International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Improved Memetic Algorithm Enabled Intelligent Multi Agent (IMAEIMA) System for Web Mining

D. Weslin, T. Joshva Devadas

Abstract*: Wider web space, the searching of a relevant data is the most curious problem for the common people accessing the web. For retreving the relevant information the user request is given to search engine. The relevant pages combined with irrelevant pages are returned to the user. The proposed work emphasizes an Improved Memetic Algorithm Enabled Intelligent Multi Agent (IMAEIMA) for searching the most appropriate pages when submitting complex queries. Improved Memetic algorithm is the traditional genetic algorithm combined with local search and random selection. In this proposed system Improved Memetic algorithm additionally enhanced with logarithmic weight function for more accuracy. Intelligent Agents are introduced in this IMAEIMA to improve its performance and accuracy by reacting intelligently based on feedback and previous experience. This system helps to retrieve relevant pages from web databases with high precision and recall. The derived architecture reveals greater precision and recall overriding the conventional search algorithms.

Index Terms*: Genetic algorithm, , Intelligent Agent, Memetic Algorithm (MA), Web Mining.

I. INTRODUCTION

Recent years World Wide Web (WWW) has aggresive growth and thereby experiences several obstructions in retrieving the relevant information from the available data set. Recent trends like Semantic Web (SW), multi-agent systems, data mining, offers knowledge beyond information retrieval from the the Web. For that, various techniques have been proposed for retrieving the most accurate pages for the submittted complex query. Improper pharse or sentence with more number of keywords is termed as complex query. The search engine searches the relevant pages by choping the query into keywords. Most of the pages are retrieved by web crawler. These pages are may be relevant or irrelevant for results. In the proposed system, user query can be processed by the Improved Memetic Algorithm Enabled Intelligent Multi Agent (IMAEIMA) system and finds the relevant pages from the web. This system receives query from the user and checks the page existence in the knowledge data base (KDB), if exists then the relavant pages are immediately sent to the user. If the page doesnot exist in the knowledge data base (KDB) the Process Agent processes the query by using search engine for relevant snippets and thus to be processed by Improved Memetic Algorithm for the most relevant links. Intelligent Agent plays a vital role in many applications. There are basically five types of agents. In the proposed system the collaborative agents are introduced to improve its performance by reacting intelligently in the environment. The collaborative agents are functioning with two elements. The grasping element upgrades the system by critic feedbacks. The enforcement element executes external action based on feedbacks. And it does not repeat the tasks that are already being done. This character helps to measure the performance improvement.

The organisation of the remaining sections in the paper is as follows: Section II presents an outline of the existing research works related to discover the relevant of information from the web. Section III illustrates the Improved Memetic Algorithm Enabled Intelligent Multi Agent (IMAEIMA) system for retrieval of relevant web pages. Section IV presents the performance analysis and discussion of the IMAEIMA system. The scope of the proposed work is described in Section V.

II. RELATED WORKS

Gatial et al. (2013) proposed a system to collect the information and also integrates by semi-automated agent technique. Agent binds the methods for design and generation of form to end user [1]. Aarti Singh, (2012) focused on clustering techniques for mining information from the web based on agent [2]. The web information is grouped into clusters relevant to the query; satisfying user needs and yields better usage of web surfing time [2]. Shakti Kundu et al, (2017) introduced a new system called WEBMINTEL, is a multi agent system with three offers: reject, accept or add to cart, and these offers are intelligently executed based on prior knowledge [3]. Melita et al. (2012) reviewed the characteristics of Genetic Algorithms, GA based search techniques, problems in implementation and available methods suitable for web search optimization. Further, they suggested success rate improvements using enhancements in further research [4]. Kolli et al. (2013) proposed a new method to find the optimal patterens from the web. This technique combines Apriori algorithm and GA operators with association rule mining approach [9]. The unrequired patterns are removed and the target patterns are extracted from the frequent patterns [9]. Raval Pratiksha et al.

Revised Manuscript Received on August 30, 2019.
* Correspondence Author
D. Weslin*, Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India.
T. Joshva Devadas, Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Virudhunagar,Tamil Nadu, India

(2014) constructed dual clusters with greater consistence between the users and pages [5]. At last the most visited pages by all its users are retrieved. The users who react as simillarly on subset of pages are listed [5]. Chuang et al. (2014) proposed a web crawling approach for the extraction of automatic address and associated information for positioning the location entity on maps [6]. Das et al. (2015) presented a new technique for identifying the query forms from webpages [7]. Their technique exhibits high accuracy in the extraction of the desired data [7]. Pooja Solanki et al. (2015) recommended a system to predict the user's navigational behaviour using the biclustering approach with greedy search and genetic algorithm to overcome the problems of grouping available in traditional clustering approach [8]. Jayaprabha et al. (2016) exploded the possibility of genetic algorithm for extracting relevant information from the web, using fitness function and ranking function with genetic algorithm [10]. The parameters for this method are extracted from the web usage and web content mining with better results [10]. Panjwani Heena et al. (2017) system used coherent dual clustering framework to trace consistent dual clusters using the click-stream information [12]. This framework can be applied to enhance web page design, information access and feature of service provided [12]. The objective of this work found high volume dual clusters with great consistence of the user and visited pages. Wang et al. (2017) suggested a crawling method based on the document frequency of the queries to mitigate the ranking bias problem [11]. Ferrante Neri et al, (2012) focused on characteristic of memetic structure particularly the coordination of memes [13]. Surekha More et al, (2014) proposed genetic algorithm based intelligent approach for web mining. Genetic Algorithm combined with local searching technique enhances the web page search [14]. The similarity between several snippets is computed and the most similar snippets are returned. Khushali D et al. (2015) proposed improved memetic algorithm using hybrid selection strategy with local search for enhancing the results [15].

III. IMAEIMA ARCHITECTURE

Nowadays, several search engines provide the necessary information from World Wide Web data bases, but the outcome need to be refined to reach the accuracy. While using complex queries the issue of getting extraneous results is observed. Query with many key words is termed as Complex query. The proposed Improved Memetic Algorithm Enabled Intelligent Multi Agent (IMAEIMA) system for web mining is based on intelligent agents coupled with Improved Memetic Algorithm (IMA). The proposed system has four intelligent agents such are:

1. User Agent (UA)
2. Process Agent (PA)
3. KDB Agent (KDBA)
4. Good Link Agent (GLA)

The user query enters into the Intelligent Multi Agent (IMAs) system, the agent immediately triggers the searching in the knowledge data base (KDB) and the relevant page exists then responds it to the user. The relevant page is not hit by agent then the agent sends the user request to the conventional search process. The conventional search engine provides the snippets for the submitted query. Successive searches are done after preprocessing the snippets and query. These preprocessed snippets constitute the first generation for Improved Memetic Algorithm (IMA). Heuristic-approach will be taken over for local search process. Apply Term frequency matrix, Weight matrix, cosine similarity and fitness function for generating the next generations. The GA processes of selection, crossover and mutation will be carried out. End of the process the system extracts the positive and passive samples automatically through Good Link agent, and updates the KDB with new snippets. The user receives the new relevant results. A greater enhancement is obtained in the search process using an Intelligent Multi Agent with Improved Memetic Algorithm (IMAEIMA). Fig.1. shows the architecture of IMAEIMA system

Procedure of proposed system:

1. User enters the Query Q.
2. Q received by Intelligent Multi Agent system (IMAs).
3. Intelligent Multi Agent system (IMAs) checks the Q with the KDB.
4. If Q exists in KDB then send it to user;
5. Else
6. Begin:
7. Send Q to Search engine
8. Apply Improved Memetic Algorithm (IMA) for further generations, using conventional search engine results.
9. After completion of Improved Memetic Algorithm (IMA) process, the results are cheked by the Good Link agent.
10. The KDB updated by KDB agent and sends the results to user.
11. Stop the process.

A. Intelligent Multi Agents

An automated software component interacting in the system to replace human intervention is termed as agent. Based on prior experience the intelligence of the agent is trained to enhance the performance of the system. This intelligent agent finds vital role in innovative real world applications. The agents are characterized by self learning, based on knowledge and their previous experience added with knowledge updation. The proposed IMAEIMA system introduces four intelligent agents to improve the performance. These are User agent, Process agent, KDB update agent and Good Link agent. Fig.2 shows the communication in between agents.

User Agent (UA): User agent establishes the connection in between user and the system. It receives the Query from an user and sends the results the user. And also it communicates with KDB agent.

Process Agent (PA): Process agent communicates with search engine, KDB agent, IMA and Good Link agent. Process agent receives the Query from KDB agent and sends this Query to IMA for further generations. After this IMA process it receives the results from IMA and sends these results to Good Link agent.

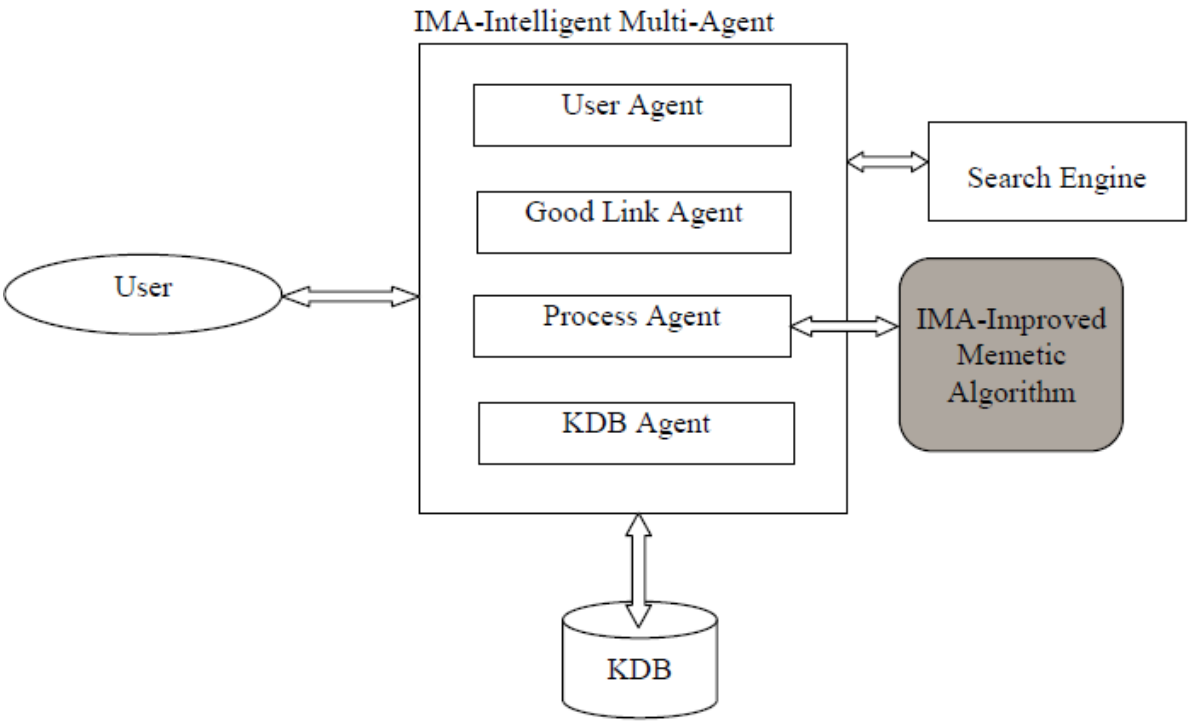


Fig.1. Architecture of IMAEIMA system

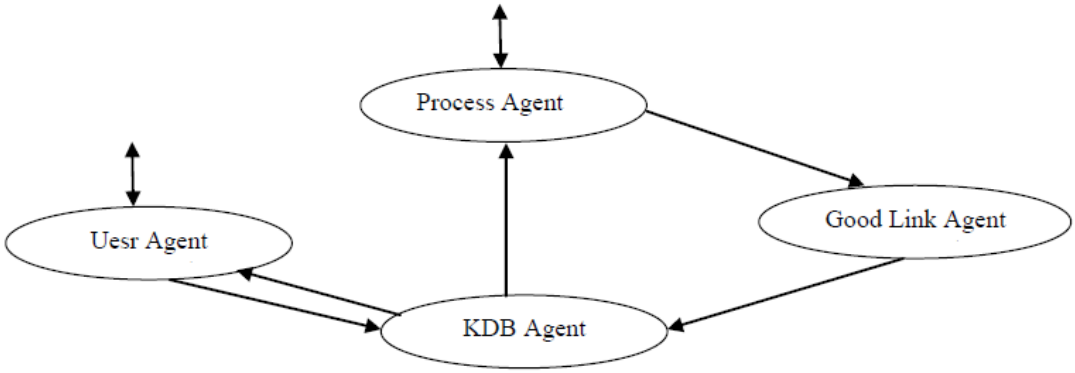


Fig.2.Communication in between Agents

KDB Agent (KDBA): KDB agent communicates with KDB, User agent, Process agent and Good Link agent. KDB agent receives the query from user agent and checks its availability in KDB, if exists reply to the user agent with results. Other wise send the query to process agent for further process. And also collects the results from Good Link agent for KDB updation.

Good Link Agent (GLA): Good Link agent communicates with KDB agent and Process agent. Good Link agent receives the results from PA and checks the positive and passive results. These results send to KDB agent

B. Improved Memetic algorithm approach

Memetic Algorithms are a population-based approach. Memetic algorithm is local search process embedded in evolutionary algorithm. GA taken for evolutionary process and heuristic function is used for local search process. Memetic algorithm is characterized by the same parameters involved in GA:

i) Population count

ii) Generation count,
iii) Crossover probability
iv) Mutation probability and along with a local search process. In addition to enhance the results hybrid-selection process is incorporated with MA is called Improved Memetic Algorithm (IMA). In IMA the population is initialized using a heuristic function or random method. A local search technique is employed to compute fitness value of each individual. The individual possessing higher quality of fitness forms the next generation population. New individuals are generated by crossover and mutation operation performed between two selected parents. The aspect of local search in memetic algorithm is to arrive at the local optimum accurately with minimum number of iterations.

Improved Memetic Algorithm Procedure:

Step 1: Initial population from problem space.
Step 2: Preprocesing of snippets and query
- Eliminate the stop words and spaces.
- Query transformed into keywords, K={k1,k2,k3,…km}

Step 3: local search process based on heuristic function

$$H(i)= \sum F_{ik} >= (1/n) (\sum F(i))\text{---------(1)}$$

where 'n' is number of snippets in the population.
Step 4: Calculate Term Frequency matrix using K and Snippets, row i represents snippet i is a term frequency vector

$$TF_i = \{TF_{ik1}, TF_{ik2}, TF_{ik3}, ..., TF_{ikm}\} \text{ ----------(2)}$$

The quality of snippet i is calculated as

$$F(i) = \sum_{i\in k} TF_{ik} \text{ ----------(3)}$$

The cosine similarity between snippet (i) and snippet(j) is tabulated from

$$s(i,j) = \frac{\sum_{i\in k} TF_{ik} * TF_{jk}}{\sum_{i\in k} TF_{ik} * \sum_{i\in k} TF_{jk}} \text{ ----------(4)}$$

Where F_{ik} is the frequency if keyword k in snippet i. The cosine similarity produces n x n matrix output for next generation of snippets.
From Term Frequency Matrix n x m (n-snippets, m-keywords in the query) generate weight matrix using, the keyword k in snippet i,

$$W_{k,si} = \begin{Bmatrix} 1+log_{10}(TF_{ik}), & if\ TF_{ik}>0 \\ 0, & Otherwise \end{Bmatrix} \text{--------(5)}$$

From S(i,j) and $W_{k,si}$ can be used for finding fitness function to measure the performance of snippet i and snippet j, it rewards the next generation chromosomes.

$$F(i,j) = \sum \frac{W_i}{W_q} s(i,j) \text{ ----------(6)}$$

Where Wi is the weight of the Q in node i and Wq is the total number of query terms in Q.
Step 5: After fitness computation the snippets are ranked to generate chromosomes passed on to next generation.
Step 6: After selection process the Crossover and Mutations are performed.
Step 7: Iterate step 3 to 7.
Step 8: Send the final results.

B.1. Hybrid-Selection strategy

In IMA the selection process follows hybrid stratergy. Initially the random population is generated. The chromosomes are sorted in order in one pool and in random in the other pool. Fig.3. explains the hybrid selection strategy. The hybrid-selection order is indexed from 1 to the size of population, denoted by i. From each set the chromosome at first position is selected. The hybrid strategy is retrieving best chromosome by tournament selection between the first chromosomes of each set P and RP. Each chromosome particiapates in the selection process, so that the chromosome with least fitness value also survives. Crossover is done several times between the two parents so that the best characters of the parents are reflected in next generation. Normally in crossover operation the child is carried out to next generation avoiding the features of the parents. The least offspring fails to reach the local minima. In the proposed algorithm the features of the parents are completely inherited by selection process while the child chromosome is crossed with the parents.

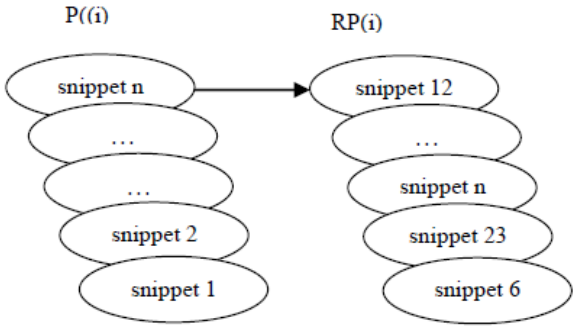


Fig.3. Hybrid Selection Sort

B.2. Cross over

GA is an evolutionary algorithm used to build an optimum result by combining the chromosomes together based on its good characters. This system uses a single point cross over technique for offspring generation. In single point cross over technique one cross over point is selected, using this cross over point swapping segments of right side of two parents produce the two off springs. In the following example 5 th position is the cross over point. So, right sides of the two parents are swapped to get offsprings.

Parent 1

1	2	3	4	5	6	7	8

Parent 2

5	6	3	1	7	8	2	4

Child 1

1	2	3	4	7	8	2	4

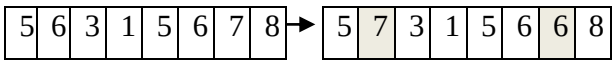
Child 2

5	6	3	1	5	6	7	8

B.3. Mutation

During mutation step, the selected individuals are used as parents to produce the children of the succeeding generation. The objective of mutation operator is maintaining diversity. The chromosomes in the next generation may be typical or atypical to the parent chromosomes. The choromosomes are discarded in the next generation, if they are atypical. The population progresses toward an optimal solution over the successive generations.

In this system order of changing technique is incorporated, in which two positions are selected and interchanged. In the below example 2nd and 7th elements are interchanged to generate offspring.



IV. PERFORMANCE ANALYSIS AND DISCUSSIONS

The proposed IMAEIMA architecture is an improvement over MA and IMA, which does not use an intelligent agent technology experience to retrieve the relevant information. In the proposed IMAEIMA architecture, the query is received by intelligent multi agent and checked with KDB for existence. After existence checking the user request is executed by conventioanl search process and first generation is computed. From this the term frequency is calculated using function (2). The quality of population is calculated using function (3). For local search process the hueristic function (1) is used, and select the population for further process. After, selection the fitness values of relevant documents to be calculated using fitness function (6) with help of weight function (5) and cosine similarity function (4). Based on fitness thershold value the best populations are processed using GA operations such as crossover and mutation. The new offsprings are added with existed population. This process continues till reach the threshold condtion based on number of iterations. These outputs are received by the intelligent multi agent and are cheked by the Good Link agent (GLA) for the positive and passive results. These results will be stored in KDB by KDB agent and simultaneously send it to user. Thus, use of intelligent multi agent reduces the over head of the system. In the IMA process the weight matrix for each keyword is calculated using logarithmic function. The logarithmic weight function helps to reduce the larger values into small values for easy and fast calculations. This provides more accuracy for finding fitness values. The local search with hybrid selection of the population and quality of population reduces the iterations.

The intelligent agent checks the final results by GLA once again reduce the error rate of non relevant documents. In information retreival system the performance can be measured by Precision and Recall. The measure of documents relevant to search is termed as Precision. Recall is the measure of the documents that are relevant to the retrieved user request.

$$precision = \frac{|\{Relevant\ documents\} \cap \{Retrieved\ documents\}|}{|\{Retrieved\ documents\}|}$$

$$Recall = \frac{|\{Relevant\ documents\} \cap \{Retrieved\ documents\}|}{|\{Relevant\ documents\}|}$$

This system uses 10 complex queries and the cross rate is 0.5 and mutate rate is 0.01 with 17 iterations. The result values are plotted in F- score graph by the experience based on the significance of the retrieved pages. These results are compared with MA and IMA results.

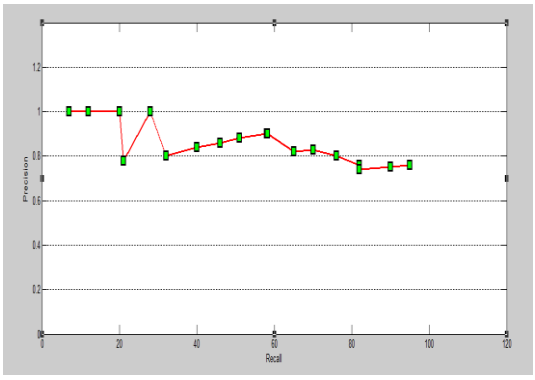


Fig.4.Precision Vs.Recall Metric for MA

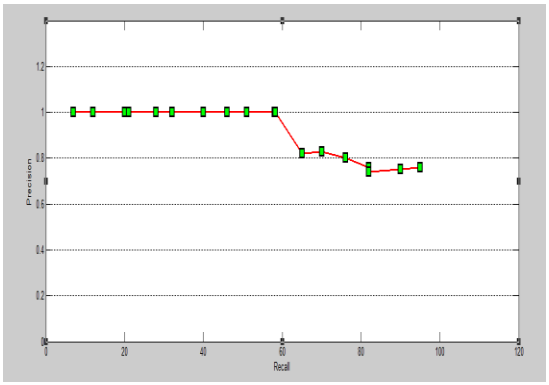


Fig.5.Precision Vs.Recall Metric for IMA

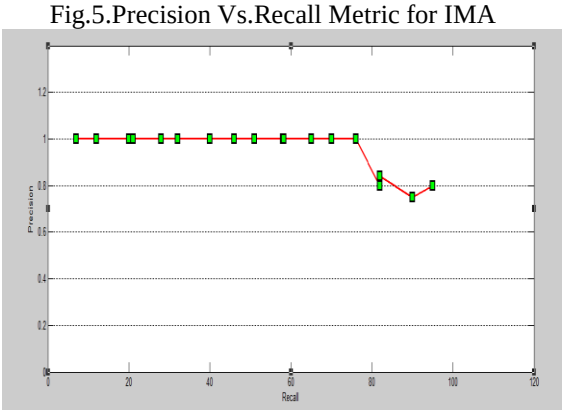


Fig.6.Precision Vs.Recall Metric for IMAEIMA

The relevancy of each snippet is the measure of its precision and recall values. The plotted results of precesion versus recall are depicted in the following Fig.4, Fig.5 and Fig.6 respectively. From these results the proposed IMAEIMA system is produced significantly better results.

V. CONCLUSION

Thus the implementation of a search engine using intelligent multi agents with improved memetic algorithm additionaly enhanced with logarithmic weight function prodeces more accurate results. The multi agents are intelligently reacts in the environment based on its prior experience and its frequent feedback system. And also it reduces the error rate using Good Link Agent. In future the improvement of the system may be based on search time and also implementation of multi search engine instead of only one searh engine and cross over combined from different search engine results.

REFERENCES

1. E.Gatial, Z. Balogh and L. Hluchy, "Agent-based Information Integration using Generated User Forms", IEEE 17th International Conference on Intelligent Engineering Systems, 2013, June 19-21.
2. Aarti Singh, "Agent Based Framework for Semantic Web Content Mining" , International Journal of Advancements in Technology http://ijict.org/ ISSN 0976-4860, Vol. 3 No.2, April 2012, IjoAT.
3. Shakti Kundu and M.L. Garg, "Web Data Mining through Software Agents" International Journal of Computer Applications (0975–8887) Volume 166 – No.5, May 2017.
4. L. Melita, Gopinath Ganapathy and Sebsibe Hailemariam, "Genetic Algorithms: an Inevitable Solution for Query Optimization in Web Mining – a Review" , The 7th International Conference on Computer Science & Education -ICCSE 2012, July 14-17, Melbourne, Australia-IEEE.
5. M. Raval Pratiksha and Mehul Barot,"A Web Page Recommendation system using GA based biclustering of web usage data", International Journal of Advance Engineering and Research Development (IJAERD) Volume 1, Issue 5, May-2014, e-ISSN: 2348 - 4470, print-ISSN:2348-6406.
6. H.M.Chuang, C.H. Chang and T.Y.Kao, "Effective web crawling for Chinese addresses and associated nformation", International Conference on Electronic Commerce and Web Technologies, Springer, 2014 pp.13–25.
7. N.N. Das and E.Kumar, "Automatic extraction of data from deep web page", International Journal of Computer & Mathematical Science, 2014, Vol. 3, pp.86–91, ISSN-2347-8527.
8. Pooja Solanki and Jasmin Jha, "Web Page Recommendation System using Biclustering with Greedy Search and Genetic Algorithm", International Journal of Innovative research in Technology - (IJIRT), June-2015, Volume 2, Issue 1, ISSN: 2349-6002.
9. Kolli Prabhakara Rao and G.Kalyana Chakravarthy, "Intelligence Service Of Web Mining With Genetic Algorithm", International Journal of Engineering Trends and Technology (IJETT) –Oct-2013, Volume 4, Issue 10, E-ISSN: 2231-5381, P-ISSN: 2349-0918.
10. P.Jayaprabha , Dr.Paulose Jacob , Dr.Preetha Mathew and P.K.Bindu, "Implementation of a Model for Web Mining Based on Web Usage", IOSR Journal of Computer Engineering (IOSR-JCE), E-ISSN: 2278-0661, P-ISSN: 2278-8727, 2016, PP 65-71.
11. Y.Wang, J. Lu, J. Chen, J. and Y.Li, " Crawling ranked deep web data sources", World Wide Web (ACM), 2017, Vol. 20, No. 1, pp.89–110.
12. Panjwani Heena and Pooja Jardosh, "Webpage recommendation in web usage mining using genetic algorithm" IJARIIE-ISSN(O)-2395-4396, 2017, Vol-3 Issue-3.
13. Ferrante Neri and Carlos Cotta, "Memetic algorithms and memetic computing optimization: A literature review", Swarm and Evolutionary Computation, 2012, Vol.2, P.No. 1–14, Elsevier.
14. Surekha More and Ujwala Bharambe , "Intelligent Web Mining Technique using Evolutionary Algorithms", International Conference on Isues and Challenges in Intelligent Computing Techniques (ICICT), 2014, P.No. 508 - 514 (IEEE).
15. Khushali Deulkar and Meera Narvekar (2015), "An Improved Memetic Algorithm for Web Search", International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), Science Direct , Procedia Computer Science 45, 52 – 59.

AUTHORS PROFILE

D. Weslin received his B.Sc in Special Mathematics and M.C.A in Master of Computer Application from the Madurai Kamaraj University in 1997 and 2001, respectively. Also, he received his M.E in Computer Science and Engineering from the Anna University, Chennai in 2006. He is currently working as an Associate Professor at Vickram College of Engineering and a Research Scholar in the Bharathiar University, Coimbatore, Tamil Nadu, India.

Dr. T. Joshva Devadas received his B.Sc in Special Mathematics, M.Sc in Computer Science and M.Phil in Computer Science from the Madurai Kamaraj University in 1984, 1996 and 2007 respectively. Also, he received his M.Tech in Computer Science and Engineering from the Pondicherry University in 2001 and completed his Ph.D in Computer Science from Madurai Kamaraj University in Madurai in 2010. He is currently a Professor in the Department of Computer Science and Engineering, School of Computing at Kalasalingam Academy of Research and Education, Virudhunagar, Tamil Nadu, India. His research interests include data mining and intelligent agents. He published more than fourteen papers in his research field.