Multi-agent System for Documents Retrieval and Evaluation Using Fuzzy Inference Systems

Galina Ivanova, Ark Andreev, Marwa A. Shouman*
Department of Computer Systems and Networks, Bauman Moscow State Technical University

Article Info

Article history:

Received Aug 16, 2016
Revised Oct 20, 2016
Accepted Nov 23, 2016

Keyword:

Fuzzy inference system
Information Retrieval
Multi-agent system
Vector Space model

ABSTRACT

Recently the World Wide Web are packed with huge quantities of information. From this view the user finds it difficult to get the relevant informations due to the increased of their quantities. This paper uses multi-agent system uses intelligent agent in order to retrieval documents from the World Wide Web. The user by this system can easily get the relevant documents which to need them.Multi-agent System is combined with fuzzy inference system for ranking documents. The documents ranking score by cosine similarity using fuzzy inference system development and implemented much simpler than the traditional method which require mathematical equations.

Corresponding Author:

Marwa A. Shouman,
Department of computer systems and networks,
Bauman Moscow State Technical University,
Email: marwashouman834@yahoo.com

1. INTRODUCTION

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources [1]. Automated information retrieval systems are used to reduce what has been called "information overload". Web search engines are the most visible IR applications. An information formal statements of information needs, for example search strings in web search engines.

Multi Agent Systems (MAS) is considered a pool of information agents. An information agent is an agent that has access to one or more information sources, and is able to store and process information obtained from these sources in order to answer queries posed by users and other information agents. The information sources may be of many types, including web services, web sites, RSS-feeds, and traditional databases [2].

Fuzzy Metagraph is an emerging technique used in the design of many information processing systems like transaction processing systems, decision support systems, and workflow System [3].

Zheng-Hua Tan has proposed a Fuzzy Metagraph (FM) based knowledge. The FM has been applied to fuzzy rule-based systems for knowledge representation and reasoning in the format of algebraic representation and FM closure matrix [4]. A.Thirunavukarasu and Dr.SUmamaheswari have proposed a Fuzzy Metagraph based Knowledge representation of Decision Support System (DSS). This method can be used in many real world applications like E-commerce, share market and disease analysis [5].

To deal with the vagueness typical of human knowledge, the fuzzy set theory [6] can be used to manipulate the knowledge in the basis. Knowledge basis in information retrieval cover a wide range of topics of which query expansion is one the main aim of query expansion is to add new meaningful terms to the initial query.

In this work we focus in the first stage on analyzing the Automatic Information Retrieval Multi agent Modeling based on Fuzzy Metagraph to make user understanding the system model. In the second stage the document relevant result from the model evaluated (ranking score) using cosine similarity in vector space model between quires and documents.

The rest of the paper is organized as follows. Section 2 explains modeling of Multi agent. Section 3 illustrates based technique and points out fuzzy metagraph information retrieval multi agent modeling technique. Section 4 explains the preprocessd model. Section 5 illustrates experimental result and section 6 concludes the paper.

2. MULTI AGENT SYSTEM MODELING

The purpose of the multi-agent system is to aid users in searching and retrieving information available on the World Wide Web. A system devoted to perform automatic information retrieval might encompass four main steps: (i) Search the World Wide Web with keyword, (ii) Extract the required information from web sources (iii) Mining the texts that extracted from the web, (iv) Store the output in database [7]. This model consists of three agents'. The first agent searches in the Internet by keywords (query words) using search engine Google and returns links by collecting the URLs of the available websites from the Internet and stores these URLs into the database. The second agent automatically retrieves document from URLs. The third agent implements the following 1) extract useful information from document retrieval 2) reprocess text using tokenization (to remove all punctuations, special characters and by replacing tabs and other non-text characters by single space), remove stop word (to remove words that are not related to the documents.) and stemming (is a heuristic process in which the end of the words or the affixes of the derivational words are chopped off to receive the base form of the word) [1,7] 3) computes term weight (tf-idf) as describe in the following section 2.1.

Agents are JADE agents capable of (i) interacting exchanging FIPA-ACL messages, (ii) sharing a common ontology in accordance with the actual application, and (iii) exhibiting a specific behavior according to their role [7].

2.1. Vector Space Model

Vector space model (VSM) is based on interpretation of both, documents andqueries, as points in a multidimensional document space [1, 7]. Cosine measure (in equation (1)) that can be interpreted as an angle between the query vector and document vector in m-dimensional document space. Similarity of a document vector to query vector equal the cosine of the angle between them [1, 8] and is given by equation (1).

$$\cos(\overrightarrow{q},\overrightarrow{d_\iota}) = SIM(\overrightarrow{q},\overrightarrow{d_\iota}) = \frac{\vec{q}}{|\vec{q}|}.\frac{\overrightarrow{d_\iota}}{|\overrightarrow{d_\iota}|}$$

$$= \frac{\sum_j \boldsymbol{w_{ij}} \times \boldsymbol{w_{q,j}}}{\sqrt{\sum_j \boldsymbol{w_{i.j}^2}}\sqrt{\sum_j \boldsymbol{w_{q,j}^2}}} \tag{1}$$

$q$is the query vector, di is the ith document vector in the collection, w_{qj} is *tf-idf* weight of term j in the query q, wij is *tf-idf*weight (term frequency – inverse term frequency) [1]of term j in the document di.

Where the two vectors d (document vector) and q (query vector) given by the flowing equation

$$\vec{d} = (w_{i,1} w_{i.2} \ldots\ldots w_{i,j})$$
$$\vec{q} = (wq_{1,} wq_2 \ldots\ldots wq_j) \tag{2}$$

If all the vectors normalized, then the cosine of the angle between two vectors is the same as their dot-product. If vector $\vec{d}$ is the document vector and vector $\vec{q}$ is the query vector, then the similarity of document D to query Q (or score of D for Q) in equation (1) can be represented as:

$$sim(q, d_i) = \cos\theta = \sum_j w_{i,j} \times w_{q,j} \tag{3}$$

3. RESULTS AND ANALYSIS (10 PT)

Fuzzy MetagraphMultiagent Information Retrieval as shown in Figure 1.

Three agent as shown in Figure 1 connect with other to build fuzzy multi-agent information retrieval modelling.The input keyword writes by the user in the stage of user interface and the three agents used to retrieve document from Google. And in this section this agent user interface and multi- agent represented by the fuzzy metagraph where every agent is represented by sets. The user interface (UI) is represented by set $\{\tilde{X}_1\}$, by it user can enter the query (keywords). Agent 1 is represented by set $\{\tilde{X}_2, \tilde{X}_3, \tilde{X}_4\}$.Agent2 is represented by set $\{\tilde{X}_5\}$, Agent3 is represented by set $\{\tilde{X}_6, \tilde{X}_7, \tilde{X}_8, \tilde{X}_9\}$ and the output is the documents retrieval indexing in matrix contain terms of each document is represent by $\{\tilde{X}_{10}\}$.

A triple X in the FUZZY G in Figure 2 represents as:

$X = \{\tilde{X}_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_4, \tilde{X}_5, \tilde{X}_6, \tilde{X}_7, \tilde{X}_8, \tilde{X}_9, \tilde{X}_{10}\}$. The meanings of the set and variables used in this a triple explained in Table 1

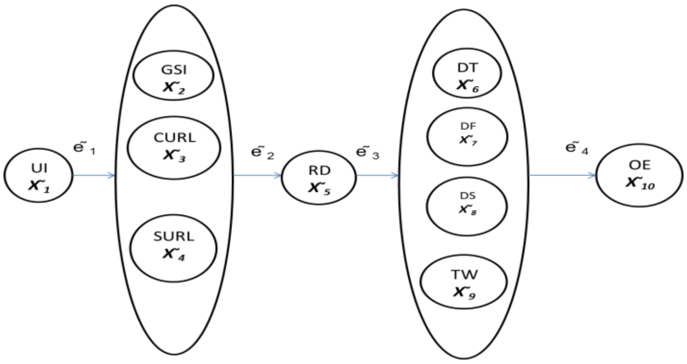


Figure 1. Fuzzy Metagraph for Each Agent Process of Multi-Agent Information Retrieval Molding

Table 1. Meaning of sets in Figure 5

Set	Variable	Meaning
$\widetilde{X}_1$	UI	User Interface to write keyword
$\widetilde{X}_2$	GIS	Google Search URL
$\widetilde{X}_3$	CURL	Google calculates URL
$\widetilde{X}_4$	SURL	Store ULR
$\widetilde{X}_5$	RD	Retrieve document from URL
$\widetilde{X}_6$	DT	DocumentTokenization
$\widetilde{X}_7$	DF	Document filtrating
$\widetilde{X}_8$	DS	Document Stemming
$\widetilde{X}_9$	TW	Term weight calculate
$\widetilde{X}_{10}$	OE	Output evaluation

The edge set can be specified as:

$\tilde{e}_1 = \{<\{\tilde{X}_1\}, \{\tilde{X}_2, \tilde{X}_3, \tilde{X}_4\}>\}, \tilde{e}_2 = <\{\tilde{X}_2, \tilde{X}_3, \tilde{X}_4\}, \{\tilde{X}_5\}>, \tilde{e}_3 = <\{\tilde{X}_5, \}, \{\tilde{X}_6, \tilde{X}_7, \tilde{X}_8, \tilde{X}_9\}>, \tilde{e}_4 = <\{\tilde{X}_6, \tilde{X}_7, \tilde{X}_8, \tilde{X}_9\}, \{\tilde{X}_{10}\}>$.As example,The in-vertex and out-vertex of $\tilde{e}$4 are In-vertex $=\{\tilde{X}_6, \tilde{X}_7, \tilde{X}_8, \tilde{X}_9\}$, *out-vertex*$=\{\tilde{X}_{10}\}$.

The simple path of the fuzzy Metagraph is represented as:

An important property of graphs is that of connectivity in Figure 1 there is a sequence of edges ($\tilde{e}_1; \tilde{e}_2, \tilde{e}_3; \tilde{e}_4$) that connects $\tilde{X}_1, \tilde{X}_{10}$, which means that a path from $\tilde{X}_1$ *to* $\tilde{X}_{10}$ exists.

Fuzzy rules are formed from fuzzy metagraph [9]. In this paper according to Figure 1 the following rules are used for fuzzy inference system (FIS) as will describe in the section 5.

$\tilde{X}_1$: weighting of term in query (tf-idf) before using membership function, w(q,t) € (0,1), w(q,t).

After membership function $\mu\, \tilde{X}_1$ € (x_1, μ)

$\tilde{X}_9$: weighting of term in document (tf-idf)) before using membership function, w(t,d), w(t,d) € (0,1) after membership function $\mu\, \tilde{X}_9$ € (x9, µ), The other sets are not used because they not have quantified values but they have qualified value.

4. PROPOSED MODEL OF FUZZY METAGRAPH MULTIAGENT INFORMATION RETRIEVAL DECISION MAKING

The structure of the proposed model is shown in Figure 2. The multi-agent is represented by fuzzy metagraph for simplicity and to understand the function of the multi-agent for retrieve information from

Google as a search engine using key word as (computer science) (show Figure 2 and section 3 illustrated that). Tthe first stage of the block in Figure 2 shows this. In the second stage the output documents from multi-agent after Document Tokenization, filtrating and stemming in agent three (show Table 1 and section 3) and calculate the weight term w(q,t) and w(t,q). There are two inputs and one output to fuzzy inference system [5] and two rules applied at fuzzy inference process. The input weight and rules and membership function considered in the third stage to fuzzifier and aggregation. The user can take from the documents retrieval what are you need and the output is ranking score documents

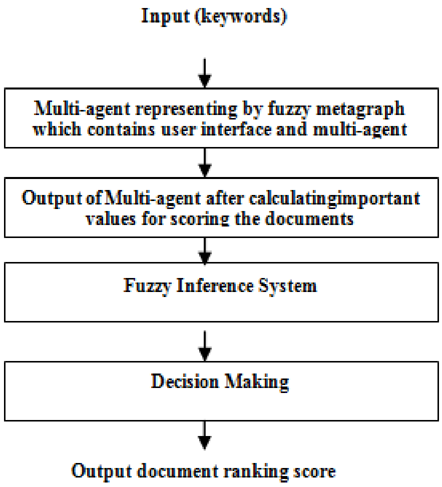

Input (keywords)
Multi-agent representing by fuzzy metagraph which contains user interface and multi-agent
Output of Multi-agent after calculatingimportant values for scoring the documents
Fuzzy Inference System
Decision Making
Output document ranking score

Figure 2. Block Diagram of the Proposed Model

5. EXPERIMENTAL EVALUATION EXAMPLE

The experiment evaluation was carried out in MATLAB platform, Mamdani-type FIS and Sugeno-type FIS and a sample of the documents retrieval from multi-agent. The experiment was ran to evaluate the ranking score of relevant documents using question (3) and keywords query as (computer science)

The input rule to the Mamdani-type FIS and Sugeno- type FIS arethe following:

1. If (w (q, t_1) is high) and (w (t_1, d) is high) then (cosine (q,d) is score high)
2. If (w (q, t_1) is low) and (w (t_1, d) is low) then (cosine (q,d) is score low)
3. If (w (t_2, d) is high) and (w (q, t_2) is high) then (cosine (q,d) is score high)
4. If (w (t_2, d) is low) and (w (q, t_2) is low) then (cosine (q,d) is score low)

Where W (t,q) is weight of term (tf-idf) of the term in the query ($Wq_{,j}$) and W(t, d) is the weight of term (*tf-idf*) of the term j in the document i W ($_{i,\ j}$) As in questions (1), (3). Also triangular membership functions was used for the linguistic input terms w (q, t_1), w (q, t_2), w (t_1, d) and w (t_2, d) as shown in the following example in Figure 3.

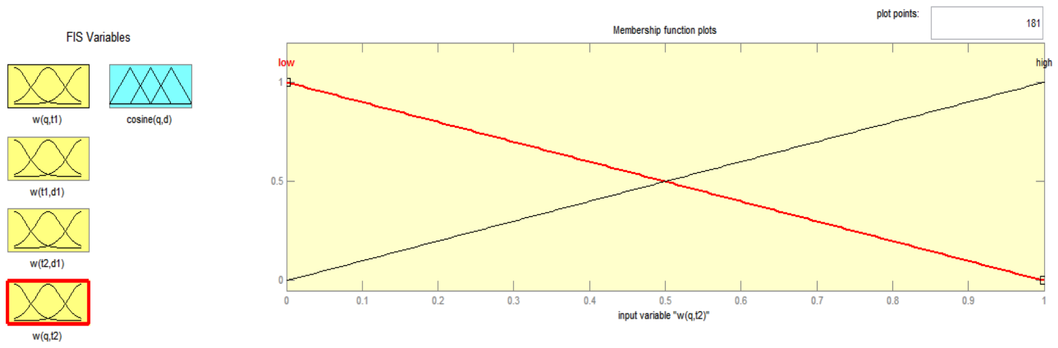


Figure 3. Example Membership Function of Input w (q, t_2)

5.1. Evaluation by Using Mamdani-Type FIS

The proposed FIS for the evaluation documents ranking score consists of four inputs (two for weigh term of document and two for weight term in the query) as shown in Figure 4 w(t$_1$,d),w(q,t$_1$),w(t$_2$,d),w(q,t$_2$). The system has one output that indicates score of document. Each of the selected input and output variables is described by a set of two linguistic fuzzy values (low and high) defined by triangle membership function, thus allowing the fuzzification procedure to convert the measured numerical value into one of the fuzzy values. Figure 3 shows one of the input w(q,t$_2$) triangle membership function and Figure 5 shows output score (cosine similarity) triangle membership functions. In the experiment the input processed by Mamdanii-type fuzzy inference system using triangle membership function and rules as previous described and as was shown in Figure 4. The input for the defuzzification process was the aggregate output fuzzy set (of sum after applied rules and the output set was a single number (centroid value) as shown in Figure 5.a and Figure 5.b. The document in Figure 5.b was highest score (centroid value 0.666) but the document in Figure 5 a was the lowest score (centroid value 0.573). The plots obtained after simulating Mamdani-type of FIS for document similarity cosine score were shown in Figures. 5.c and Figure 5.d.

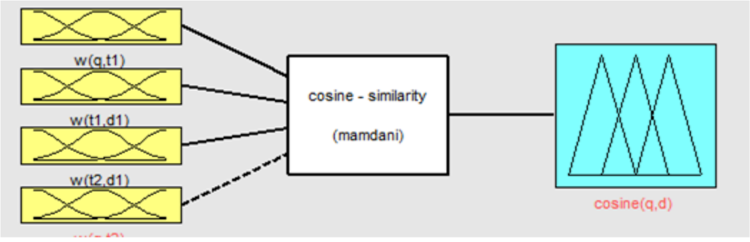
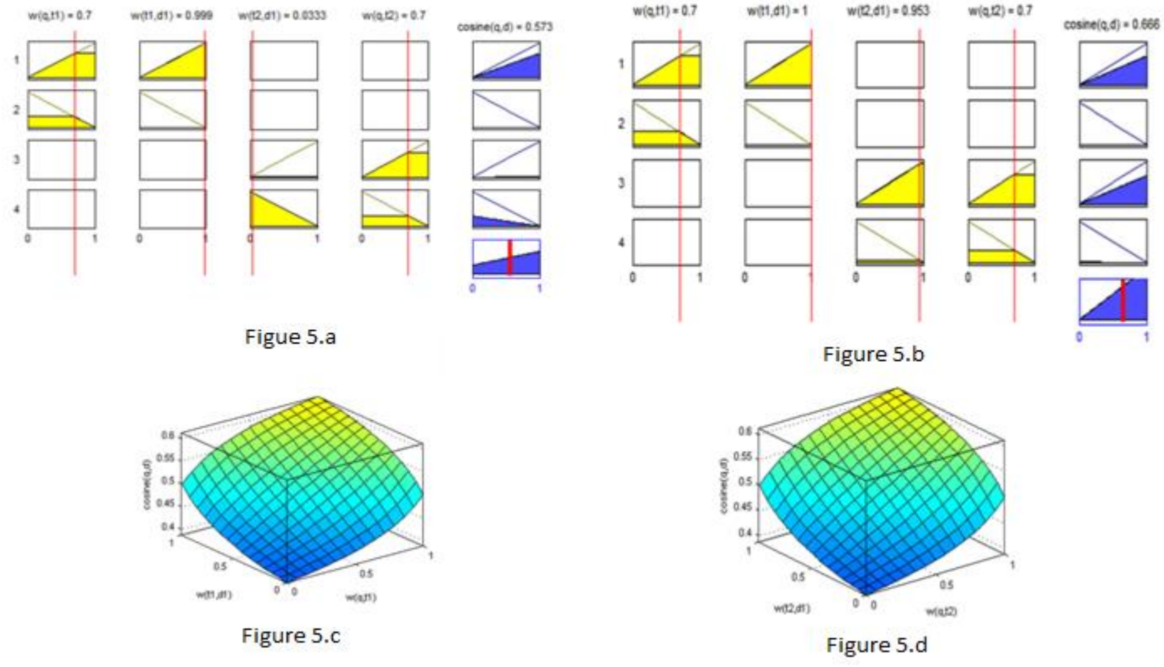


Figure 4. Mamdani –Fuzzy Type Fuzzy Inference System Using Membership Function and Rules



Figue 5.a

Figure 5.b

Figure 5.c

Figure 5.d

Figure 5. Result Experiment Output using Mamdani -type FIS

5.2. Evaluation by Using Sugeno -Type FIS

The initial steps and the setting of Sugeno-type FIS are same as of Mamdni-type FIS. It also consists of four inputs (two for weigh term of document and two for weight term in the query) as shown in Figure 6 w(t$_1$,d),w(q,t$_1$),w(t$_2$,d),w(q,t$_2$) and produces one output that indicates the similarity or the ranking score (cosine). Each of the selected input variables is described by a set of two linguistic fuzzy values, defined by triangle membership function as in the case of Mamdani-type fuzzy inference system(as already shown in Figure 3). Unlike the output value range of the Mamdani-type fuzzy inference system, the range of Sugeno-type output is between 0 and 1.The output of this system can only be either constant or linear in this FIS, so two linguistic fuzzy values for the output are “Low”, and “High” which can be constant low score 0 and high score 1. The rule base for Sugenotype FIS is the same as for Mamdani-type FIS. In the experiment the input processed by Sugeno –type fuzzy inference system using triangle membership function and rules as described. The output set was a single number (weighted average) as shown in Figure 7.a and Figure 7.b. The document in Figure 7. b was highest score (weighted average 0.999), but the document in Figure 7.a was the lowest score ((weighted averag 0.714). The plots obtained after simulating Sugeno -type of FIS for document similarity cosine score were shown in Figures 7.c and Figure 7.d.

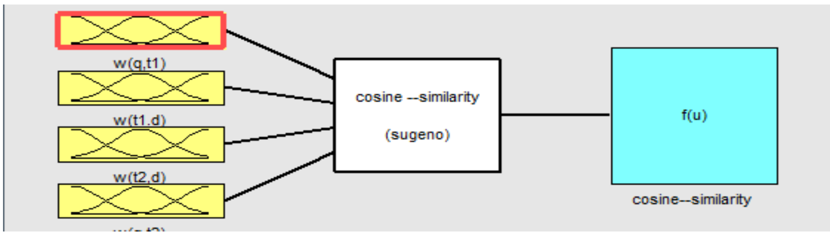


Figure 6. Sugeno –Fuzzy Type Fuzzy Inference System using Membership Function and Rules

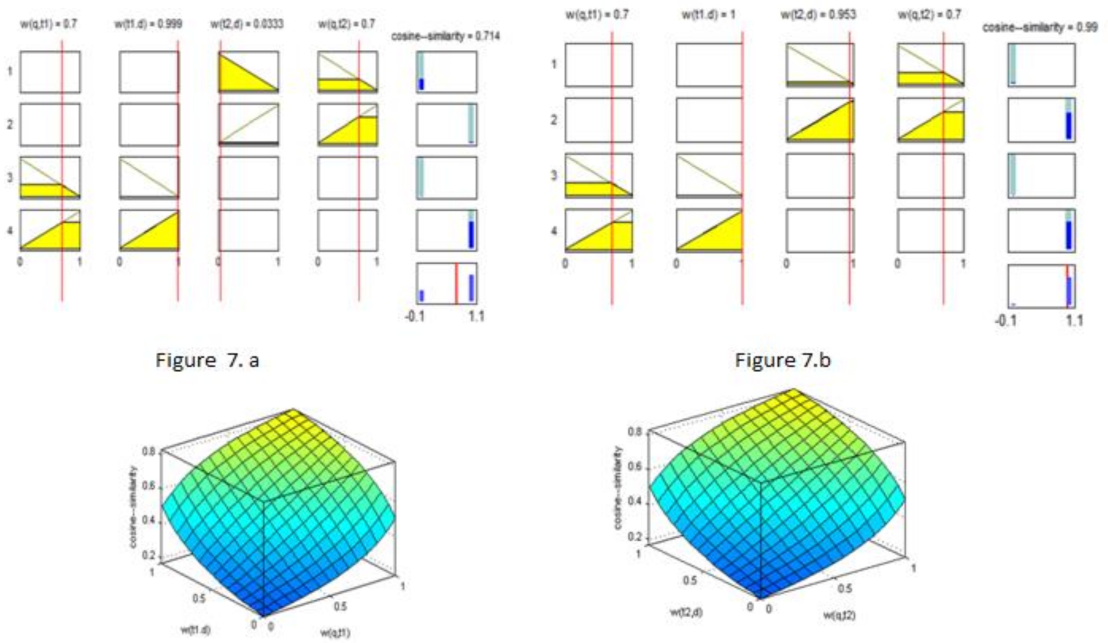


Figure 7. a

Figure 7.b

Figure 7.c

Figure 7.d

Figure 7. Result Experiment Output using Sugeno -Type FIS

6. CONCLUSION

Multi-agent system have presented to retrieves automatically multi-documents (text) and extract the useful information from the text information according to the users interests in a web-based environment by using keywords. Fuzzy Metagraph for Automatic Information Retrieval Multi agent modeling have been analyzed. We presented traditional method of defining the cosine measuring similarity between query and document to evaluate the document ranking score. Fuzzy Metagraph for Automatic Information Retrieval Multi agent modeling was combined with fuzzy inference system to construct a model for document ranking score. The documents ranking score cosine similarity using fuzzy inference system development and implemented much simpler than the traditional method which require mathematical equations. It has been concluded from this paper that for the evaluate document ranking score using similarity method (cosine), Mamdani-type FIS and Sugeno-type FIS works similarly. Membership functions and rules are same for both the FIS, only difference is that output membership functions for Sugeno-type FIS can only be either constant or linear and also the crisp output is generated in different ways for both the FIS. Sugeno-type FIS is better results better than Mamdani-type. Both the models are simulated using 4 rules and four input membership Functions. Also only one output value (centroid value) is used.in the case of Mamdani-type FIS and (average weight value) in the case of Sugeno-type FIs for document ranking score.

REFERENCES

[1] A Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. 2001; 24: 35–43.

[2] M Woolridge. Introduction to Multiagent Systems. John Wiley and Sons, 2001.

[3] A Thirunavukarasu, S Uma Maheswari. Fuzzy metagraph based clustering techniques. *Paripex-Indian Journal of Research*. 2013; 2: 117-119.

[4] Z Tan. Fuzzy Metagraph and Its Combination with the Indexing Approach in Rule-Based Systems. *IEEE Transactions on Knowledge and Data Engineering*. 2006; 18: 829-841.

[5] A Thirunavukarasu, S Uma Maheswari. Fuzzy metagraph based knowledge representation of decision support system. *International Journal on Computer Engineering and technology*. 2012; 3: 157-166.

[6] P Dashore, S Jain. Fuzzy Metagraph and Hierarchical modeling. *International Journal on Computer Science and Engineering*, 2011; 3: 435 –449.

[7] GS Ivanova, AM Andreev, VI Nefedov, MA Shouman, EV Egorova. Automatic search for information using multi-agent system. *Electromagnetic Waves and Electronic Systems*. 2015; 2: 33-38.

[8] P Raghavan, H Schütze, D Manning. Introduction to Information Retrieval. Cambridge University Press, 2008.

[9] A Thirunavukarasu, S Uma Maheswari. Technical Analysis of Fuzzy Metagraph based Decision Support System for Capital Market. *Journal of Computer Science*. 2013; 9: 1146-1155.