Modern Applied Science; Vol. 13, No. 11; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Fuzzy Logic System for Retrieval of Information in Electronic Libraries

Ali Mohammad H. Al-Ibrahim[1]

[1]Computer Science Department, Faculty of Information Technology, The World Islamic Sciences and Education University (WISE), P.O. Box 1101, Amman 11947, Jordan

Correspondence: Ali Mohammad H. Al-Ibrahim, Computer Science Department, Faculty of Information Technology, The World Islamic Sciences and Education University (WISE), P.O. Box 1101, Amman 11947, Jordan, E-mail: ali.alibrahim@wise.edu.jo

Received: April 2, 2019 Accepted: October 18, 2019 Online Published: October 22, 2019

doi:10.5539/mas.v13n11p76 URL: https://doi.org/10.5539/mas.v13n11p76

Abstract

This research represents one of the steps aimed to address one of the most important challenges on the Web and digital libraries, which is compute the rank of the document's, and its importance, and their relevance to the user and to meet their needs for information, and so by taking advantage of the vast potential of logic Fuzzy in dealing with this kind of problems, and provide high flexibility for the user to clarify the issues and areas that interested them.

This research is concernd on the design and implementation of a proposal for the information retrieval system, called Fuzzy Information Retrieval System(FZIRS). This system is designed to deal with a huge distributed database on a group of computers (servers) associated with the Intranet network specially designed to work the system, which includes different types and sizes of text files.

The proposed system has the ability for mining of data mining from the database and retrieve useful information from them and that meet the user's needs well.

This accomlished through the applying of the proposed algorithms for indexing operations and calculate the rank of documents and generate keywords operations and display the retrival results, which showed high quality when calculating results compared with other Information retrieving algorithms.

Keywords: information retrieval system, fuzzy logic, data mining, distributed database information retrieving algorithms, Fuzzy Information Retrieval System(FZIRS)

1. Introduction

Computer plays an important role in the design and construction of modern information systems, to achieve speed, accuracy and confidence and authority to information system, which lead to it high efficiency in performance. It also conducted a complex calculations, which are difficult to implement manually in addition to superior ability to store a huge amount of information in an orderly manner so as to facilitate retrieval very slim times as electronic computer can accomplish all the other tasks carried out by the information system, including the safe investigation and data integrity and full security against loss or damage by beneficiaries.

The large development in the number of systems intended for automated libraries of all kinds shows the importance of this technology for libraries and have many of the surveys conducted to the number of automated systems in libraries, especially in the United States. Where the survey was conducted in 1964, which proved that there are 25 system automatically works in libraries in America. In 1971, a survey was conducted last known (Lark Survey) was the census (1366), an automated system for libraries (between integrated systems or parts of systems in the (506) Library.) (Jassim Mohammed & Sabah Mohammed, 2002)

In 1984, another survey was conducted, proved that there are about of (30000) thirty thousand special automated system libraries (that most of these systems operate on a personal computer PC), and noted the extent of the figure in 13 years is the time difference between the Lark Survey 1971 and 1984, where the number of automated systems for libraries multiplied by 59 times, and the fact that it was due to two main reasons:

1. The actual needs of modern societies of the need to control the huge and growing amount in the desired traded by these communities and gets the necessary information from them quickly and the accuracy of the information and efficiency. And benefit from the economic, social and cultural development objectives.

2. The great potential offered by computers mechanism and modern communications technology of the large storage capacity and processing speed that has become measured in simple parts of the second and multiplied hundreds of times the speed of data exchange, which decrease dramatically the cost of these operations, in addition to the possibility of remote processing and the use of computers and data banks and others.

The history of the use of automated systems in libraries dates back to 1935, the year saw the introduction of the first data processing machine in the library, where Texas University of using Punched Card devices in Circulation System, followed by the Boston Public Library to use punched cards to analyze some statistics supply and rolled automated systems then used as digital libraries digital Libraries, and the first person who's ask for using this type of device in libraries both Melvin J. Voight in charge of the University of California library, along with Clay L. Perry of the computer center at the same university in 1962, where he was the pilot venture is to convert the number of records (700) series to machine-readable form in addition to print a monthly list of numbers plus a complete list of everything library has.

Its notes that most of the automated systems that appear were parts of the systems, that is, they were not integrated systems that can include all library operations at the same time but it has been dealing with only one part of the library operations such as indexing or Circulation, but with 1961 the national Library of Medicine work with and Medical Literatur Analysis and Medlars Retrival system and review system functions in an attempt to automation of all functions of the library addition to conducting searches bibliographic and issuing Meoicus index as well as the operations of the indexing mechanism and borrowing mechanism and assist in the acquisition and adjust patrols and thus the emergence of the First

system an integrated in libraries in 1966. Not only that, it changes and developments of interest in the field of library and information that evolution that has taken place on the services provided by libraries, where the space agency Nasa tested the first system to broadcast selective information SDI works on the automated computer, where the user select topics that wish to consult by The automated system compared between subjects of articles and descriptors objectivity and the beneficiary by providing it with a list of articles that match his interests that he previously selected (Jassim Mohammed & Sabah Mohammed, 2002).

So the public libraries among the first institutions that was adopted the information retrieval systems. Retrieval of information used in public library systems developed initially by academic institutions and then by commercial vendors, in the recent years major developments has been the completion in the field of information technology, and especially after the Web appearance of continuous improvement in the performance of search engines making it the most popular e-libraries in all disciplines. And ensure that the First generation of libraries systems strategies and primitive techniques made the search depends on the author's name and address.

But the second generation, the search strategy has been included: 1. Subject Title. 2. Keywords. 3. Many complex queries. While third generation focus on: 1. graphical interfaces(GI). 2 - Electronic forms. 3. Hypertext properties. 4 - Open System Architecture: It is a system that has the ability to continue change for its operations and internal structure (Yates, R. A. & B. Ribeiro-Neto., 1999).

2. Aim of the Research

Information retrieval process has great importance in providing services and transfer information and the contents of the scientific web pages, as well as the contents of the documents in digital libraries to users by means of quick and easy research and methodological tools that achieve retrieval of information for use in the form that meets their actual needs and conviction fully in their field of study or their research.

Thus, here we are trying to answer the following questions which form the main focus of the search:

1. What is a work mechanism in information retrieval systems technology? What are the main constituent parts?

2. How can we get good results meet the needs of users?

3. What are the reasons for the escalation of retrieval results related compared with other related results?

4. How to improve and develop strategies and retrieval operations in order to give more accurate results to users?

5. What are the algorithms used to calculate the documents ranking and which give high quality results?

3. Information Retrieval Systems

Information retrieval science is one of branches of old knowledge, comparable to the feet of computer science itself, if not the oldest. This definition is in the information retrieval of the oldest definitions field, if not the oldest among them, is what was mentioned Meurs (1950), who define information retrieval "is the name of the process or the way in which enables the user to convert his need for information to the actual list of documents in the store (such as a hard drive) which contains useful information for the user (Fabio Crestani & Gabriella Pasi., 1997).

Information retrieval system in general includes two parts:

1. Archive Script: which is a set of text units (which is often named as the "document.").

2. Retrieval engine: The system user submit applications that describe the type of documents required, and the retrieval engine matching the applications with the documents in the archive script, and then returns the user to the college part of a group of documents that give the best match with the user needs (Bhaskar Karn., 1998).

And the main requirements of these systems are:

1. The user must be able to enter any word from the natural language, any words, or phrase to the system without the need to engage in indexing and search operations, and this has the implication that the retrieval of the full text and indexed every word in the document automated .

2. System must arrange (Rank) documents retrieved by the possibility of their usefulness to the user request.

3. The system must support the re-formulation of a mechanism to search sentences based on previous information to the user (feedback) (D. Hiemstra., 2001).

Based on the foregoing, information retrieval can be considered as one of computer science fields system, which aims to store a large amount of information and allow quick access to it, and this information can be of any type: text, audio, or video (Van Rijsbergen, C.J., 1979).

4. Soft Computing and Information Retrieval

The term Soft Computing(SC) defined by (L.A. Zadeh., 1994) Zadeh, which refers to the cooperation between useful mechanisms to solve problems that require some sort of intelligence away from traditional computing methods. SC offers a total of appropriate technologies to solve the ambiguity, lack of objectivity, and inaccuracies in many of the existing problems.

The retrieval of information is aimed to modeling, design, and implementation of systems capable of providing fast and efficient services to access based on large amounts of information content. It also aims to assess the relationship of the elements of information the user needs actress words inquiry. This task is very difficult and complicated because of the lack of punctuated in objectivity, uncertainty, and lack of precision.

SC includes different methods group to solve such problems, Fuzzy logic , genetic algorithms, neural networks, rough sets, and Bayesian (Ricardo Baeza-Yates., 2003)

As with the first beginnings of the fuzzy rough sets theory by L. A. Zadeh in 1965, it has been the completion of clear progress in the utilization of concepts Fuzzy rough sets in multiple fields of visual and sensory science. Fuzzy rough sets were used and extensively in the science of computer vision and artificial intelligence. Fuzzy rough sets dealing with ambiguity and thumb buried in the comprehension of the human system, and provide excellent system structure to describe, analyze, and interpret the mysterious events and doubtful. Where the human visual system is basically a fuzzy system, because we understand and interpret the visible world mysterious around us (Acharya, T. & Ray, A. K., 2005), information retrieval takes advantage of fuzzy logic in information blocking operations, extracting texts, inquiring language models (Ricardo Baeza-Yates., 2003).

5. Digital Books

The digital book been scientific revolution that could change the way of history in the twenty one century, more than other invention of the printing machine , as digital book (E.Book) became in hand, is easy to deploy, easy distribution and get it, easy to read and see, and Digital Book in fact only scientific material, from letters and pictures, reports and statistics, are preserved in some kind of many kinds of files.

The digital book differs in its features and characteristics of the paper book in many points, the most important of these points are（Ahmad Ziad, 2010):

1. Get the book without price or a very low price.

2. Ease of reading the book on the computer, it can, in many cases, change the size of the letters, and control the degree of light , to relieves the reader.

3. Access to the required idea in the book, using the tool "Search," and mediated by a single word, to find out the positions of the book required

4. saving millions of books in limited space in laser disks or hard drives (Hard) embedded within the computer or by independent portable drives, allowing download millions of books on travel, trips and moving them from place to place with ease.

5. Reading, writing and documentation speed, and achieving scientific knowledge revolution

6. The Proposed System for Retrieving Information Using Logic Fuzzy

The system (Fuzzy Information Retrieval System (FZIRS)) is designed to deal with large distributed data base on a group of computers (servers) associated with a network Intranet, which includes different types and sizes of text files. This system has the ability to dig for the data in the data and retrieve base useful ones that meet user needs efficiently with information.

7. Create a Text Document

The index building process is one of the key processes for efficient retrieval of documents and this process requires two major steps:

1. Document analysis, includes determine the important words, and her candidacy, and this requires analysis of each document in the database (e.g. web documents) and the organization of these documents in the form of key elements of the document (title, author's name, source) and how to represent the information in it, on tabular form, critical information in text, charts, graphics, or images. The decision that must be taken is to identify any of the information or parts of the document will be indexed and which are not included in the indexing process.

2. analyze the strings or terms, and the decision must be taken is to identify any words (or phrases) can be used as a reference to the document for the best representation of the moral contents of the documents.

So before we begin the process of cataloging the collection of the preparations and settings that are on the document in order to prepare it for indexing process, which vary depending on the quality of the text document to be indexed ,as in Figure (1): the creation of the document.

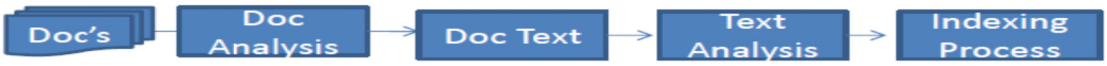


Figure 1. the creation of the document.

In order to analyze text documents, it has been proposed of special data structure data of the proposed algorithm, and hold onto this structural important information available in the document, which will facilitate the indexing process and the calculation of the rank of the document, and this structure included two parts:

1. Special part in words, which includes information on every word found in the documents (Term Frequency).
2. Documents in which the word found(document frequency).

Figure 2 the proposed structure of the data, which will be addressed later.

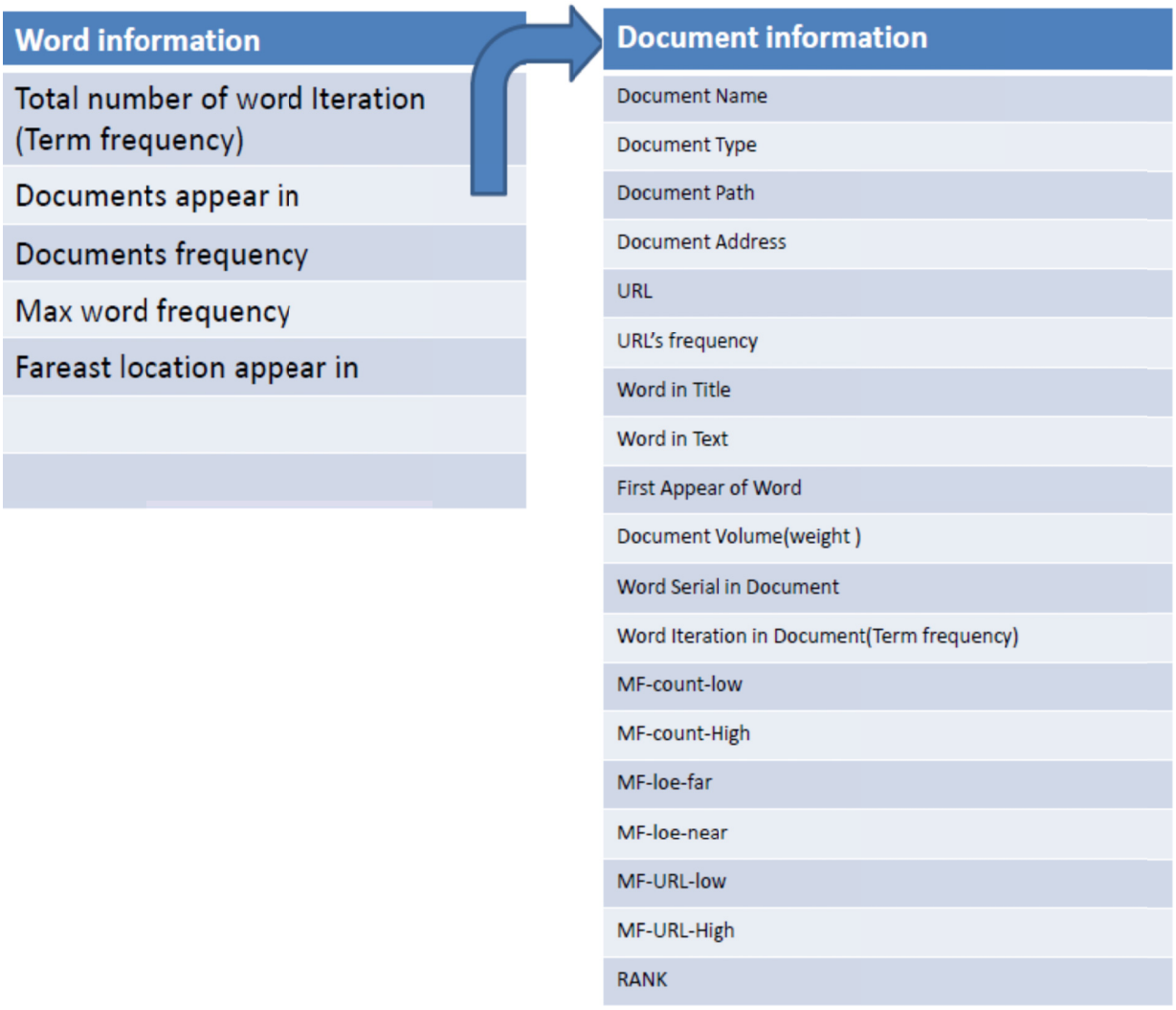


Figure 2. proposed structural data, which will be addressed later

8. System Architecture

System architecture consists of the structure shown in Figure (3), where we note in this architecture, there are five computers represent the servers that contain the database, as each server contains its own database, which will be indexed later. Also note that one of these servers is the main server, which will be housed in the end, all indexing files of the servers. There are an unknown number of user computers (Clients) connected to the network with servers through the HUB, because the work is done on the intranet (ECN), as we need to Access Point device connected to the network for the purpose of wireless communications for other Clients.

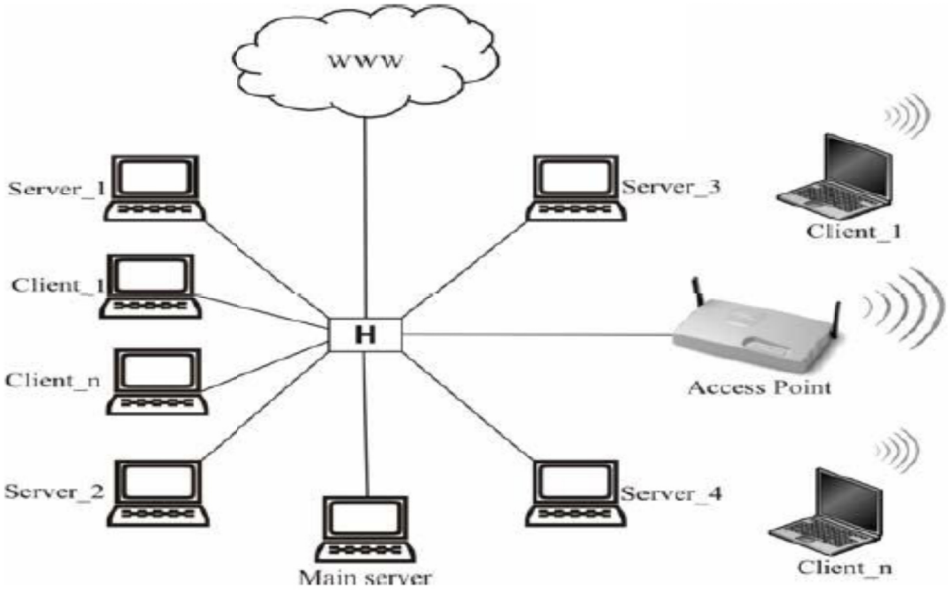


Figure 3. FZIRS architecture

With respect to the client's computers, they do not contain the database, but include interfaces that enable users to communicate with the main server, which will be used for information retrieval that meets their needs.

9. Algorithms are Proposed

As the designer of FZIRS system in this research consists of two programs, one of whom works on computers servers and the other on computers clients, it was necessary to work specific software algorithms depending on the function performed by each of them, and following the basics ideas of the work of both of the two programs and algorithms.

9.1 The Proposed Algorithm for Indexing Process

1. Initially determine the database path, by define the number of files in it.

2. determine the file type as if the file type of HTML you'll have to remove the decorations tags located within the textual content of the file, this process is carried out using a regular expression. Where in the case of other types of files you do not need to do this operation.

3. Extract text from files, with the use of a regular expression that deleting special characters of text.

4. Then deducted words from the text as special regular expression of English characters, also examine each word using another regular expression excluded other special words that will not ne within the indexing process.

5. After obtaining the correct word, store their information in private matrix by words information which its structure declares in Figure (2).

6. Before information is stored make sure that this information was not stored in advance, so if you had previously stored their information is added to the existing entry.

7. Compute the term (word) frequency in the file using the algorithm Boyer- Moore.

8. This process is repeated until all files are indexed and the content of the terms (words).

9.2 The Proposed Algorithm to Process the Expense of Rank

Computation process of rank, based on the Fuzzy logic operations, and as follows:

- Fuzzification

Fuzzification: it is in the process compute of membership functions(MF) for each word present in the matrix of the words, and illustrated in Figure (4) functions used in this process, and by the following formulas:

$$MF\_count\_low = 1 - (1/(1 + \exp(-((count - (\max\_count / 2)))))) \quad \ldots\ldots 1$$

$$MF\_count\_high = 1/(1 + \exp(-((count - (\max\_count / 2))))) \quad \ldots\ldots 2$$

$$MF\_loc\_near = 1 - (loc / end\_loc) \quad \ldots\ldots 3$$

$$MF\_loc\_far = loc / end\_loc \quad \ldots\ldots 4$$

$$MF\_URL\_low = 1 - (1/(1 + \exp(-((fileURLcount - (\max\_no\_of\_URLs / 2))))))) \quad \ldots 5$$

$$MF\_URL\_high = 1/(1 + \exp(-((fileURLcount - (\max\_no\_of\_URLs / 2)))))) \quad \ldots\ldots 6$$

MF_count_low: Membership Function to the small number (count) of iteration of the Term(word) in the text file.

MF count high: Membership to the large number (count) of iteration of the term (word) in the text file.

MF_loc_near: Membership Function of the term (word) location nearest to the beginning of the text file.

MF_loc_far: Membership Function of the term (word) location Fareast to the beginning of the text file.

MF_URL_low: Membership Function to the small number of iteration of the connections (URLs) in the text file.

MF_URL_high: Membership Function to the large number of iteration of the connections (URLs) in the text file.

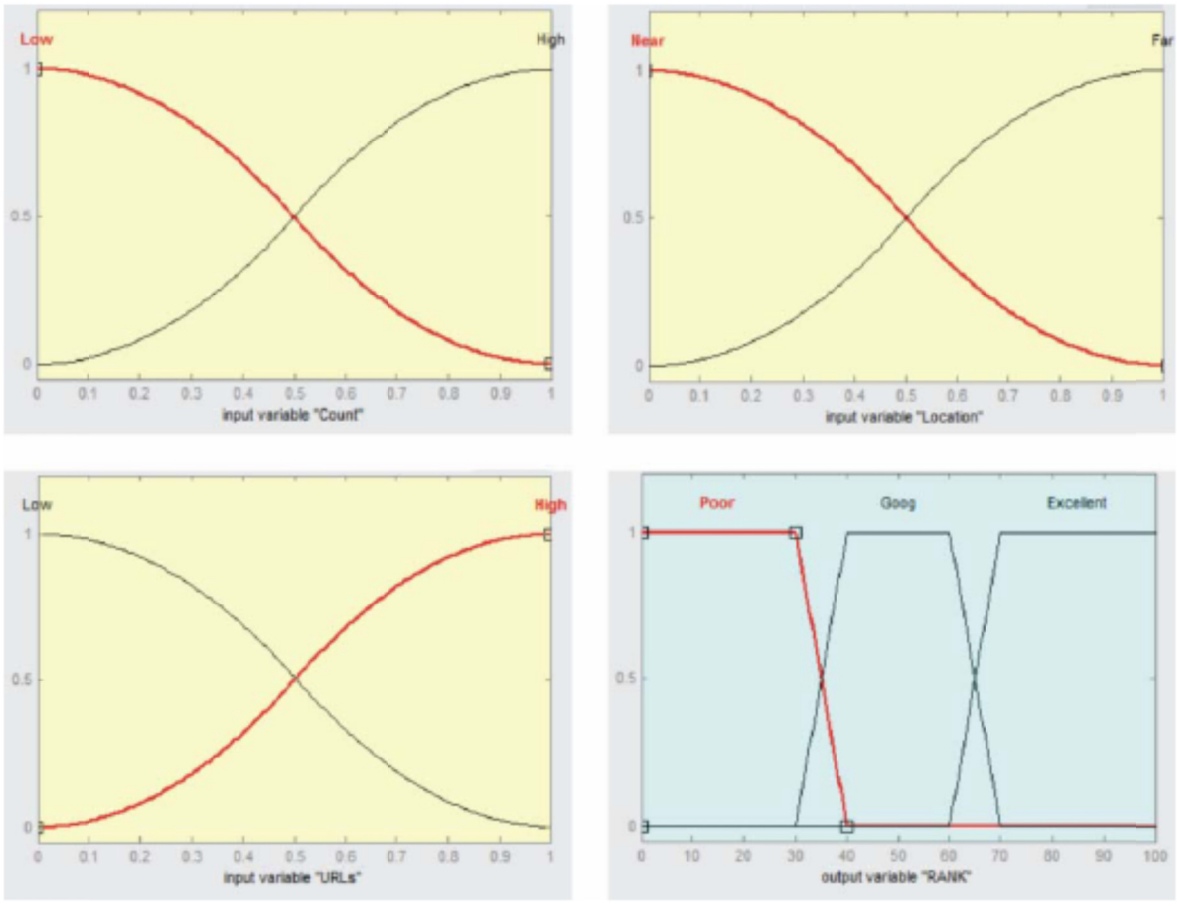


Figure 4. Membership functions (MF)

❖ Knowledge bases:

After the completion of the Fuzzification the knowledge base process is ready which consisting of Fuzzy aggregates (poor, good, excellent), which represents the result of the implementation of the Knowledge bases of

Fuzzy logic, using eight bases with files from HTML type, where in the other cases there are four Knowledge bases used . These rules are illustrated in Figure (5).

1. If (Count is Low) and (Location is Far) then(RANK is Good)
2. If (Count is Low) or (Location is Far) then(RANK is Poor)
3. If (Count is High) and (Location is Near) then(RANK is Excellent)
4. If (Count is Low) and (Location is Near) then(RANK is Good)

1. If (Count is Low) and (Location is Near) and(URLs is Low) then(RANK is Poor)
2. If (Count is Low) or (Location is Near) or (URLs is High) then(RANK is Good)
3. If (Count is Low) or (Location is Far) or (URLs is Low) then (RANK is Poor)
4. If (Count is Low) or (Location is Far) or (URLs is High) then (RANK is Poor)
5. If (Count is High) or (Location is Far) or (URLs is High) then (RANK is Good)
6. If (Count is High) or (Location is Far) or (URLs is Low) then (RANK is Poor)
7. If (Count is High) or (Location is Near) or (URLs is High) then (RANK is Excellent)
8. If (Count is High) or (Location is Near) or (URLs is Low) then (RANK is Good)

Figure 5. Knowledge bases Logic Fuzzy

❖ Aggregation:

After the implementation of previous Knowledge bases (rules), Aggregation process of the results is done using Aggregation OR-MAX, as shown in Figure (6)

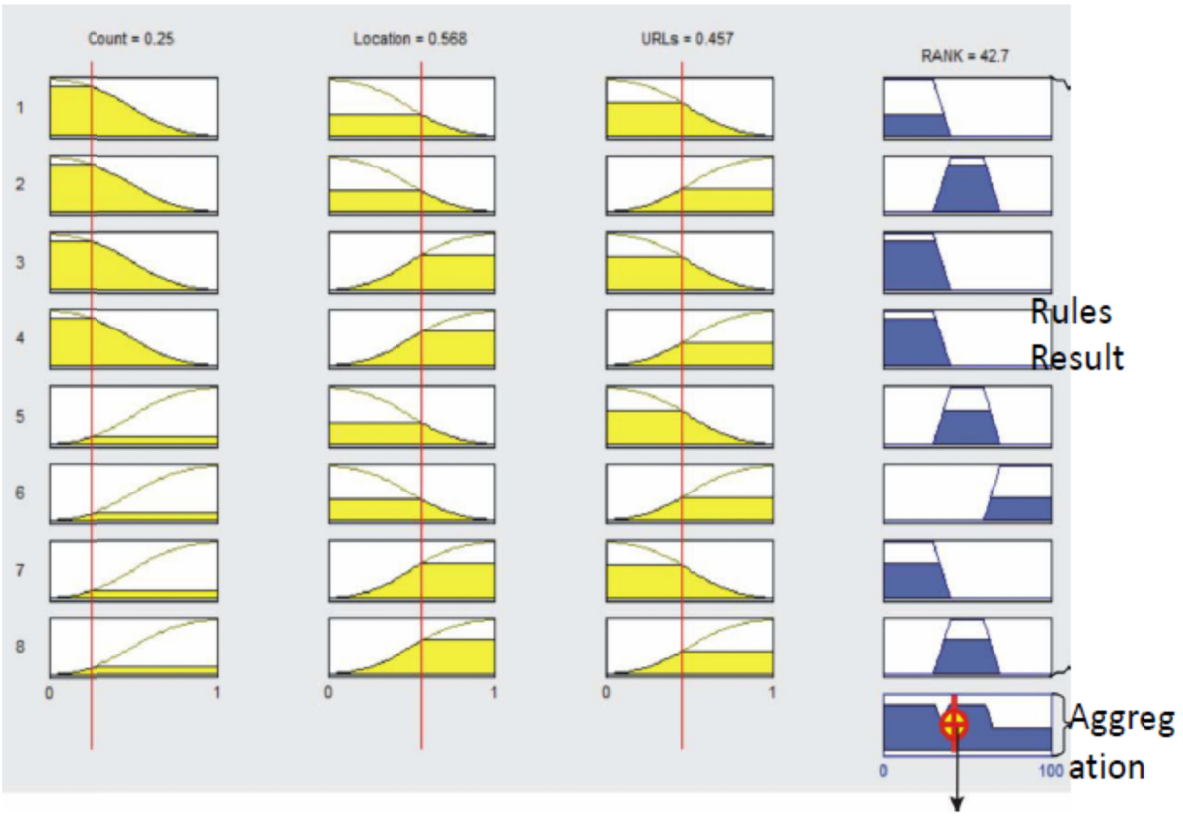


Figure 6. implementation of the rules and the process of collection and calculation of the value of grade

❖ Defuzzification

Compute the Rank for each file of the word depending on the Centroid way as in Figure (6), according to the following formula

$$RANK = \frac{\sum Xi\mu(X)}{\sum \mu(X)} \quad \text{.....} \quad ٧$$

If the word found in the title and in the content of the file, the rank of it will become a double.

9.3 The Proposed Algorithm for the Search and Retrieval

When the user enters the query statement and selects a Search button, the screening process begins for the query statement. This phrase is entered through a SmartTextBox, which is a class helps the user in correct process for printing errors that may be present in the query statement, and figure (7) shows the flowchart of the search process.

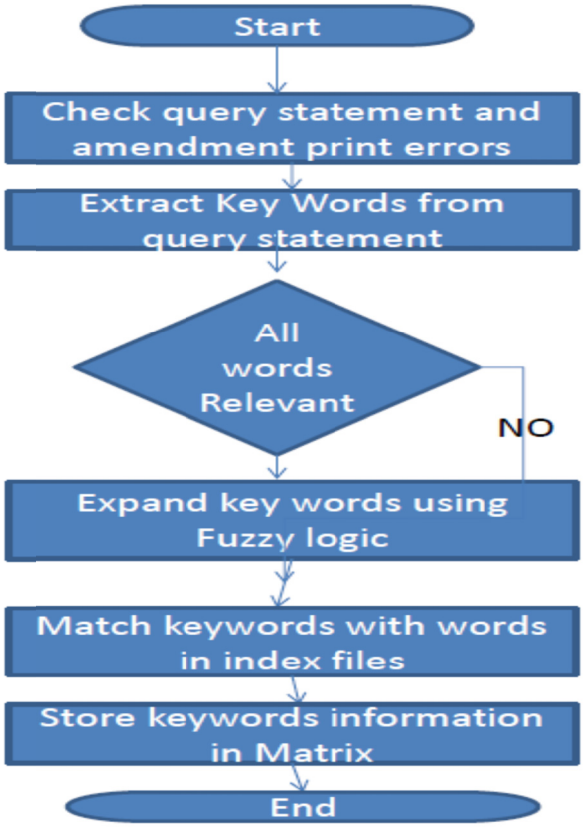


Figure 7. Flowchart for the search process

After that call a custom function to work to create a matrix keywords information, this information shall include the keyword, the number of files that include this word, the number of servers, which includes information on this floor, and information files in which the word appears. Others also create a matrix to save the recovered files information, as shown in Figure (8).

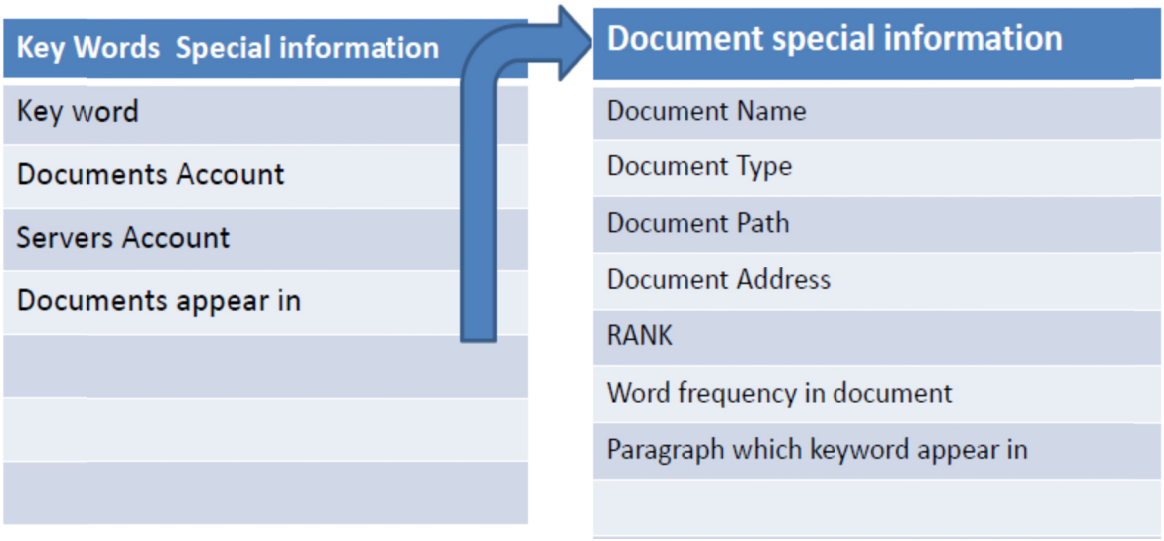


Figure 8. keywords structural matrix.

Keywords are extracted from the query statement using the phrase more than a regular expression to delete tags and special symbols, and then get on the ferry depending on regular expression with special English alphabet. Then the statement is breakdown into keywords and stored in its own matrix.

If the Checkbox of All relevant word index will be expanded keywords contained within query statement, using the method of private logic Fuzzy search, which represent a simplified search algorithm based on its principle on the Levenshtein-Distance and used to get closer to the match between the strings. And give Levenshtein-Distance the matching between two literal.

The search process continue until to be completed all keywords examination, also the process of comparing words added to the new list in order to make sure it is not already appear in the list.

The values are set to own files matrix related to key words and complete fill keywords matrix.

10. FZIRS Implementation and Discuss the Results System

The FZIRS system has great importance, because of its speed and accuracy in indexing, search and retrieval operations, and easily used by the users. It has been implemented and programming the system using the programming language C# within the package of programming languages Visual Studio 2008 with frame work of NET Framework Ver. 3.5. Language C# supports writing programs with multithreading, which is used to design and implement the designed system required, the system depends on the client \ server model.

For the application of the proposed system must achieve the following:

1. Provide computer with special specifications because we need computer where the main memory size of 8GB RAM or more and 64bit operating system to be in order to be able to deal with this memory, 2.9GHz processor, hard disk and at least 320GB. This computer is called the main server.

2. Prepare database consisting of text documents on your hard drive for the main server.

3 upload NET Framework Ver. 3.5 on the main server.

4. Upload the main server software of the system of FZIRS process of indexing on the main server.

5. Linking the main server with a wireless access point for the purpose of creating Intranet network.

6. After that main server program executed for the purpose of obtaining indexing files that we keep on your hard drive with a main server.

Then upload the user program (the second part of the proposed system) on users' computers associated with the Intranet network in order to provide search and retrieval service.

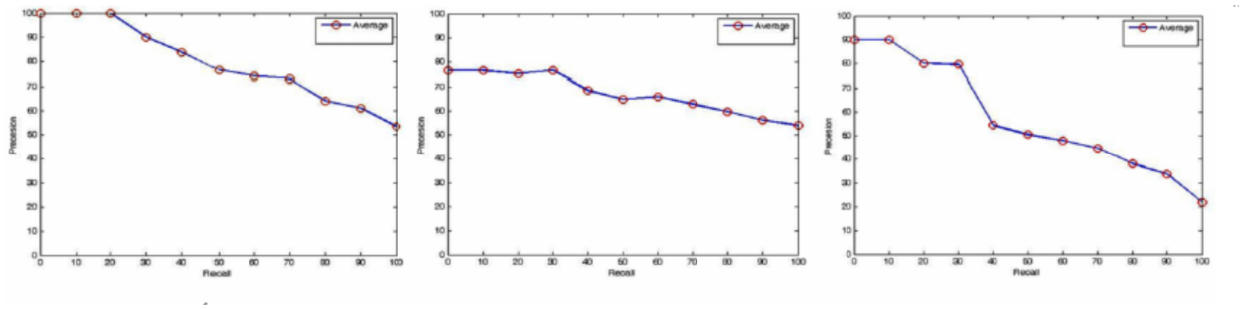
The Fuzzy logic system Tested on a set of ten query statements, and used the same query statements in the system does not support the Fuzzy logic , and another system supports research in the desktop interface, which is a program that indexing and searching based on lucene written language C#, and compute the average of recall and precision for all query statements, and the results are as in table 1 and charts in Figure (9). Then the results

of the three algorithms compiled in a chart one shown in Figure (10) in order to demonstrate the quality and efficiency of the proposed algorithm using fuzzy logic.

And then representation the results of the three algorithms by the histogram in Figure (11), and the proposed algorithm and program Lucene globally in the histogram shown in Figure (12).

Table 1. Results of the ten query statements

Recall	*Precision without using Fuzzy logic*	*Precision using Lucene program*	*Precision using Fuzzy logic*
10%	76.25%	90%	100%
20%	75.13%	79.98%	100%
30%	76.40%	79.50%	90%
40%	68.40%	54.37%	83.70%
50%	64.80%	50.38%	76.40%
60%	65.90%	47.82%	73.90%
70%	62.70%	44.57%	72.90%
80%	59.60%	37.96%	64.03%
90%	56.20%	33.73%	61.10%
100%	53.80%	21.82%	53.40%



a b c

Figure 9. The recall and precision for ten i query statements

a- using Fuzzy logic, (b) without using the Fuzzy logic C- using Lucene program.

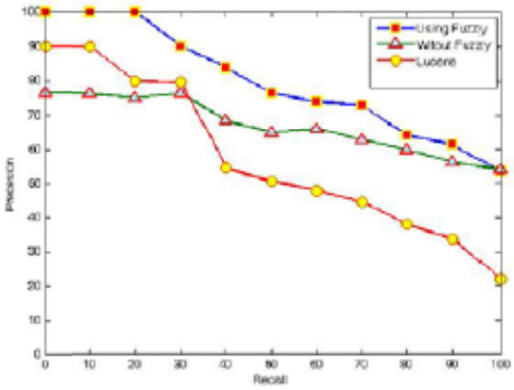


Figure 10. chart of the results of the three algorithms which demonstrates the quality and efficiency of the proposed algorithm in information retrieval

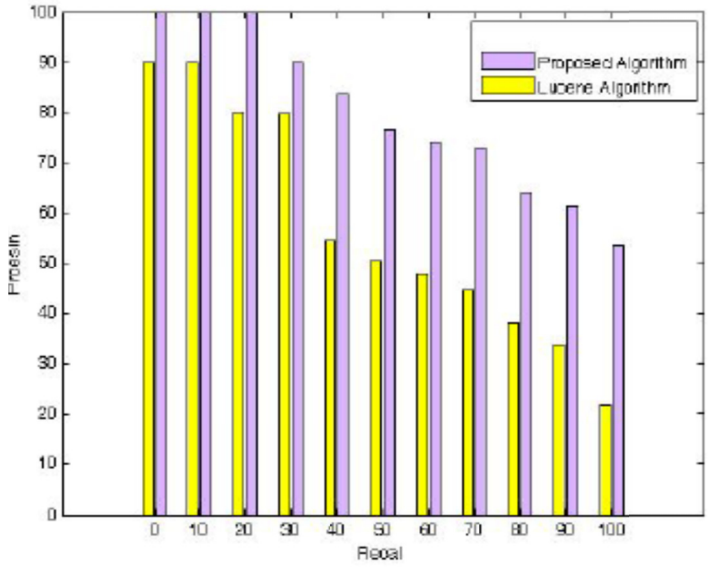


Figure 11. chart of the results of proposed algorithm and lucene algorithm

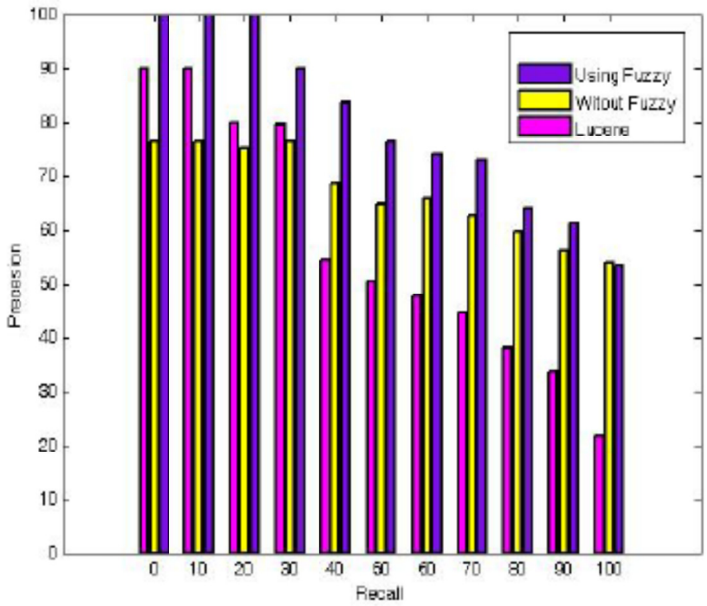


Figure 12. chart of the results of the three proposed algorithms(using Fuzzy, without Fuzzy, Lucene).

11. Conclusions

In this research we have reached to the following conclusions:

1. The use of the theory of Fuzzy logic in information retrieval has at least one of the following characteristics compared with the classic ways:

a. Fuzzy Relevant relationships be more reflective of real values, also be built more realistic process.

b. In the event that the number of documents is very large, the Fuzzy logic directs the user to the more important required documents.

c. Fuzzy query provides the user the opportunity to get to the big topics that concerned.

2. The proposed algorithm is applied for the purpose of calculating the rank document gave accuracy and high flexibility.

3. using Regular expressions provided easy process in document analysis process and in determining the symbols and words excluded.

4. Use the Boyer-Moore algorithm has helped to repeating compute quickly and accurately.

5. The proposed indexing algorithm which requires more time to implement to make the process of search and retrieval are faster.

6. When a user enters a query of the system, the system checks the query statement during the entry process and thus ensures access to accurate results without the user's operations correction and amendment occupancy.

7. The use of the proposed algorithm led to generate the key words to increase the accuracy of the results.

8. The process of expanding the keywords using Fuzzy logic and Levenshtein-Distance algorithm to improve the quality of recovery by expanding the number of relevant documents.

References

Acharya, T. & Ray, A. K. (2005). Image Processing: Principles and Application, John Wiley & Sons, Inc, Hoboken, New Jersey.

Ahmad Ziad (2010). the Arabic language and Digital Book, a professor of modern Arabic literature at the University of halab, a member of the World Council of the Arabic language.

Bhaskar Karn. (1998). Information Retrieval System Using Fuzzy Set Theory – The Basic Concept , Assistant Professor, Department Of MIS(Management Information Systems), Birla Institute Of Technology, Mesra, Ranchi.

D. Hiemstra. (2001). Using Language Models for Information Retrieval, Universiteit Twente. ISBN 90-75296-05-3, ISSN 1381-3617; No. 01-32 (CTIT Ph.D. Thesis Series) Subject headings: information retrieval, natural language processing.

Fabio Crestani & Gabriella Pasi. (1997). Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks , Department of Computing Science, University of Glasgow.

Jassim Mohammed & Sabah Mohammed (2002). *Introduction to Library and Information Science,* University of Sana'a, 146-143.

L.A. Zadeh. (1994). Fuzzy logic, neural networks and soft computing , Communication of ACM. https://doi.org/10.1145/175247.175255

Ricardo Baeza-Yates. (2003). Information retrieval in the Web: beyond current search engines, Center for Web Research, Department of Computer Science, University of Chile, Blanco Encalada, 2120 Santiago, Chile , Received 1 January 2003; accepted 1 July 2003.

Van Rijsbergen, C.J. (1979). Information Retrieval Butterworths , Department of Computing Science, University of Glasgow, second edition.

Yates, R. A. & B. Ribeiro-Neto. (1999). Modern Information Retrieval , Addison-Wesley

Copyrights