Web Content Mining Using Genetic Algorithm

Faustina Johnson and Santosh Kumar

Department of Computer Science & Engineering
Krishna Institute of Engineering & Technology
Ghaziabad-201206, India
johnson.faustina@yahoo.in, santosh_k25@rediffmail.com

Abstract. Web mining is the application of data mining techniques on the web data to solve the problem of extracting useful information. As the information in the internet increases, the search engines lack the efficiency of providing relevant and required information. This paper proposes an approach for web content mining using Genetic Algorithm. Genetic Algorithm is being used for wide range of optimization problems. Evolutionary computing methods help in developing web mining tools which extract relevant and required information. It has been shown experimentally that the proposed approach is able to select good quality web pages as compared to the other existing algorithms proposed in the literature. The proposed approach considers several parameters like time website existed, backward link, forwards links and others for selecting good quality web pages.

Keywords: Genetic Algorithm, Web Mining, Backward Links, Forward Links.

1 Introduction

The World Wide Web is enormous and growing exponentially day by day. Finding the relevant and required information is tedious task. Information is so vast that it cannot be directly used for business purposes. Web mining is an approach in which data mining techniques are applied on the web data. Web mining approaches can be used in extracting the relevant information from the huge internet database. The data on the web is heterogeneous varying from structured to almost unstructured data like images, audios, and videos. There is enormous amount of redundant information available on the web resulting in multiple pages containing almost same or similar information differing in words and/or formats. A significant amount of information on the web is linked. Hyperlinks exist among web pages within a site and across different sites. Based on the nature of data in the web mining is categorized into three main areas: Web Content Mining, Web Structure Mining and Web Usage Mining. Web content mining search automatically and retrieves information from a huge collection of websites and online database using search engine. The data for content mining lies in various formats text, image, audio, video metadata and hyperlinks. Web Structure Mining is discovering the model underlying link structures (topology) on the web e.g. discovering authorities and hubs. Web Usage Mining mines the log file and data associated with a particular website to discover and analyze the user access patterns.

S. Unnikrishnan, S. Surve, and D. Bhoir (Eds.): ICAC3 2013, CCIS 361, pp. 82–93, 2013.

The data used for web content mining includes both text and graphical data. Based on the searching content mining is divided into two types. These are Web Page Content Mining and Search Result Mining [22, 23]. Web page content mining is the technique of searching the web via content. Search result content mining further searches the pages from a previous search. Evolutionary approaches have been used for web content mining [27]. The proposed approach uses different technique using Genetic Algorithm (GA) for web content mining. GA is a branch of Artificial Intelligence which was inspired by Darwin's theory of living organisms in which successful organisms were produced as a result of evolution [8, 13]. So GA is search algorithm based on the natural selection and natural genetics. The main significance of GA is the survival of the fittest which is also known as natural selection. It is different from other search methods in that it searches among population of points and works with coding of parameter set rather than parameter values themselves. There are problems which we cannot determine a priority to the sequence of steps leading to a solution. Search is a best method for such problems. There are two methods to perform search. These are blind strategies and heuristic strategies. Blind strategies do not use information about the problem domain. Heuristic strategies use additional information to guide the search. Two main issues in search strategies are exploring the search space and exploiting the best solution. Exploration is the method of searching new sources and exploitation is the method of using known sources. Hill Climbing is an example of exploitation and random search is an example for exploration. GA makes a remarkable balance between exploration and exploitation of the search space. The major steps in GA include generation of a population of solution, finding the objective function and the application of genetic operators such as reproduction, crossover and mutation. GA implementation starts with a population of chromosomes which are randomly generated. According to the fitness function the chromosomes are evaluated. The chromosomes with better solution are given more chance to reproduce than the chromosomes with poorer solution [13]. This paper proposes a GA based approach for web content mining to get the Top-T web links and the proposed algorithm has been compared with the algorithm proposed in [27] hereafter known as MA algorithm. It has been shown experimentally that the proposed approach performs better than the MA algorithm.

In Section 1.1 Literature review is done. Section 2 discusses GA based approach. Section 2.2 describes the proposed algorithm. An example based on the proposed approach is given in Section 3. Section 4 & 5 are experimentation and conclusion respectively.

1.1 Related Work

Web content mining is the most challenging area in the field of web mining. A lot of work has been done still the search engines lack in their efficiency and accuracy in responding the user queries. Evolutionary approaches can play a critical role in the mining of web content data. In [27], the authors proposed an algorithm (MA algorithm) for content mining. They have considered web search as a general problem of function optimization. Using the fact that the web is a graph in which nodes are web pages and edges are the links between these web pages. The search space in the optimization problem is a set of web pages. Evaluation or fitness function is done on a set

of web pages. In the beginning individuals are generated with a heuristic creation operator by querying standard search engine to obtain pages. Individuals are selected or deleted based on fitness function. And then it gives birth to offsprings after crossover is performed. Evaluation function is based on the link quality and page quality. They are obtained using the number of keywords given by the user and mean number of occurrences in link. In [2] a genetic relation algorithm (GRA) was performed for additional searching of documents according to user interest. Evolutionary GRA optimizes the relationship between hyperlinks in web pages. GRA provided the search strategy with minimal user intervention.

Genetic Algorithm is an adaptive heuristic search algorithm which was inspired by Darwin's survival of the fittest. It is based on the evolutionary ideas of natural selection and genetics. GA can be used to solve the optimization problem. GA mimics the process of the nature such as Selection, Crossover, Mutation and Accepting. The most commonly used methods for selecting chromosomes for crossover are Roulette Wheel selection, Boltzmann selection, Tournament selection, Rank selection, Steady state selection. One Point, Two Point, Uniform, Arithmetic and Heuristic crossovers can be applied on selected chromosomes [8]. For data mining optimization in educational web based system, GA utilizes the data from educational web based system and predicts the final grade and classifies them according to their grades. So here GA is useful to optimize prediction accuracy [12]. GA can be applied to different types of mining such as Content Mining, Structure Mining and Usage Mining. There are several research trends and techniques in the field of web content mining [4, 6].

Web Content Mining can be done on structured and semi-structured data. It explains how web content mining helps in extraction of data in a simple way. Web consists of data such as audio, text, video etc. HTML document is semi-structured data and data in the form of tables is structured data. Techniques used for extraction of structured data are Web Crawler, Wrapper Generation and Page Content Mining. Techniques used for semi-structured data are top-down extraction using OEM (Object Exchange Model), WICCAP [5]. Web Mining helps in resource discovery, information selection and preprocessing, generalization, analysis and visualization. The technique of discovering usage pattern from web data is known as web usage mining. It consists of three phases such as pre-processing, pattern discovery and pattern analysis [3, 7]. Pre processing consist of usage pre-processing, content pre-processing, structure pre-processing. Pattern discovery consists of statistical analysis, association rules, clustering, classification, sequential patterns, dependency modeling. According to the user's interest additional searching is done using a genetic relation algorithm. The main feature of this algorithm is that it can optimize the relationships between the web hyperlinks using evolution theory. It satisfies the user by giving more quality pages than search engines. It uses only minimum user interaction and it provides similar hyperlinks according to the users need [2]. The technique of finding data objects which differ significantly from the rest of the data are known as outlier mining. This helps in identification of competitors which in turn help in the development of electronic commerce. WCOND - Mine using n-grams without a domain dictionary algorithm is proposed for mining web content outliers. N-grams means n contiguous character slice of a string which is divided into smaller substrings each of size n .The result show that this algorithm is capable of detecting web content outliers from web datasets [9].

Information retrieval, Information extraction and machine learning were some of the techniques used to extract knowledge from the web. In [3] these techniques were compared with web mining. In [24] useful information was selected by information retrieval by indexing text. Relevant document is selected by Information retrieval whereas information extraction focuses on extracting relevant facts. Information retrieval system and information extraction system is a part of web mining. Preprocessing phase is supported by information extraction before web mining. It also helps in indexing which further helps in retrieval. Machine learning indirectly supports web mining by improving text classification process better than traditional information retrieval process [25]. Information in the web is structured to facilitate effective web mining. Web mining is decomposed into resource discovery, information extraction, and generalization [10]. A multi object combination optimization model is constructed which handles GA for negotiation problem then considers global convergence, concurrency features. A new GA based approach is proposed to hybridize local meta-heuristic mechanism which helps to speed up the search process [11]. Web design patterns are useful tool for web data mining. Web pages are analyzed to find out which useful information is included in the web page. To find out useful information two methods were used. One is information retrieval and the second one is information extraction. Information retrieval is used to extract useful information from large collection of web pages. Information extraction is used to find structure information. It uses patterns for web page description and this description is used for another task. It detects pattern instances in web pages and then it is compared with information extraction [22]. After extracting information, an algorithm is used to match large number of schema in databases. In [1] an algorithm is explained which matches correlated attributes with less cost. The positive and negative correlated attributes are distinguished by Jaccard measure.

2 The Approach

2.1 Genetic Algorithm

John Holland invented Genetic Algorithms in 1960 and at the University of Michigan. The algorithm got developed by Holland and his students in the 1960 and 1970. GA is an evolutionary algorithm. Evolution is a method of searching a solution from an enormous number of possibilities. In Biology, enormous set of possibility means it is a set of possible genetic sequences. It is a solution of finding out fittest organisms. Evolution can also be explained as a method for designing innovative solution for complex problems [25]. GA was inspired by Darwin's theory of evolution. GA is used to solve optimization problems such as numerical optimization and combinatorial optimization [8]. Optimization is the problem of finding best solution. Best Solution implies that there is more than one solution and the solutions are of not equal value [26]. A GA mimics the process of natural evolution. The five phases are Initial Population, Fitness Function, Selection, Crossover and Mutation. The most beneficial part of GA is the crossover.

Biological background of GA is that all living beings are made up of cells. And cells in turn are made up of chromosomes and each cell contains the same set of one or more chromosomes. They are made up of strings of DNA which acts as a blueprint for the organism. Chromosomes are further divided into genes. The genes are responsible for exhibiting particular trait such as eye color. Alleles are different possible setting of a trait. Brown, blue, hazel colors of eye are examples for alleles. The position of each gene in the chromosome is known as locus. Complete set of Genetic material is called organism's genome. Genotype is the particular set of genes in a genome. The physical expression of genotype is called the phenotype.

The major steps in GA are the generation of initial population of solutions, finding the objective function, and applying genetic operators [13]. These are shown in Fig. 1.

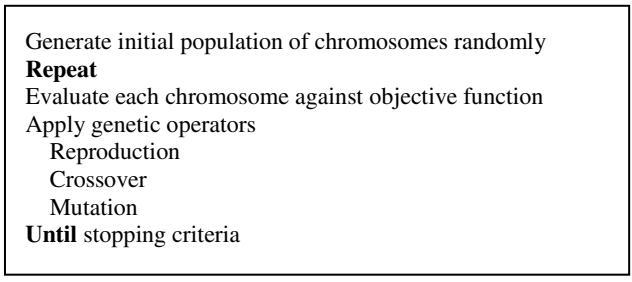

Fig. 1. Basic Genetic Algorithm

Crossover is the main operator in GA which results in generation of new offsprings. Crossover exchanges genes between parents and produces new offsprings. Mutation is another tool in GA. It is a process of changing the gene in a chromosome which results in generation of a new offspring. The offspring which survive the most are considered to be more fit. So fitness value is calculated on the basis of survival of the fittest [8, 25].

Here in the web content mining problem, the Chromosomes are a set of web pages. And each web page is the gene of the chromosome and locus is the position of web page connected to a Chromosome. GA is used to find the Top-T web links according to the need of the user.

Chromosome1 = (l_{11}, l_{12}, l_{13}, l_{14}, l_{15}, l_{16}, l_{17}, l_{18}, l_{19}, l_{20})

Where $l_{11,}$ l_{12}…. are web links.

2.1.1 Chromosome Representation

Initial population is the set of candidate solutions. Each candidate solution is represented by a chromosome. The structure of a chromosome is set of Top-T web links as given below.

26	25	52	51	11	40	36	34	30	61

Fig. 2. Top-10 Web Links

2.1.2 Fitness Function

Fitness function is a quality measurement derived from a gene. Fitness function quantifies the optimality of a solution (chromosome), so that the particular solution may be ranked against all other solution. Function depicts the closeness of a given solution to the desired result [8].

1. No of Keywords such as K_1, $K_{2,.....}$ K_n,
   Frequency of each keyword in a document d is f_1, $f_{2,}$,…. f_n respectively.
2. The time the website existed in the internet
3. Number of backward links
4. Number of forward links

$Cost_{Keywords} = \sum_{j=1}^{Top\ T} \sum_{i=1}^{n} ft$

Where n=10 and f_i is the total frequency of keywords in documents in Chromo some Cj.

$Cost_{time=} \sum_{i=1}^{n} Tl$

Where n =10 andT_l is the time the web page existed in the net.

$Cost_{backward\ link =} \sum_{m=1}^{n} Bm$

Where n=10 and B_m is the number of backward links.

$Cost_{\ forward\ link} = \sum_{p=1}^{n} Fp$

Where n=10 and F_p is the number of forward link.

Cost Function

$$F(x) = C1. \sum_{j=1}^{Top\ T} \sum_{i=1}^{n} ft + C2. \sum_{i=1}^{n} Tl + C3. \sum_{m=1}^{n} Bm + C4. \sum_{p=1}^{n} Fp$$

Where C1, C2, C3 and C4 are Constants to adjust the different parameter.

Fig. 3. Parameters for the cost function

2.1.3 Selection

The Selection is a method of choosing chromosomes for performing crossover. The proposed algorithm is using binary tournament selection. Two individuals are randomly selected from the initial population. One out of these two randomly selected chromosomes is selected for crossover. The chances of selection of fitter individual is more because of the selection of predefined parameter k as 0.75.The process is described in Fig.4.

Choose a parameter ‘k’ (say= 0.75)
Choose two individuals randomly from population
Choose a random number ‘r’ between 0 and 1
if r < k **then**
Select the fitter among the two individuals
else
Select the less fitter individual
end if

Fig. 4. Algorithm for Tournament Selection

2.1.4 Crossover

Crossover is the genetic operator which combines two chromosomes to produce new offsprings. Crossover is used to exchange the genetic material of two chromosomes. Crossover is used to explore the search space. There are several types of crossover like One Point, Two Point, Uniform, Arithmetic, Heuristic and cyclic cross over. Here cyclic crossover has been used.

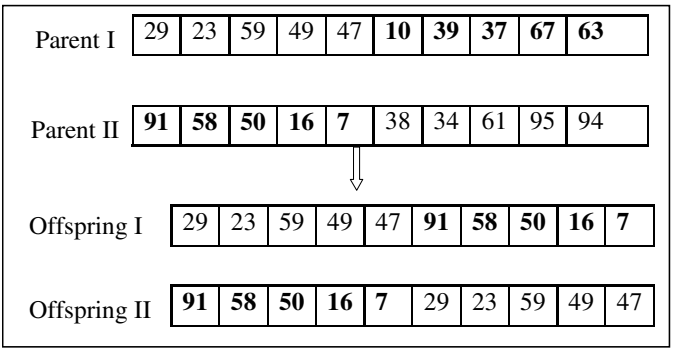

Fig. 5. Crossover

So we get two new chromosomes having different web links from their parents [8].

2.1.5 Mutation

Mutation maintains genetic diversity from one generation to the next. Mutation alters one or more gene value in a chromosome from its initial state. This results in entirely new gene. As a result GA arrives into a better solution. Mutation is used to exploit the search space. Different types of Mutation operators are Flip gene, Boundary, Uniform, Non uniform and Gaussian [8].

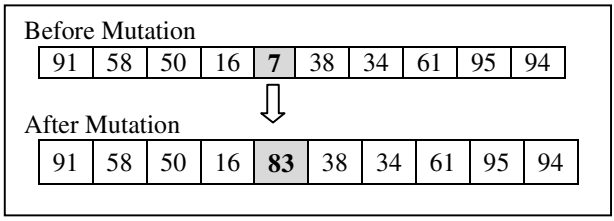

Fig. 6. Mutation

An example based on flip gene mutation is given in Fig. 6.

2.2 Proposed Algorithm

The proposed algorithm takes number of Top-T web links, Initial population size, number of generations and mutation rate as input parameter and generates Top-T web links as an output. The algorithm starts by generating initial population randomly. The

binary tournament selection is used for selecting two chromosomes from the population for generating population for crossover. The algorithm proceeds by performing crossover followed by mutation to generate new population. The process is repeated for pre-specified number of generations. The whole process is shown in Fig. 7.

Input:
- Number of top-T Web links, TopT
- Initial Population Size, InitialPopulationSize
- Number of generations, N
- Mutation Rate, MR

Output:
Set of Top-T Web links

Notations:
- Tournament Size, TS
- Crossover Rate, CR
- Crossover Point, CP
- Cost Function

F(x) = C1. $\sum_{j=1}^{T} \sum_{i=1}^{n} fl$ + C2. $\sum_{i=1}^{n} Tl$ + C3. $\sum_{m=1}^{n} Bm$ + C4. $\sum_{p=1}^{n} Fp$

Method:
Randomly Generate Initial Population, InitialPopulation[InitialPopulationSize]
For Generation=1 to N
a) //Applying Tournament Selection
  For I=1 to CR
    Select TS number of chromosomes randomly from InitialPopulation[]
    Select a random variable, rand_var, between 0 and 1
    k=CR/100
    If rand_var < k **Then**
    CrossoverPopulation[I]= fitter(TS randomly selected chromosomes)
    Else
    CrossoverPopulation[I]=less_fitter(TS randomly selected chromosomes)
  End For
b) //Performing Crossover
  For J=1 **To** InitialPopulationSize
    Parent1=ChooseRandom(CrossoverPopulation[])
    Parent2= ChooseRandom(CrossoverPopulation [])
    (offspring1, offspring2)=Crossover(Parent1, Parent2)
  //Copy offsprings to new generation
    NewGeneration[J]=offspring1
    NewGeneration [J+1]=offspring2
    J = J +2
    End For
c) Apply mutation with mutation rate, MR
d) //Copy new generation to initial population
    InitialPopulation [] = NewGeneration[]
End For
Return InitialPopulation [0]

Fig. 7. Proposed Algorithm (PA) for Selecting Top-T Web Links

3 An Example

The proposed algorithm has been implemented using JDK1.7. The program was run for 50 generations with different crossover rates. Fig. 8 shows quality of PA for generation 1, 2, 49 and 50 for crossover rate of 75% obtained from the execution. The initial population with their quality values is shown against each chromosome. Ci represents chromosomes. Each Chromosome consists of 5 pages. The initial population consisting 10 chromosomes are represented in the tabular form. The initial population as in Generation 1 has different quality values and as the generation increases the quality values get converged.

Generation 1

	Chromosome					Quality
C1	3	2	1	9	8	1539.56
C2	3	2	1	5	4	1666.82
C3	1	5	4	9	8	1534.38
C4	1	7	6	4	9	1812.96
C5	2	7	6	5	4	1782.52
C6	3	2	7	6	9	1704.07
C7	3	1	7	4	9	1680.09
C8	2	7	6	5	8	1708.62
C9	3	1	6	5	8	1717.14
C10	3	2	6	4	9	1653.88

Generation 2

	Chromosome					Quality
C1	2	7	6	5	8	1708.62
C2	1	7	6	4	9	1812.96
C3	3	2	7	6	9	1704.07
C4	3	2	7	6	9	1704.07
C5	3	2	7	6	9	1704.07
C6	3	2	7	6	9	1704.07
C7	3	2	7	6	9	1704.07
C8	3	2	1	9	8	1539.56
C9	3	2	7	6	9	1704.07
C10	3	2	1	9	8	1539.56

Generation 49

	Chromosome					Quality
C1	3	2	7	6	9	1704.07
C2	3	2	7	6	9	1704.07
C3	3	2	7	6	9	1704.07
C4	3	2	7	6	9	1704.07
C5	3	2	7	6	9	1704.07
C6	3	2	7	6	9	1704.07
C7	3	2	7	6	9	1704.07
C8	3	2	7	6	9	1704.07
C9	3	2	7	6	9	1704.07
C10	3	2	7	6	9	1704.07

Generation 50

	Chromosome					Quality
C1	3	2	7	6	9	1704.07
C2	3	2	7	6	9	1704.07
C3	3	2	7	6	9	1704.07
C4	3	2	7	6	9	1704.07
C5	3	2	7	6	9	1704.07
C6	3	2	7	6	9	1704.07
C7	3	2	7	6	9	1704.07
C8	3	2	7	6	9	1704.07
C9	3	2	7	6	9	1704.07
C10	3	2	7	6	9	1704.07

Fig. 8. Computation of Quality for Generation 1,2,49,50

The cost function of the proposed Genetic Algorithm is:
F(x) = C1.CostKeywords+C2.Costtime+C3.Cost backward link+C4.Cost forward link
Where C1, C2, C3 and C4 are Constants to normalize the different parameters according to significance.

C1=0 .09 C2= 0.05 C3=0.08 C4=0.07

In the above example, CostKeywords=2453.0, Costtime=25834.0, Cost backward link=1198.0, Cost forward link=1368.0

Using the Cost Function

$$F(x) = 0.09*2453.0 + 0.05*25834.0 + 0.08*1198.0 + 0.07*1368.0 = 1704.07$$

4 Experimentation

Fig. 9 shows the comparison of the existing GA based approach with the proposed GA based approach for selecting Top-T web pages. Both the algorithms were implemented using JDK 1.7 in Windows 7 environment. The two algorithms were compared by conducting experiments on an Intel based 2 GHz PC having 3 GB RAM. The comparisons were carried out on Quality of web pages selected by the two algorithms. The experiments were performed for selecting the top-5 to top-10 web-pages over 500 generations using different crossover rate. The graphs are plotted with Top-T pages on the X-axis and Quality of the web pages on the Y-axis for different crossover rate which

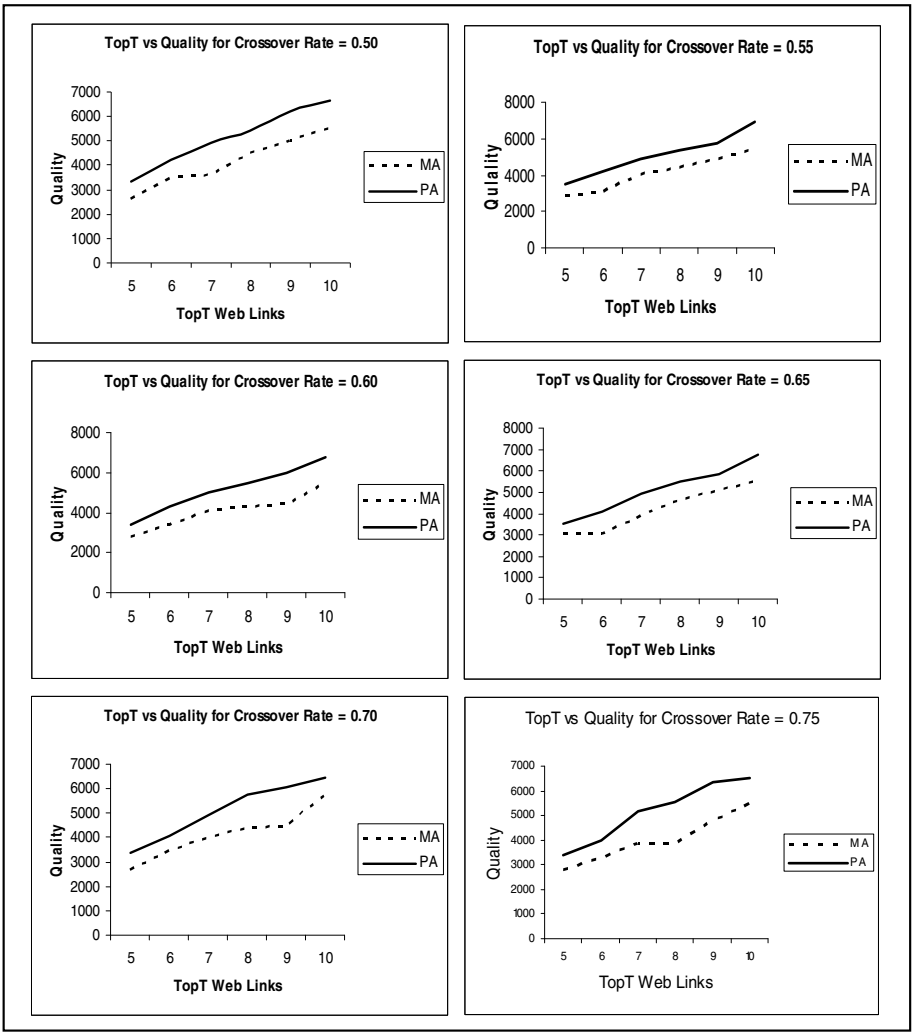

Fig. 9. Comparison of PA vs MA for crossover rate=0.50, 0.55, 0.60, 0.65, 0.70, 0.75

ranges from 0.50 to 0.75. The quality of the existing GA based approach was calculated based on the frequency of the keywords and their mean quality on the other hand the quality of PA was calculated based on frequency of keywords, time the website existed on the internet, number of backward links and forward links. Based on the significance of the factors, multiplicative constants have been used. The experimental results show that proposed algorithm performs better than the existing algorithm.

5 Conclusions

The proposed GA based algorithm is a new approach to select Top-T web links considering several important parameters like number of forward links, number of backward links, keywords and the time website existed. It helps to get relevant and required web pages. As the factors in the cost function increases, the GA provides better quality web links. Further, it has been shown experimentally that the web pages selected by the proposed algorithm are better than the existing algorithm MA. As the Top-T pages increases the quality of the proposed algorithm also increases.

References

1. Ajoudanian, S., Jazi, M.D.: Deep Web Content Mining. World Academy of Science, Engineering and Technology 49 (2009)
2. Gonzales, E., Mabu, S., Taboada, K., Hirasawa, K.: Web Mining using Genetic Relation Algorithm. In: SICE Annual Conference, pp. 1622–1627 (2010)
3. Kosla, R., Blockeel, H.: Web Mining Research: A Survey. SIGKDD Explorations 2, 1–15 (2000)
4. Liu, B., Chiang, K.C.: Editorial Special Issue on Web Content Mining. ACM Journal of Machine Learning Research 4, 177–210 (2004)
5. Nimgaonkar, S., Duppala, S.: A Survey on Web Content Mining and extraction of Structured and Semi structured data. IJCA Journal (2012)
6. Singh, B., Singh, H.K.: Web Data Mining Research: A Survey. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–10 (2010)
7. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000)
8. Chakraborty, R.C.: Fundamentals of Gentetic Algorithms. Artificial Intelligence (2010)
9. Agyemang, M., Barker, K., Alhajj, R.S.: WCOND-Mine: Algorithm for detecting Web Content Outliers from Web Documents. In: 10th IEEE Symposium on Computers and Communication, pp. 885–890 (2005)
10. Etzioni, O.: The World Wide Web: Quagmire or Gold Mine? Communications of the ACM 39(11), 65–68 (1996)
11. Zhi, Z., Jun, J., Fujun, Z., Qiangang, D.: A New Genetic Algorithm for Web-based Negotiation Support system. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 209–214 (2003)

12. Bidgoli, B.M., Punch, W.F.: Using Genetic Algorithms for Data Mining Optimization in an Educational Web Based System. In: Cantú-Paz, E., Foster, J.A., Deb, K., Davis, L., Roy, R., O'Reilly, U.-M., Beyer, H.-G., Kendall, G., Wilson, S.W., Harman, M., Wegener, J., Dasgupta, D., Potter, M.A., Schultz, A., Dowsland, K.A., Jonoska, N., Miller, J., Standish, R.K. (eds.) GECCO 2003. LNCS, vol. 2724, pp. 2252–2263. Springer, Heidelberg (2003)
13. Mathew, T.V.: Genetic Algorithm. pp. 1–15 (2005)
14. Khalessizadeh, S.M., Zaefarian, R., Nasseri, S.H., Ardil, E.: Genetic Mining: Using Genetic Algorithm for Topic Based on Concept Distribution. World Academy of Science, Engineering and Technology (2006)
15. Juang, C.F.: A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design. IEEE Transactions on System, Man and Cybernetics, 997–1006 (2004)
16. Dallal, A.A., Shaker, R.: Genetic Algorithm in Web Search Using Inverted Index Representation. In: 5th IEEE GCC Conference & Exhibition, pp. 1–5 (2009)
17. Nasaaroui, O., Dasgupta, D., Pavuluri, M.: S2GA: a soft structured Genetic Algorithm and its application in Web Mining. In: Fuzzy Information Processing Society. IEEE Proceedings, pp. 87–92 (2002)
18. Toth, P.: Applying Web-Mining Methods for Analysis in Virtual Learning Environment (2006)
19. Liu, B.: Web Content Mining. In: The 14th International World Wide Web Conference, Japan, May 10-14 (2005)
20. Nick, Z.Z., Themis, P.: Web Search using a Genetic Algorithm. IEEE Internet Computing 5(2), 18–26 (2001)
21. Kudelka, M., Snasel, V., Lehecka, O., Qawasmeh, E.E.: Web Content Mining Using Web Design Patterns (2008)
22. Dunham, M.H.: Data Mining Introductory and Advanced Topics. Pearson Education, India (2006)
23. Van, C.J.: Information Retrieval. Butterworths (1979)
24. MitChell, T.: Machine Learning, ch. 1-9. McGraw Hill (1997)
25. Mitchell, M.: An Introduction to Genetic Algorithms, ch. 1-6. MIT Press, pp. 1–203 (1998)
26. Haupt, R. L.: Practical Genetic Algorithms, ch. 1-7. John Wiley & Sons Inc., pp. 1-251 (2004)
27. Marghny, M.H., Ali, A.F.: Web Mining Based on Genetic Algorithm. In: AIML Conference, pp. 82–87 (2005)