MetaFusion: An Efficient MetaSearch Engine using Genetic Algorithm

Dr. Daya Gupta
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
daya_gupta2005@yahoo.co.in

Devika Singh
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
devika.singh91@gmail.com

***Abstract*— World Wide Web is a dynamic source of information which is expanding its content at a staggering rate. Individual search engines are not able to handle the exponential nature of web. Hence meta-search engines are used to solve the problem of low web space information coverage rate of individual search engines. A meta-search engine is a kind of search tool that dynamically dispatches user query to the underlying search engines, hence providing parallel access to multiple search engines and then aggregate the results to present single consolidated result list to user. In this paper , a novel meta-search engine, *MetaFusion* , has been proposed. The proposed algorithm combines fuzzy AHP with genetic algorithm to get more comprehensive and optimized results. Experimental results shows that relevancy of results returned by MetaFusion is more than several existing research Metasearch engines. The precision of MetaFusion is more when compared with Dogpile and Infospace.**

Keywords*— *MetaFusion, Information retrieval, metasearch, search engine,MCDM, Fuzzy AHP,Genetic Algorithm

I. INTRODUCTION

These days common man are using World Wide Web to search needed information using variety of Search engines like Google ,Bing , Yahoo, Ask , Lycos etc. World Wide Web is a huge repository of data consisting of billions of web documents that are distributed over multiple web servers. The information on web is increasing day by day and due to this web coverage given by individual search engines have been constantly decreasing[1]. For individual search engines only return 45% of relevant results[2]. The research in last decade has focused on improving the search procedure. As a result meta-search engines have been proposed. Meta-search engine is basically considered to be a fusion tool which commences its session when user poses query to its interface. After that MSE processes the query and submits the refined query to multiple underlying search engines. The underlying search engines accesses the network resources and then return back their respective results to MSE. Then MSE aggregates the returned results into single consolidated rank list by using certain aggregation algorithm. Various research Meta-search engines are available online like Dogpile , Infospace , Excite, DuckDuckGo etc. which are used by users to serve their needs. The basic functioning of meta-search engine is shown in Fig. 1. [3]

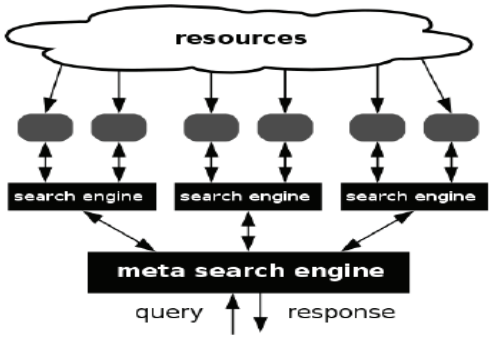

Fig. 1: Meta-search Engine Functioning [3]

Thus, a meta-search engine is an improvement over a single search engine since it broadens the search coverage and increases search accuracy by allowing the extraction of more appropriate results with the same amount of effort. The advantages of using MSE is: 1)increases search effectiveness. 2) improves search accuracy. 3) increases users convenience by allowing him to access multiple search engines for just one query. 4) solves real time search issues related to web search.

Early meta-search engine models like MetaCrawler[4,5], Borda-Fuse[6] etc. were centered around assigning weight scores to documents , thereby using relatively straightforward technique for result merging. Then came next generation of MSEs like Weighted Borda Fuse[6] where individual search engines are also assigned weights to reflect their performance. Recent models of MSE,[7] were based on OWA and use multi –criteria decision making. It addresses the issues related to missing documents. If a document is missing from one search engine's rank list but present in other, it does not makes it less relevant because different search engines cover different portions of web search space. Another recent model of MSE ,MetaSurfer [8], was proposed by Tayal et al. in 2014,uses modified EOWA along with FAHP to perform metasearch. Very recently MetaXplorer[9] , was proposed by Daya Gupta and Neha Dimri in 2015 to allow for imprecise and uncertain comparisons. It was based on Intelligent OWA operator along with FAHP to evaluate document score.

In this research we are improving forgoing MSE by applying multi optimization algorithm. This paper present a new meta-search engine, named as MetaFusion, which is capable of handling dynamic web environment. The proposed

978-1-5090-3251-8/16/$31.00 ©2016 IEEE

algorithm uses Fuzzy Analytical Hierarchy Process(FAHP) along with Genetic algorithm to retrieve comprehensive results. FAHP reflects human thinking and addresses the uncertainty of information while making decisions in MCDM problems. Genetic Algorithm is a self adapting global optimization parallel search algorithm which imitates biological evolution process i.e. crossover ,mutation, selection [10].The main challenge during information retrieval is to find most appropriate set of documents with respect to user query. Hence Genetic Algorithm(GA) is used for result merging which uses average weight of document in underlying search engine as fitness function. The documents are ranked in decreasing order of their fitness value i.e. most relevant document have higher fitness value and is present at top position in rank list.

The rest of the paper is organized as follows: Section II describes related work. Section III describes the proposed work. Section IV presents experimental results and Section V concludes the paper.

II. RELATED WORK

In this section various previous meta-search engine models have been discussed.

MetaCrawler is one of the first meta-search engines developed by Erik Selberg and Oren Etzioni at the University of Washington, Seattle in 1995[4,5]. The steps involved in ranking computation is described below:

1. User query is processed and forwarded to underlying search engine's such as Lycos, Excite, Yahoo etc.
2. The documents are assigned weights i.e "confidence score" such that top most document in each search engine's result list gets highest value of confidence score.
3. Then results are merged by adding corresponding values of confidence scores.
4. Finally the duplicates are removed and result is displayed to user.

Borda-Fuse Model was proposed by Aslam and Montague [6] in 2001 for result aggregation. The steps involved in ranking computation is described below:

1. Each search engine is considered as a voter , which ranks a set of *n* documents according to their relevance.
2. The top most document is assigned n points , the second one is assigned n-1 points and so on procedure continues.
3. The documents that are missing in search engine result list are assigned remaining points evenly.
4. Then for each document , we add the corresponding Borda Points obtained from different search engines .
5. Display the result in decreasing order of Borda Point value. Hence top most document has highest Borda Points.

The major drawback of Borda-Fuse Model is that it considers election process to be homogeneous. Hence came new model of MSE i.e. Weighted Borda Fuse in which different weights are assigned to different search engines based on their performance[6].

Commercially available MSE such as Dogpile uses hybridized combination of parallel and serial techniques to perform metasearch[11].User query is pre-processed and dispatched to multiple search engines. Certain intelligent processing algorithm such as duplicate detection and removal, ranking etc. is applied onto returned results and document in decreasing order of preference is displayed to user. Another commercially available MSE Infospace does intelligent predictive search analysis on returned results to rank the documents[11].

Very recent model of MSE ,MetaXplorer [9], was proposed by Daya Gupta and Neha Dimri to allow for performing metasearch on user query. The steps involved in ranking computation is described below:

1. User posted query is refined and dispatched to several underlying search engines.
2. In-OWA operator was used to assign importance degree to search engines.
3. Missing document's are assigned weight by taking weighted mean of that document in those search engines where they appear.
4. FAHP is applied to perform pair-wise comparison of documents and hence document scoring is done.
5. The overall document preference is obtained by multiplying document score with search engine's importance degree.

III. PROPOSED WORK

The aim of our proposed meta-search engine, MetaFusion is to present a set of relevant documents with respect to user query. We propose an approach which uses Fuzzy AHP along with Genetic Algorithm to select most appropriate set of relevant documents. Now we discuss some basic concepts related to OWA operator ,Fuzzy AHP and Genetic Algorithm. This section present overview of some related terms used in our model.

A. OWA Operator:

OWA operator of dimension n is defined as a function F: $R^n \rightarrow R$(where R = [0,1]), with associated weighing vector W , where $W=[W_1\ W_2\ W_3 \ldots W_N]$, such that

1) $W_i \in [0,1]$
2) $\Sigma\ W_i = 1$
3) $F(A_1,A_2,A_3 \ldots .. ,A_n) = W_1*B_1 + W_2*B_2 + W_3*B_3 \ldots\ldots + W_n*B_n$

where Bi is the i^{th} largest value in $A_1, A_2, ...,A_n$.

B. Fuzzy AHP

Thomas Saaty proposed Analytic Hierarchy Process (AHP) , to solve complex multi-criteria decision making problems[12]. Saaty demonstrated the benefits of applying pair wise comparison in MCDM problems[13]. AHP is a divide and conquer based method which selects one or more available alternatives and ranks them based on several criterias each having different importance values. In AHP we break a problem into sub-problems and then solve these sub-

problems one by one [14]. But, AHP can't handle the uncertainty and imprecision associated with the decision maker's perception. Therefore in 1988 Fuzzy Analytical Hierarchy Process(FAHP) was introduced which reflects human thinking while making decisions[15]. Fuzzy AHP uses linguistic quantifiers while making comparisons instead of crisp numbers . Hence, crisp judgments gets transformed into fuzzy judgments. Triangular fuzzy number (TFN) is expressed by three real numbers i.e. (l,m,u) where l denotes the minimum value, m denotes most expected value and u denotes the maximum value that describes fuzzy event.

C. Genetic Algorithm

Genetic algorithm is a heuristic search based technique that is inspired by Darwin's theory and mimics natural evolution process[16]. Genetic Algorithm(GA) starts its operation with a set of random candidate population i.e. initial population. Then we apply various genetic operations like crossover, mutation and selection on this set of solution . The process continues iteratively until we get optimized result or maximum number of generations have been reached. GA is highly robust and self- adaptive algorithm ,hence, solves the more complex problems in optimum manner. The flowchart of Genetic Algorithm is shown in Fig. 2.

The GA has been applied in various domains of real world, for example: Multi objective optimization [17], Feature selection by applying multi-objective genetic algorithms [18], Job Scheduling problems [19], Wireless Sensor Networks [20],and Cloud Computing [21] etc.

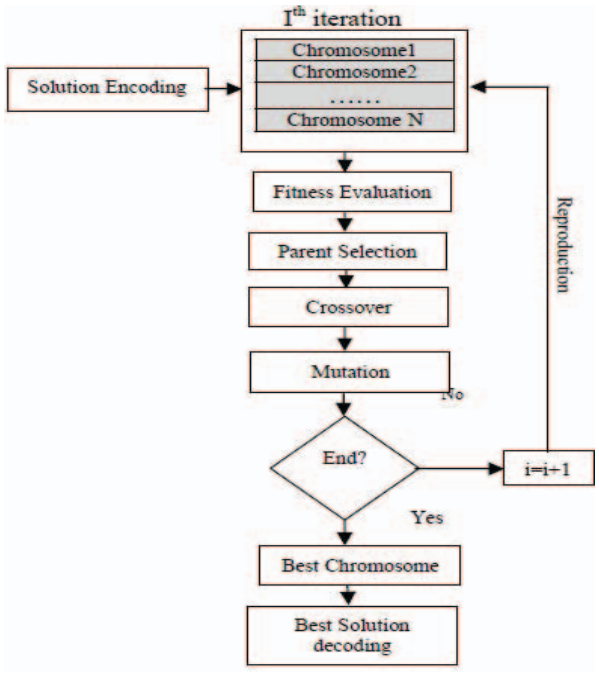

Fig. 2: Flowchart Of Genetic Algorithm

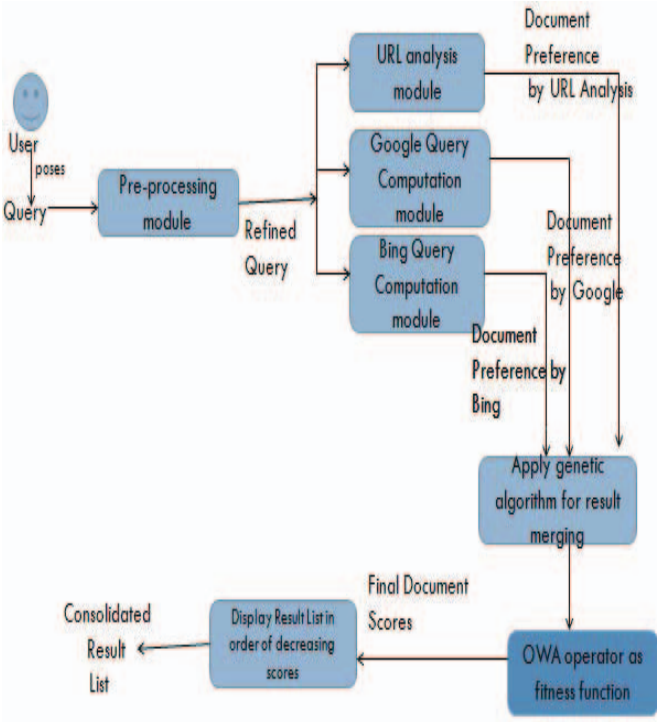

Fig. 3: Working of Proposed approach

D. Proposed Algorithm:

The proposed model of MetaFusion mainly consists of two phases i.e. Training phase and Query Execution phase. In our proposed MSE we have considered two underlying search engines where query is dispatched i.e. Google and Bing. The working of our proposed Meta-Search Engine is shown in Fig. 3.

The training phase allows to learn search engine's importance degrees through examples and thus, makes the proposed model adaptable to the changing environment. The training phase assigns weight values to Google and Bing according to their performance. The training algorithm could be run periodically or as per user feedback in order to reflect the changes in the environment.

Steps involved during Query Computation Phase is shown below:

Step 1: User submits query to MetaFusion interface, and then preprocessing module refines query to remove redundant terms, stopwords etc.

Step 2: The refined query is forwarded to Google and Bing Query Computation module where Fuzzy AHP is applied to their respective result list. Fuzzy AHP consist of following steps:

- Form pair-wise comparison matrix of size N×N by using TFN (l,m,u) where N is total number of unique documents present in Google and Bing.TFN are depicted as follows:
  a. Least Important (LTI) (1,1,3)
  b. Less Important (LSI) (1,3,5)

c. Equally Important (EI) (3,5,7)

d. More Important (MEI) (5,7,9)

e. Most Important (MSI) (7,9,9)

- Apply alpha-cut method to form interval performance matrix [αleft, αright] and is computed as follows:

$$\alpha left = [\alpha*(m-l)]+l \quad (1)$$

$$\alpha right = u-[\alpha*(u-m)] \quad (2)$$

where α is confidence factor which ε between[0,1].

- Obtain Crisp Judgement Matrix ,C_λ , by using following equation:

$$C_\lambda = \lambda*\alpha right+(1-\lambda)*\alpha left \quad (3)$$

where λ is optimism index of decision maker and λ ε[0,1].

- Then, Normalize C_λ by dividing each element by corresponding column-sum. After that add each row to get document score.

- Multiply document score with Google weight and Bing weight.

Step 3: The URL analysis is also performed simultaneously on the returned results. The contents of research paper or journal is consider to be more relevant than content of textbook.

Step4 : Then Genetic Algorithm is applied next to merge the result and form the single consolidated rank list of documents based on the fitness value of document. In the proposed work OWA operator have been used as Fitness function. The steps involved in Genetic Algorithm is shown below:

A. Calculate OWA operator weight using following equation:

$$W_i = \left(\frac{i}{n}\right)^{\propto} - \left(\frac{i-1}{n}\right)^{\propto} \quad (4)$$

where n is number of criteria and α ε[0,1].

B. Now compute the *Ordered Weighted Average* and average for same using following equation:

$$OWA_{i,j} = \sum_{j=1}^{N} w_i \times dscore_{i,j} \quad (5)$$

where dscore is document score obtained by FAHP.

$$Avg(OWA_{i,j}) = \frac{\sum_{k=1}^{X^U} OWA_k}{X^U} \quad (6)$$

where X^U is maximum number of documents a search engine can return.

C. Fitness of an individual document is obtained as follows:

$$fitness(i) = \begin{cases} \beta, CP > fitness(i) \| fitness(i) = \max Fit \\ \left| \frac{Avg(OWA)_i - CP}{\max Fit - CP} \right|, otherwise \end{cases}$$

where β ε[0,1] is a real random number, CP denotes cut point as $0.5 \times maxFit$ and $maxFit$ is the maximum value of $Avg\ OWA$.

D. Apply crossover operation and generate offspring chromosomes. In proposed work three point crossover is used.

E. Apply polynomial mutation operation over generated offspring vector to bring diversity in population.

F. Evaluate newly generated chromosomes and select best amongst them based on basis of their fitness value i.e. "survival of fittest".

IV. EXPERIMENTAL RESULTS

The proposed model of MSE combines the results of two well known search engines i.e. Google and Bing. We have compared the performance of our proposed model with existing research MSEs by considering over 100 test queries from different domains of real world. Some of these sample queries are listed in Table I.

The performance of MetaFusion is compared with existing MSEs i.e. Dogpile and InfoSpace in terms of precision. The interface of MetaFusion is shown in Fig. 4.

Neurobiology	Cognitive Science	Multi Agent System	Photosy-nthesis
Expert System	Branch Prediction	Human genetics	Robotic Fusion
Optical technology	Image forensics	Semantic Analysis	Partical Swarm Intelligence
Branch Prediction	Pattern Recognition	Cryptography	Ontology

TABLE I : SOME SAMPLE QUERIES

Fig. 4: Proposed model "MetaFusion" Interface

The results obtained by MetaFusion , Dogpile, InfoSpace on the query “Cosmochronology” is shown in Fig 5 ,6 and 7 respectively.

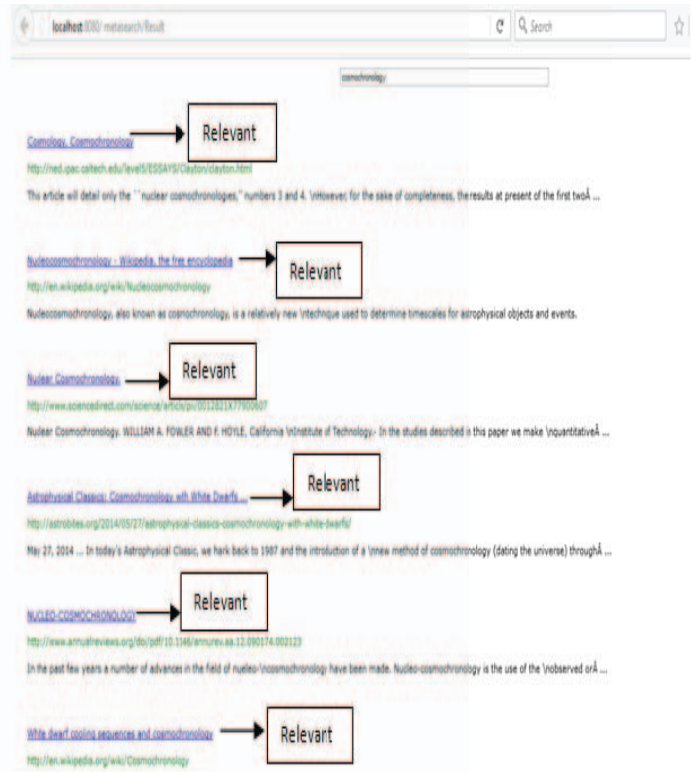

Fig. 5: Results of Proposed Approach “MetaFusion”

Fig. 6: Dogpile Results

Fig. 7: Infospace Results

The precision value of our proposed MetaFusion model is greater than existing MSE which is shown in Fig. 8.

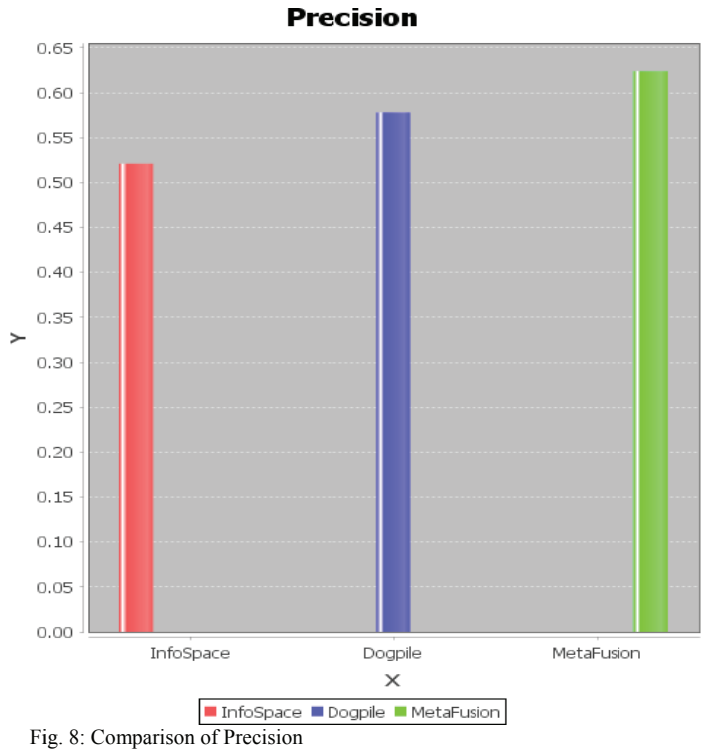

Fig. 8: Comparison of Precision

V. CONCLUSION

World Wide Web contains enormous number of documents and is major source of information dissemination. It is very challenging task to retrieve relevant set of documents from large database. Hence our proposed meta-search engine model ,MetaFusion , tries to retrieve the relevant information according to user query. The previous model of MSE, MetaXplorer uses FAHP for result aggregation whereas our proposed algorithm uses Fuzzy AHP along with Genetic algorithm to generate the aggregated rank list and arrange the documents in order of their decreasing fitness value i.e. "survival of fittest". Hence the document at the top will be having higher fitness value. In our proposed work Genetic algorithm is applied for multi-optimization and thus it is more efficient. The performance of MetaFusion is compared with available research meta-search engines over set of 100 test queries which are taken evenly from different research domains. The results shows that MetaFusion has the highest precision of 0.6341 when compared with available research MSEs Dogpile and Infospace. In future , we can extend this research work by considering more number of search engines in result aggregation. Also in future , we can apply any other multi-optimization algorithm for result merging.

REFERENCES

[1] R.M. Losee, "When information retrieval measures agree about the relative quality of document rankings", Journal of the American Society of Information Science, vol. 51, issue 9, pp. 834-840, 2000.

[2] A.H. Keyhanipour, B. Moshiri, M. Kazemian, M. Piroozmand, and C. Lucas, " Aggregation of Web search engines based on users' preferences in WebFusion", Knowledge-Based Systems, vol. 20, issue 4, pp. 321–328, 2007.

[3] Available Online at: http://en.wikipedia.org/wiki/Metasearch_engine

[4] E. Selberg, O. Etzioni, "Multi-Service Search and Comparison Using the MetaCrawler", Proceedings of the 4th International World Wide Web Conference, pp. 195-208, 1995.

[5] E. Selberg, O. Etzioni, "The MetaCrawler Architecture for Resource Aggregation on the Web", IEEE Expert, 1997.

[6] J. Aslam and M. Montague, "Models for Metasearch", Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, pp. 276-284, 2001.

[7] E. D. Diaz, A. De, V.V. Raghavan, "A Comprehensive OWA based Framework for Result Merging in Metasearch", 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, Canada, Springer, pp. 193-201, 2005.

[8] Devendra Tayal, Amita Jain, Neha Dimri, Shuchi Gupta, "MetaSurfer: a new metasearch engine based on FAHP and modified EOWA operator", International Journal of System Assurance Engineering and Management, Springer, 2014.

[9] Daya Gupta, Neha Dimri, " MetaXplorer: An Intelligent and Adaptable Metasearch engine using a novel OWA operator" ,Mtech Thesis, DTU 2015.

[10] JH Holland, Adaptation in natural and artificial systems, MIT Press, 1975.

[11] SUN Ying-cheng ,LI Qing-shan, "The Research Situation and Prospect Analysis of Meta-search Engines 2012",International Conference on Uncertainty Reasoning and Knowledge Engineering,IEEE 2012

[12] T. L. Saaty, The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.

[13] T.L. Saaty, "Relative Measurement and its Generalization in Decision Making: Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors - The Analytic Hierarchy/Network Process". Review of the Royal Spanish Academy of Sciences, Series A, Mathematics, vol. 102, 2, pp 251-318, December 2007.

[14] A. Ishizaka and P. Nemery, Multi-criteria decision analysis: methods and software, John Wiley & Sons, Ltd, Chichester, West Sussex, UK, 2013

[15] M.S Dinesh, K.ChidanandaGowda and P.Nagabhushan, "Fuzzy Hierarchical Analysis for Remotely Sensed data", Geoscience and Remote Sensing Symposium Proceedings, IEEE, vol. 2, pp. 782-784, 1998.

[16] M. Gordon."Probabilistic and Genetic Algorithms in Document Retrieval" Commun. ACM, 31(10):1208–1218, 1988.

[17] N.Srinivas, K. Deb, Multiobjective optimization using nondominated sorting in genetic algorithms. Evolutionary Computation, 2(3), 221-248.

[18] K.Waqas,R.Baig, S. Ali,"Feature subset selection using multiobjective genetic algorithms," Multitopic Conference, 2009. INMIC 2009. IEEE 13th International , vol., no., pp.1,6, 14-15 Dec. 2009.

[19] Jinho Kim, Zong Woo Geem, "Optimal scheduling for maintenance period of generating units using a hybrid scatter-genetic algorithm," Generation, Transmission & Distribution, IET , vol.9, no.1, pp.22,30, 1 8 2015.

[20] K. Biswas,V. Muthukkumarasamy,K. Singh, "An Encryption Scheme Using Chaotic Map and Genetic Operations for Wireless Sensor Networks," Sensors Journal, IEEE , vol.15, no.5, pp.2801,2809, May 2015.

[21] Shih-Chia Huang, Ming-Kai Jiau,Chih-Hsiang Lin, "A Genetic-Algorithm-Based Approach to Solve Carpool Service Problems in Cloud Computing," Intelligent Transportation Systems, IEEE Transactions on, vol.16, no.1, pp.352,364, Feb. 2015.