The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

Web Based Information Retrieval using Fuzzy Logic

Rahul Shankar Khokale
Assistant Professor,
Department of Computer Science & Engineering,
PIGCE, NAGPUR, India
softrahul@gmail.com

Dr. Mohd. Atique
Associate Professor,
PGDCSE, SGBAU,
AMRAVATI, India
mohd.atique@gmail.com

***Abstract*—Information retrieval on the internet is necessity of today's quintessential technocrats. Enormous information is readily available on the internet. Information retrieval is the key application of internet, as it provides knowledge to the knowledge seekers. The volume of data on the internet is very large, and to fetch most appropriate and relevant information are the challenges in WBIR (Web Based Information Retrieval). Many times, internet applications need to deal with large amount of data collected from non-technical users and is imprecise and incomplete. In this paper, Web based information retrieval based on Fuzzy logic is presented. In the proposed algorithm we are determining Page Scores, and by using HITS algorithm, Hub Score and Authority Score computed. The Fuzzy Inference System consists of three inputs such as Page Rank Score, Normalized Hub Score, Normalized Authority Score, and two outputs which are Relevant Document and Non-relevant Document. User query which can be vague or imprecise will be analyzed by using fuzzy inference rules and the optimum query will be generated for web crawlers so as to produce the desired web documents effectively and efficiently. The performance can be evaluated on the basis of precision and recall parameters.**

Keywords-Information Retrieval; Fuzzy Logic, Page-Rank Algorithm, HITS Algorithm, Web Crawlers, Precision;Recall

I. INTRODUCTION

Search on the web is a daily activity for many people throughout the world. Search and communication are most popular uses of the computer. The field of computer science that is most involved with research and development for search is information retrieval. Or the process of searching specific information among a large number of information items is known as Information Retrieval (IR). "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information." (Salton,1968). Information Retrieval tasks include:

- Ad-hoc search
- Filtering
- Classification
- Question and Answering

Types of information items are documents, Web pages, online catalogs, structured records, multimedia objects.
Web Based Information Retrieval (WBIR) means the process of searching for relevant documents or information among the large number of documents on internet. The primary purpose of establishing an information retrieval system lies in assisting the users to efficiently acquire desired information. Users of IR systems expect to find the most relevant items to a certain query. The computing parameters such as recall and precision are used for effectiveness appraisal of these systems [1]. Generally, an information retrieval system does not present an ideal behaviour. Users often receive large result sets, and they have to spend a considerable time to find these items which are really relevant to their initial queries. Indeed, this kind of searching information will neglect relevant documents that do not contain the index terms which are specified in the user's queries. The architecture of information retrieval system is shown in figure 1.

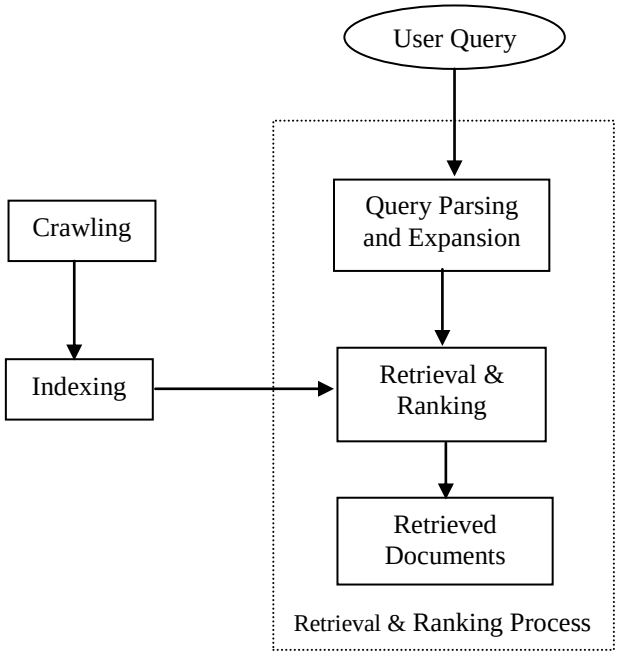

Figure 1 : Architecture of IR system

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

1.1 Information Retrieval on the Web

Retrieving information from the web can prove to be difficult because of the size and abstractness of data contained on the web. Web retrieval is made increasingly difficult when adding in factors such as word ambiguity (where a single word can take on multiple meanings), and the large amount of typographical errors contained within web information. It is estimated that one in every two-hundred words, on an average web site, will contain a textual error

1.2 Dimensions of Information Retrieval

Information Retrieval on web is more than just text, and more than just web search although these are central. People are doing IR work with different media, different types of search applications, and different tasks. New applications increasingly involve new media e.g., video, photos, music, speech. Like text, content is difficult to describe and compare text may be used to represent them (e.g. tags).

Table 1 : Dimensions of IR

Content	Applications	Tasks
Text	Web search	Ad-hoc search
Images	Vertical search	Filtering
Videos	Enterprise search	Classification
Scanned Docs.	Desktop search	Question
Audio	Forum Search	Answering
Music	P2P search	
	Literature search	

1.3 Issues with Information Retrieval

There are several key issues involving information retrieval. These issues are relevance, evaluation, and information needs. However, these are not the only issues involving information retrieval. Other issues such as performance, scalability and occurrences of paging update are other common information retrieval issues. Relevance is the relational value of a given user query to the documents within the database. Relevance of a document is normally based on a document ranking algorithm. These algorithms define how relevant a document is to a user query by using functions that define relations between the query given and the documents collected in the index. The evaluation of the feedback given by the information retrieval system is another issue with information retrieval. The behaviour of the system may not meet the expectations of the user or the documents returned from the system may not all be relevant to a query. Depending on the system and the user, the results of a query should be in a format that most fits the data being searched and returned. Information needs is how the user interacts with the information retrieval system. The data within the system should be able to be accessed easily and in a way that is convenient to the user. Retrieving too much information might be inconvenient in certain systems, also in other systems not returning all relevant information may be unacceptable.

1.4 Fuzzy Logic Systems

The theory of Fuzzy Logic was first raised by the mathematician Lotfi A. Zadeh in 1965. This theory is a result of the insufficiency of Boolean Algebra to many problems of the real world. As most of the information in the real world is imprecise, and one of humans' greatest abilities is to effectively process imprecise and "'fuzzy" information. According to the Oxford English Dictionary, the word Fuzzy is defined as blurred, indistinct, imprecisely defined, confused or vague. Fuzzy systems are knowledge based or rule based systems. The heart of a fuzzy system is a knowledge base consisting of the so called fuzzy IF-THEN rules. A fuzzy IF-THEN rule is an IF-THEN statement in which some words are characterized by continuous membership functions. Fuzzy logic is conceptually easy to understand. The mathematical concepts behind fuzzy reasoning are very simple. What makes fuzzy nice is the "naturalness" of its approach and not its far-reaching complexity. Fuzzy logic is flexible. With any given system, it's easy to massage it or layer more functionality on top of it without starting again from scratch. Fuzzy logic is tolerant of imprecise data. Everything is imprecise if you look closely enough, but more than that, most things are imprecise even on careful inspection. Fuzzy reasoning builds this understanding into the process rather than tacking it onto the end. Fuzzy logic can model nonlinear functions of arbitrary complexity. A fuzzy system can be created to match any set of input-output data.

II. RELATED WORK

Fuzzy Logic deals with uncertainty or imprecision. Fuzzy Logic can be a best approach suited for information retrieval because, most of the time, user query is vague or imprecisely written. Maryam Hourali et al. presented an Intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology. They have proposed a novel approach for fuzzy ontology generation with two uncertainty degrees. By implementing linguistic variables, uncertainty level in domain's concepts (Software Maintenance Engineering (SME) domain) has been modeled, and ontology relations have been modeled by fuzzy theory consequently. Then, we combined these uncertain models and proposed a new ontology with two degrees of uncertainty both in concept expression and relation expression. The generated fuzzy ontology was implemented for expansion of initial user's queries in SME domain [1]. Parry has implemented fuzzy ontology for information retrieval which is focused on

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

medical documents retrieval. This ontology has fuzzy values in its relations[2]. Zhai et al. presented a fuzzy ontology for semantic information retrieval in e-commerce domain, and a semantic query expansion method is used for this purpose. Their framework includes three parts: concepts, properties of concepts, and values of properties in which property value can be either standard data types or linguistic values of fuzzy concepts [3]. They also implemented fuzzy ontology for semantic information retrieval in supply chain management, traffic information retrieval, and intelligent transportation systems fields [4–5]. Leite and Ricarte presented a framework to encode a geographic knowledge base composed of multiple-related ontologies whose relationships were expressed as fuzzy relations.[6] An ontology-based spatial query expansion method was proposed by Fu et al., which considered a geographical ontology to expand geographic terms. Various factors are taken into account to support intelligent expansion of a spatial query, including types of spatial terms as encoded in the geographical ontology, types of non spatial terms as encoded in the domain ontology, as well as the semantics of the spatial relationships and their context of use. [7]. Bratsas et al. used a fuzzy query expansion and a fuzzy thesaurus to solve the Medical Computational Problem (MCP). In the experiments, the system was capable of retrieving the same MCP for distinct descriptions of the same problem. The system uses a unique fuzzy thesaurus for query expansion[8]. Sunhong Kim and Sangyong Han proposed the approach for inferring trust between users in the web-based social network using fuzzy logic. They have suggested a method of inferring trust between users in web-based social network information retrieval system using fuzzy logic for reliable information retrieval and user's trust detection[9]. M.J. Martín-bautista et al., present a study of the role of user profiles and fuzzy logic in web retrieval processes. Flexibility for user interaction and for adaptation in profile construction becomes an important issue. They focus their study on user profiles, including creation, modification, storage, clustering and interpretation[10]. Padmini Srinivasan and Donald H. Kraft presented the concept of vocabulary mining deals with using the domain vocabulary, perhaps a controlled vocabulary, with the goal of improving the performance of an information retrieval system in response to a user query. They have carried out research on Vocabulary Mining for Information Retrieval my means of Extensions and Generalizations of Combining Fuzzy Sets and Rough Sets[11].

III. PROPOSED WORK

In this proposed work, Web-Based Information Retrieval (WBIR) system based on Fuzzy logic is presented. We have used Page-Rank algorithm and HITS algorithm to generate inputs to the proposed FIS. The Fuzzy Inference System consists of three inputs such as Page Rank Score, Normalized Hub Score, Normalized Authority Score, and two outputs which are Relevant Document and Non-relevant Document.

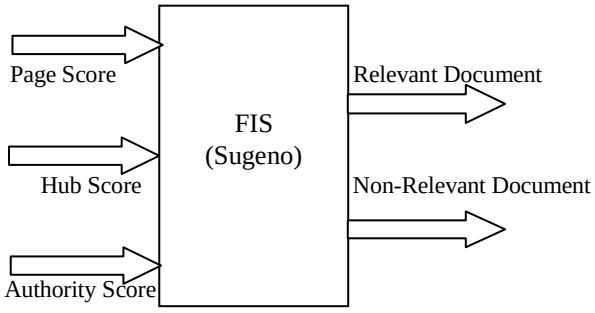

Figure 2 : Fuzzy Inference System model for WBIR

The Objectives of using Fuzzy Logic for Web-Based Information Retrieval are given below:

- To produce the most relevant documents as per user interest.
- To enhance the effectiveness of document search on the internet.

2.1 Algorithm

1) Initialize the PageScore() =0;
2) Enter the User Query for which documents are to be searched;
3) Let W= {w_1, w_2,.....,w_N } be the number of words belong to User Query.
4) Web crawler starts with a list of URLs to visit called *Seeds* and downloads the respective page from the Internet at the given URL.
5) A Web crawler parses through the downloaded page and retrieves the links to other pages. Each link in the page is defined with an HTML anchor tag.
6) Each link is added to the list of URLs to visit called the *Queue or Frontier.*

7) Calculate the Page Score as follows:

```
For each w_i ϵ W,  do
    if  w_i  is present in the document;
        PageScore()=PageScore() + 1;
            else
            PageScore() is unchanged;
    end
```

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

8) Use HITS algorithm to compute Hub Score and Authority for each retrieved page.
9) Apply PageScore(), Hub Score and Authority Score as inputs to the proposed FIS.
10) Simulate the FIS model on the given inputs and find whether the given document is relevant to non-relevant
11) Display all the relevant documents.

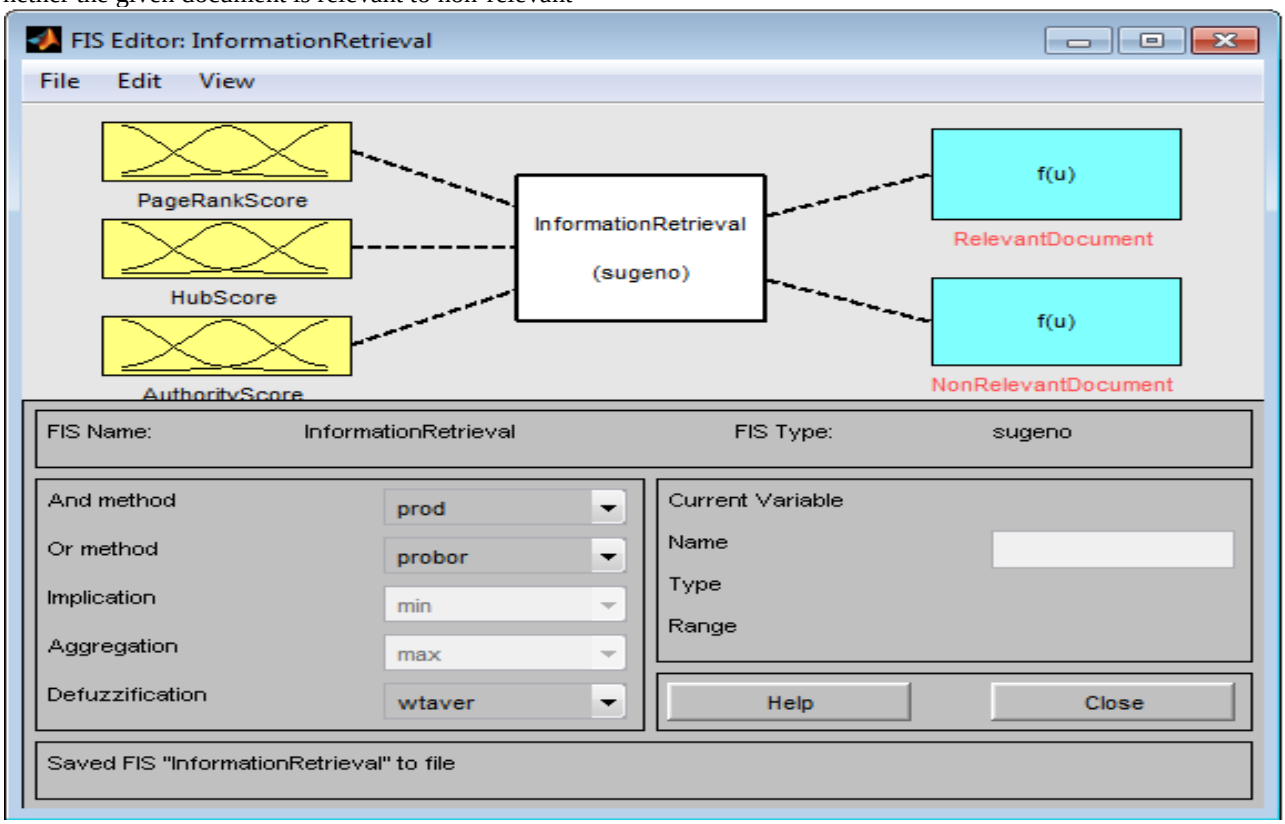

Figure 3 : Fuzzy Inference System for Information Retrieval

The Fuzzy Inference System is modeled and implemented in MATLAB as shown in Fig.3.

The membership functions for the three inputs variables: Page Score, Hub Score and Authority Score are as follows

Table 2 : Membership functions for input variables

Input	0 - 0.3 (MF)	0.2 – 0.7 (MF)	0.6 – 1.0 (MF)
Page Score	low	medium	high
Hub Score	bad	average	good
Authority Score	bad	average	good

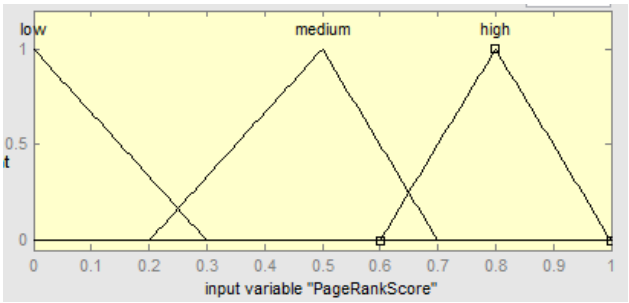

Figure 4: Membership function for Page Rank Score

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

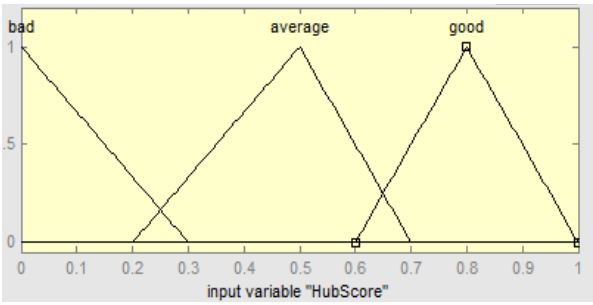

Figure 5: Membership function for Hub Score

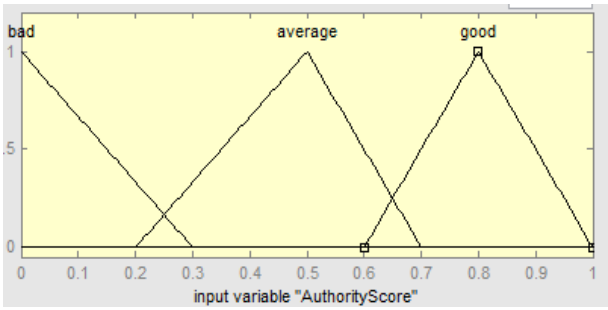

Figure 6: Membership function for Authority Score

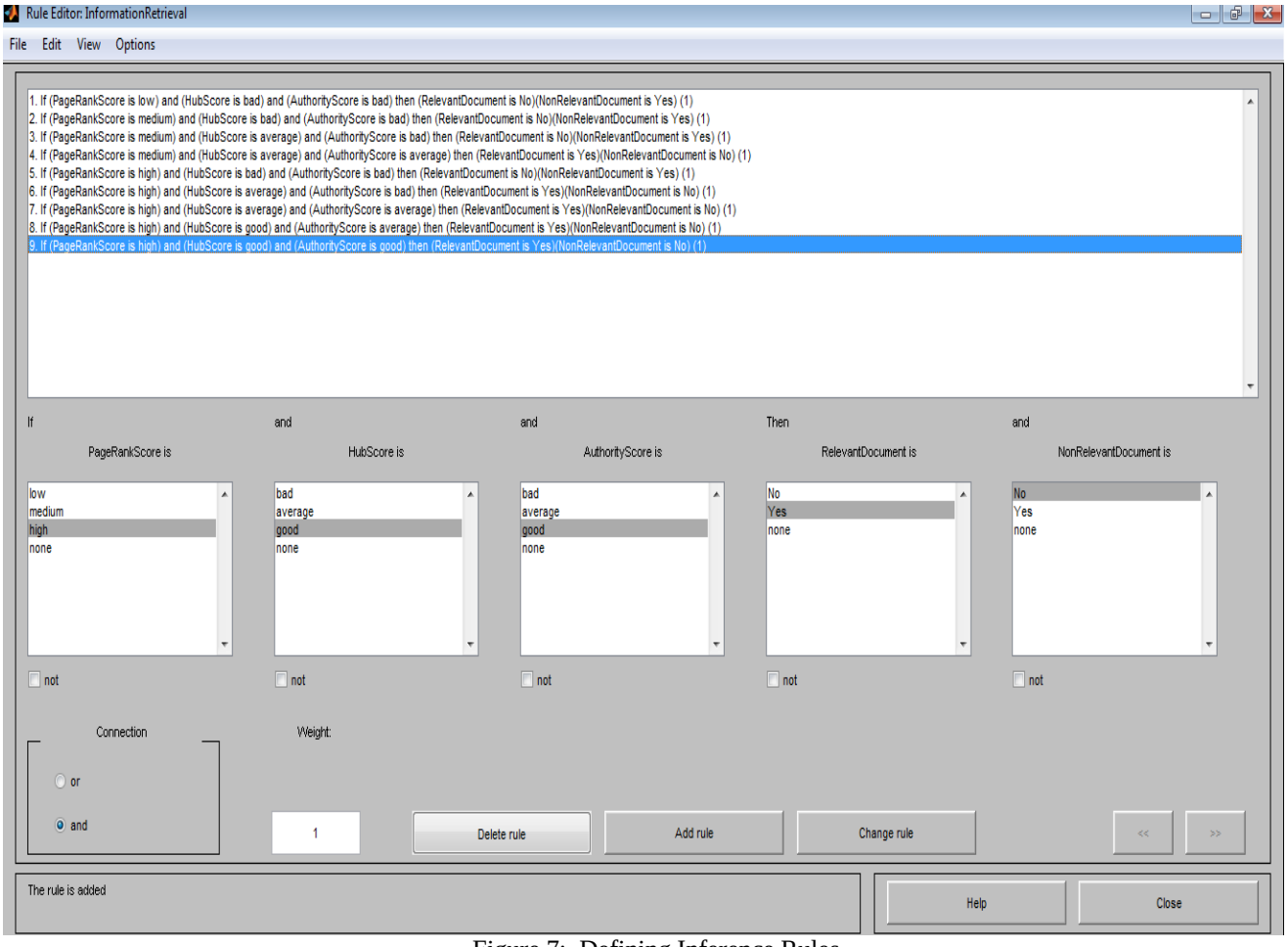

Figure 7: Defining Inference Rules

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

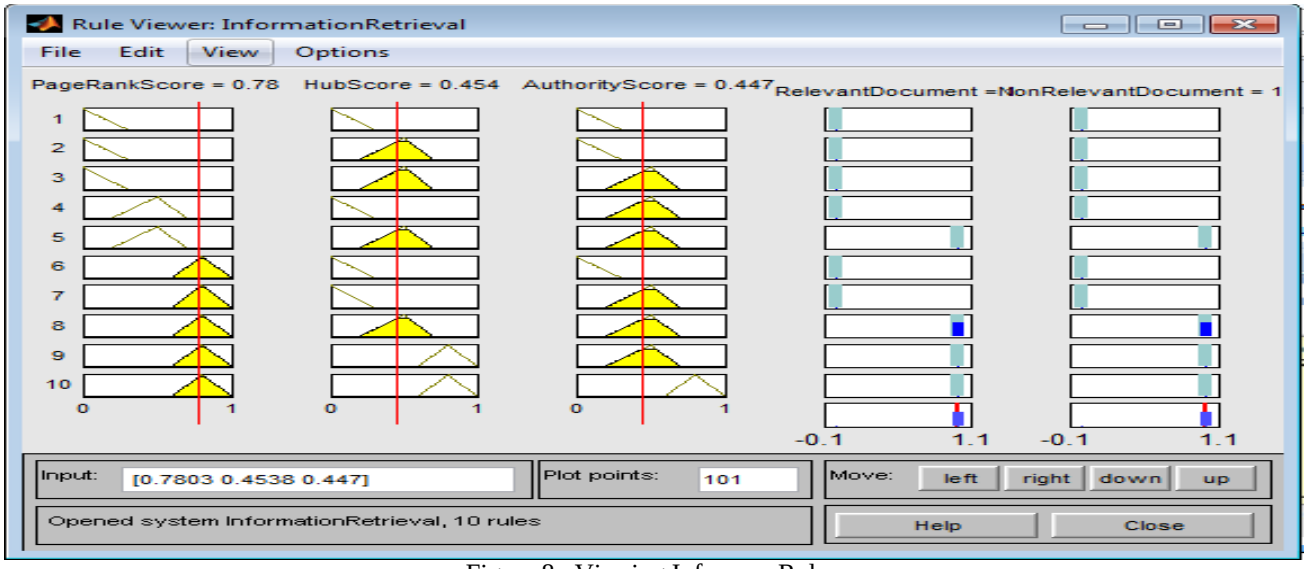

Figure 8: Viewing Inference Rules

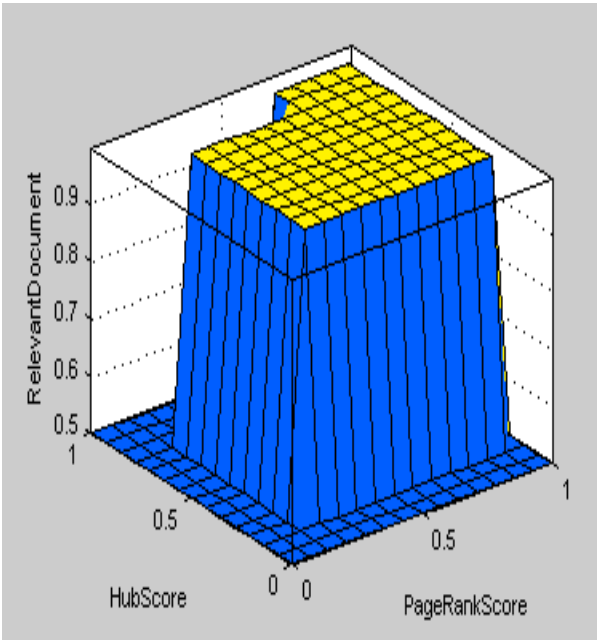

Figure 9 : Surface View plot

IV. CONCLUSION

In this paper, the authors have tried to present the importance of efficient web based information retrieval. Those who are information seekers from the internet, they are facing the problems of acquiring relevant documents from voluminous web contents. The Web Based Information Retrieval system using Fuzzy logic can help us to produce the most relevant pages or documents efficiently and effectively. The performance can evaluated on the basis of relevance factors, precision and recall. However, there is lot of scope for the further improvement of this proposed work.

REFERENCES

[1] Maryam Hourali and Gholam Ali Montazer, "An Intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology", Hindawi Publishing Corporation Advances in Fuzzy Systems Volume 2011, Article ID 683976, 11 pages doi:10.1155/2011/683976

[2] D. Parry, "A fuzzy ontology formedical document retrieval," in Proceedings of The Australasian Worhhop on Data Mining and Web Intelligence, 2004.

[3] J. Zhai, Y. Liang, Y. Yu, and J. Jiang, "Semantic information retrieval based on fuzzy ontology for electronic commerce,"Journal of Software, vol. 3, no. 9, 2008.

[4] J. Zhai, Q. Wang, and M. Lv, "Application of fuzzy ontology framework to information retrieval for SCM," in Proceedings of The International Symposium on Information (ISIP '08), pp.173–177, May 2008.

[5] J. Zhai, Y. Yu, Y. Liang, and J. Jiang, "Traffic information retrieval based on fuzzy ontology and RDF on the Semantic Web," in Proceedings of The 2nd International Symposium on Intelligent Information Technology Application (IITA '08), pp. 779–784, December 2008.

[6] M. A.A. Leite and I. L.M. Ricarte, "Document retrieval using fuzzy related geographic ontologies," in Proceedings of

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.12 e-ISSN: 2251-7545

the International Conference on Information and Knowledge Management, pp. 47–54, 2008
[7] G. Fu, C. B. Jones, and A. I. Abdelmoty, "Ontology-based spatial query expansion in information retrieval," Lecture Notes in Computer Science, vol. 3761, pp. 1466–1482, 2005
[8] C. Bratsas, V. Koutkias, E. Kaimakamis, P. Bamidis, and N.Maglaveras, "Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions," in Proceedings of The 29th IEEE Annual International Conference on Engineering in Medicine and Biology Society, vol. 2007, pp. 3794–3797, 2007
[9] Sunhong Kim and Sangyong Han,"The method of Inferring Trust in Web-based social Network using Fuzzy Logic", Proceedings of the International workshop on Machine Intelligence Research (MIR Day, GHRCE, Nagpur), 2009
[10] M.J. Martín-bautista and M.A. Vila, D.H. Kraft and J. Chen, J. Cruz, "User Profiles and Fuzzy Logic for Web Retrieval",
Journal of Soft Computing, v. 6, n. 5, 2002, pp. 365-372
[11] Padmini Srinivasan and Donald H. Kraft, "Extensions and Generalizations of Combining Fuzzy Sets and Rough Sets: Vocabulary Mining for Information Retrieval**",** Technical Paper, Department of Computer Science, Louisiana State University, Baton Rouge, LA, USA, 2003

.