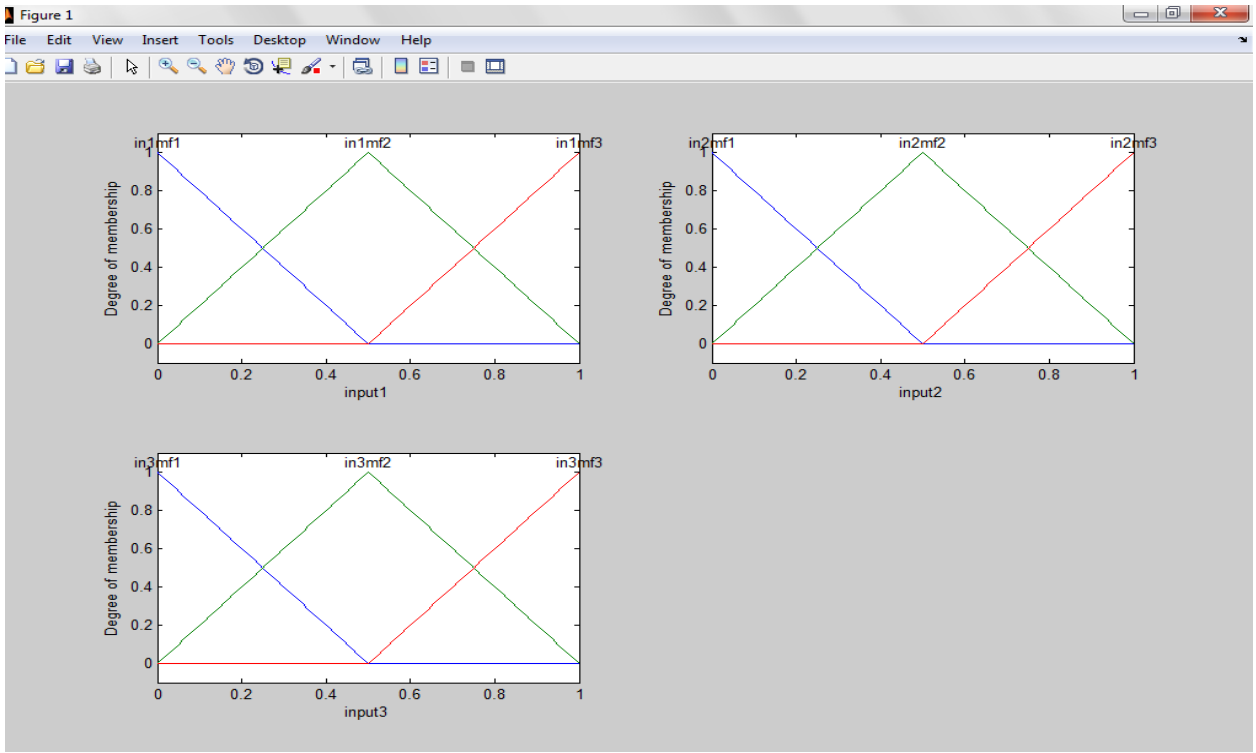

Fig.23. Input Variable Fuzzy Membership for Proposed Anfis Filter System before Calibrate

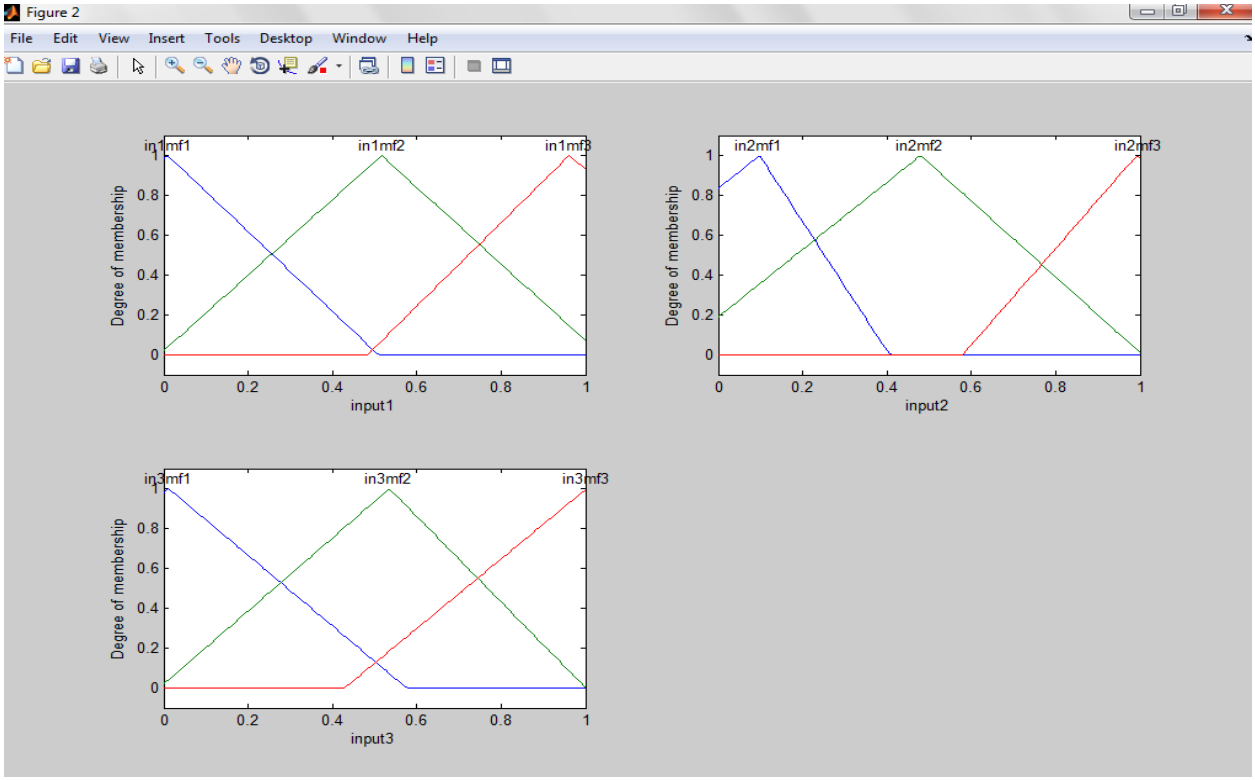

Fig.24. Input Variable Fuzzy Membership for Proposed Anfis Filter System after Calibrate

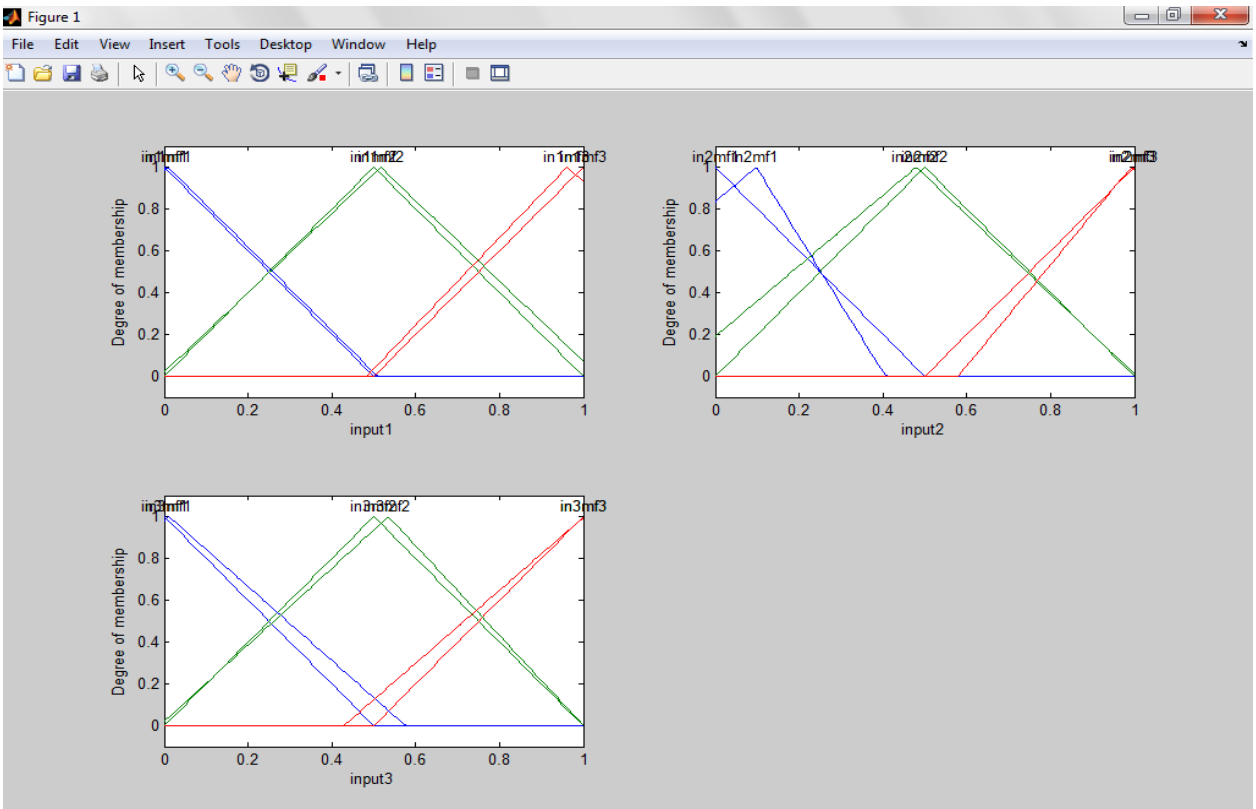

Fig.25. Input Variable Fuzzy Membership for Proposed Anfis System Before, After Calibrate

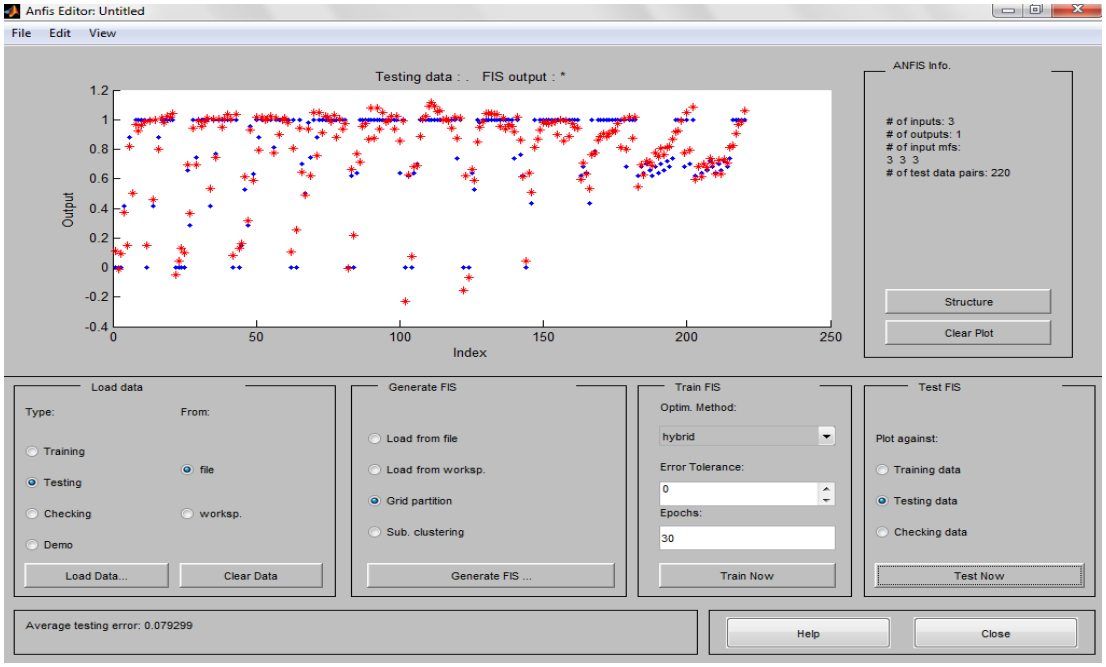

Fig.26. Calculate RMSE for test data with proposed Anfis system

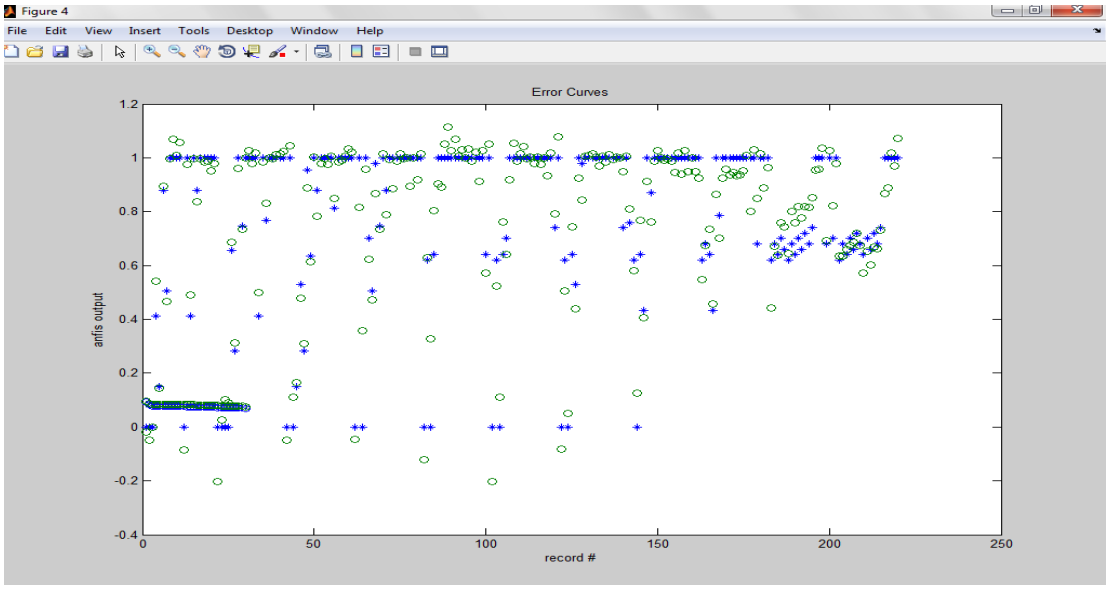

Fig.27. Predicted and Actual Outputs for Anfis System

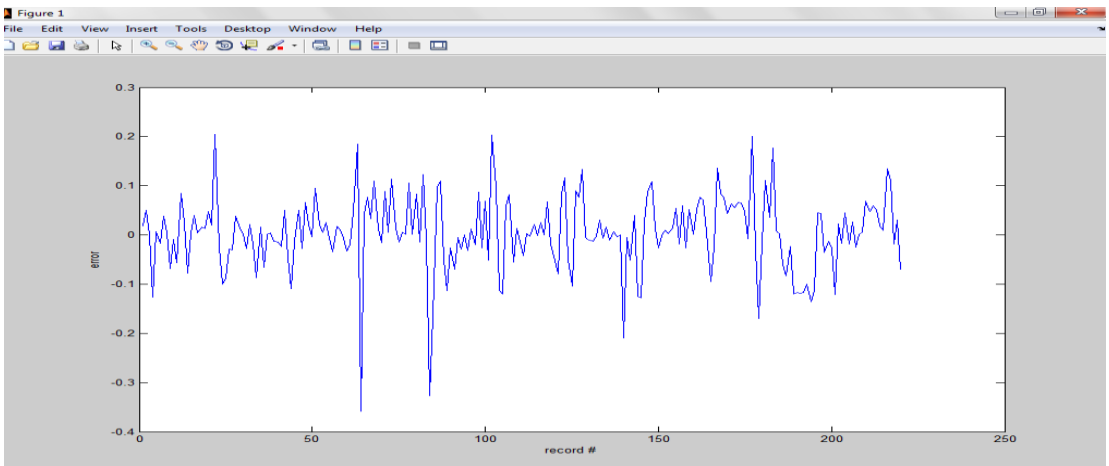

Fig.28. Errors for Outputs with Anfis System

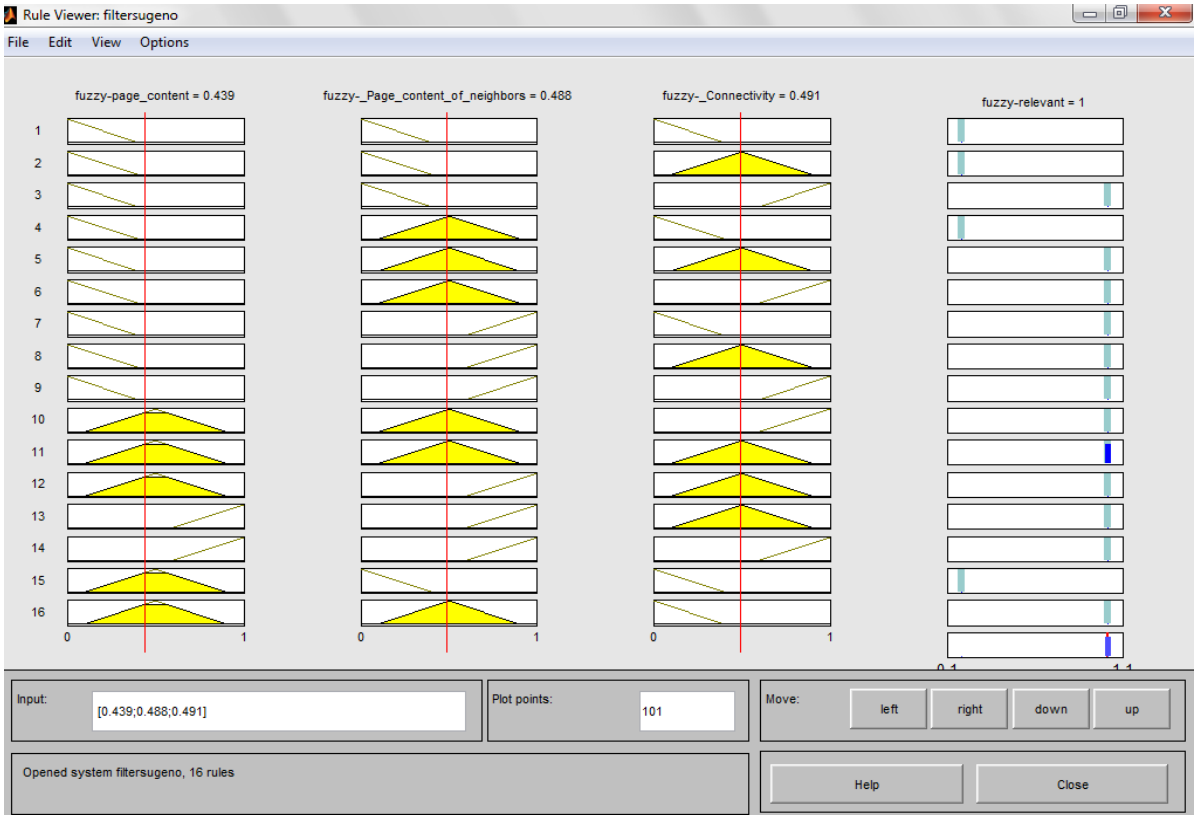

Fig.29. The exact amount of proposed system for fuzzy page relevant

In the fuzzy proposed system using Sugeno per fuzzy page content=0.439, fuzzy page of neighbours=0.488, fuzzy connectivity=0.491 value of fuzzy relevant 1 respectively (Fig. 29). Surface View plot for fuzzy relevant, fuzzy page connectivity, fuzzy page content in Fig. 30. After applying 220 data pages corresponding test data error value of 0.861 was calculated (Fig.31) but anfis 0.079

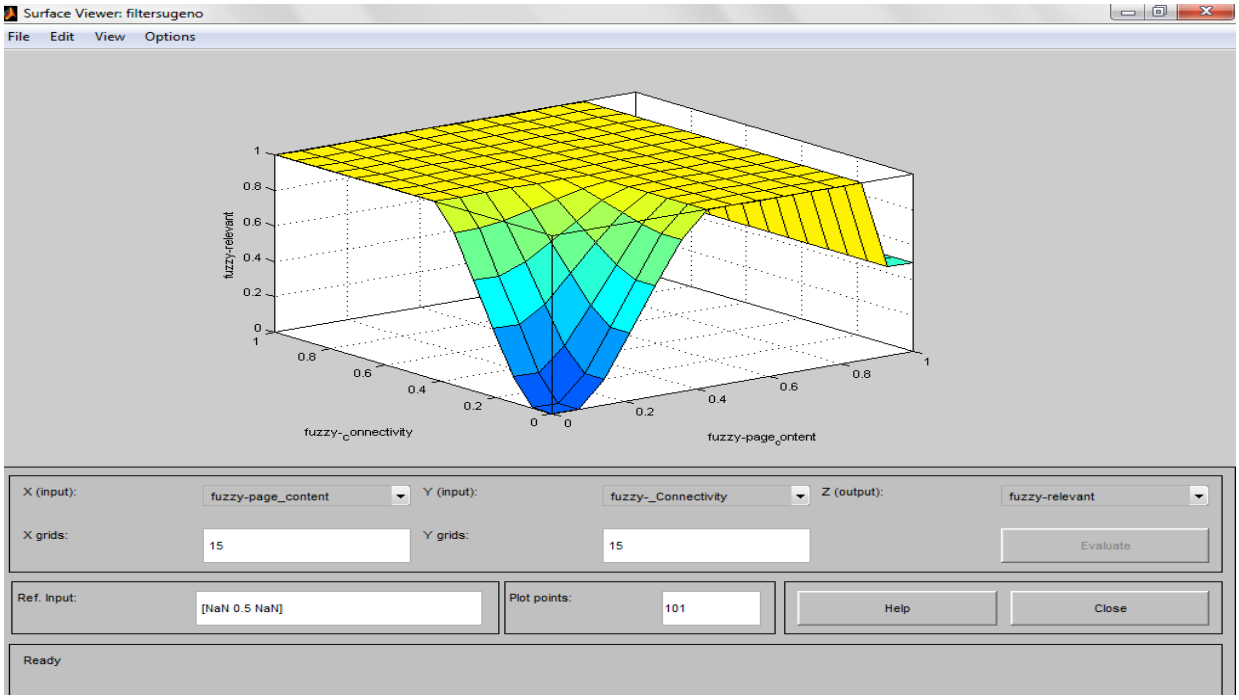

Fig.30. Surface View Plot for Fuzzy Relevant, Fuzzy Page Connectivity, Fuzzy Page Content

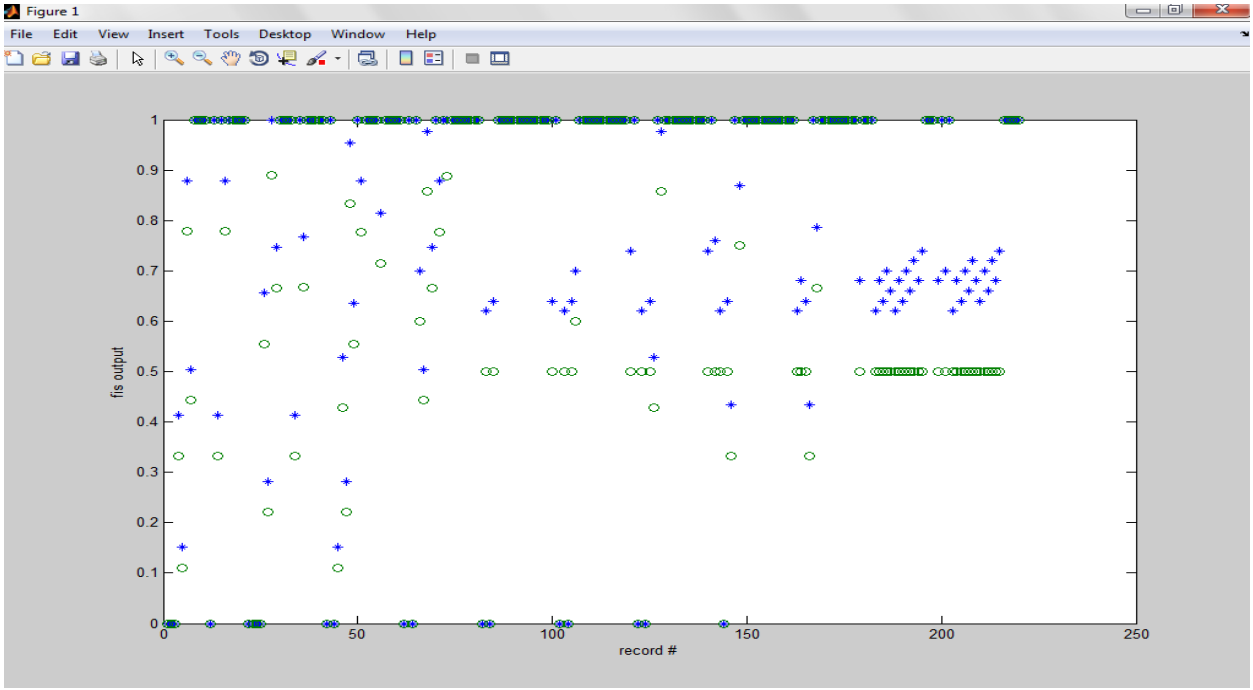

Fig.31. Predicted and Actual Outputs for Fuzzy Sugeno Relevant System

V. Conclusion And Future Directions

In this paper we show a method to retrieve Web pages, according to a query search engine. We have considered content, pages neighbours and links features. After the phase fuzzifying, we present, two Sugeno fuzzy systems and neural fuzzy systems Anfis for web page retrieval. the accuracy of the proposed system can be enhanced by Increasing the number of training data. Given that Anfis system error was less (Rmse=0.079), We believe that our proposed system can be used as an expert system to retrieve Web pages relevant to the query used by the search engines and web content management .

Another method of classification such as svm, other properties related Web pages, variety of inputs and changes in the membership function of the membership functions can be considered as a future research work.

References

[1] Chaffee, Jason, and Susan Gauch. "Personal ontologies for web navigation." In Proceedings of the ninth international conference on Information and knowledge management, pp. 227-234. ACM, 2000.

[2] Cho, Junghoo, Hector Garcia-Molina, and Lawrence Page. "Efficient crawling through URL ordering." Computer Networks and ISDN Systems 30, no. 1 (1998): 161-172.

[3] Baujard, O., V. Baujard, S. Aurel, C. Boyer, and R. D. Appel. "Trends in medical information retrieval on internet." Computers in Biology and Medicine 28, no. 5 (1998): 589-601.

[4] McCallumzy, Andrew, and Kamal Nigamy. "Text classification by bootstrapping with keywords, EM and shrinkage." (1999).

[5] Scarselli, Franco, Lucia Di Noi, and Markus Hagenbuchner. "Web Spam Detection by Neural Networks for Structures."

[6] Beel, Joeran, and Bela Gipp. "Academic search engine spam and Google Scholar's resilience against it." Journal of electronic publishing 13, no. 3 (2010).

[7] Iraji, Mohammad Saber, and Homayun Motameni. "Object Oriented Software Effort Estimate with Adaptive Neuro Fuzzy use Case Size Point (ANFUSP)." International Journal of Intelligent Systems and Applications (IJISA) 4, no. 6 (2012): 14.

[8] M. Chau, H. Chen, Comparison of three vertical search spiders,IEEE Computer 36 (5) (2003a) 56–62.

[9] Li, Rongmei. "Improving Web Page Retrieval using Search Context from Clicked Domain Names." In Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on, pp. 393-397. IEEE, 2009.

[10] Dubey, Hema, and B. N. Roy. "An Improved Page Rank Algorithm based on Optimized Normalization Technique." (2011).

[11] Bhamidipati, Narayan L., and Sankar K. Pal. "Comparing scores intended for ranking." Knowledge and Data Engineering, IEEE Transactions on 21, no. 1 (2009): 21-34.

[12] Sharma, Dilip Kumar, and A. K. Sharma. "A Comparative Analysis of Web Page Ranking Algorithms." International Journal on Computer Science & Engineering (2010).

[13] Minnie, D., and S. Srinivasan. "Intelligent Search Engine algorithms on indexing and searching of text documents using text representation." In Recent Trends in Information Systems (ReTIS), 2011 International Conference on, pp. 121-125. IEEE, 2011.

[14] Qiu, Zhanzi, Matthias Hemmje, and Erich J. Neuhold. "Using link types in Web page ranking and filtering." In Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on, vol. 1, pp. 311-320. IEEE, 2001.

[15] Chau, Michael, and Hsinchun Chen. "A machine learning approach to web page filtering using content and structure analysis." Decision Support Systems 44, no. 2 (2008): 482- 494.

[16] Khokale, Rahul Shankar, and Mohd Atique. "Web Based Information Retrieval using Fuzzy Logic."
[17] Zadeh, Lotfi A. "Fuzzy logic= computing with words." Fuzzy Systems, IEEE Transactions on 4, no. 2 (1996): 103-111.
[18] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30, no. 1 (1998): 107-117.
[19] Chakrabarti, Soumen, Byron Dom, and Piotr Indyk. "Enhanced hypertext categorization using hyperlinks." In ACM SIGMOD Record, vol. 27, no. 2, pp. 307-318. ACM, 1998.
[20] Mohammad Saber Iraji,Majid Aboutalebi,Naghi. R. Seyedaghaee,Azam Tosinia,"Students Classification With Adaptive Neuro Fuzzy", IJMECS, vol.4, no.7, pp.42-49, 2012.
[21] Ghosh, Ashish, B. Uma Shankar, and Saroj K. Meher. "A novel approach to neuro-fuzzy classification." Neural Networks 22, no. 1 (2009): 100-109.

Authors' Profiles

Mohammad Saber Iraji received B.Sc in Computer Software engineering from Shomal university, Iran, Amol ; M.Sc1in industrial engineering (system management and productivity) from Iran, Tehran and M.Sc2 in Computer Science. Currently, he is engaged in research and teaching on Computer Graphics, Image Processing, Fuzzy and Artificial Intelligent, Data Mining, Software engineering and he is Faculty Member of Department of Computer Engineering and Information Technology, Payame Noor University, I.R. of Iran.

Hakimeh Maghamnia is Computer software engineering Graduated from Department of Computer Engineering and Information Technology, Payame Noor University,I.R. of Iran.

Marzieh Iraji is Information technolog engineering Graduated from Department of Computer Engineering and Information Technology, University College of Rouzbahan, Sari, Iran.

How to cite this paper: Mohammad Saber Iraji, Hakimeh Maghamnia, Marzieh Iraji,"Web Pages Retrieval with Adaptive Neuro Fuzzy System based on Content and Structure", IJMECS, vol.7, no.8, pp.69-84, 2015.DOI: 10.5815/ijmecs.2015.08.08

