IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Comparison of Meta-heuristic Algorithms for Web Link Prioritization

Kamika Chaudhary[1*], Neena Gupta[2]

[1,2]Department of Computer Science, Kanya Gurukul Campus(Gurukul Kangri Vishwavidyalaya) Dehradun, India

[*]*Corresponding Author: Kamika.agrohi@gmail.com, Tel.: 9557575347*

DOI: https://doi.org/10.26438/ijcse/v7i6.319324 | Available online at: www.ijcseonline.org

Accepted: 12/Jun/2019, Published: 30/Jun/2019

Abstract— Technological advancement in all the fields leads to the problem of information overload in front of internet user. It has become extremely difficult for them to reach to the information which is most relevant to their need. Information gigantic size makes the user to wander from one web page to another in order to reach to their target information. This leads to wastage of user time and also reduces the interest of user from the search engine, websites as well as internet. The problem of accessing user relevant web pages falls into the category of NP-Complete problems. Web Mining, an application of data mining is utilized to find the solution of this issue of information extraction. For retrieving the relevant data top T web links needs to be prioritized. Here we propose a memetic algorithm and simulated annealing algorithm for selecting the most relevant web document. Both the algorithms are compared on the basis of their performance experimentally and results shows the domination of one over another.

Keywords—Web Mining, NP-complete, Memetic algorithm, Simulated Annealing algorithm

I. INTRODUCTION

As the technology advances more and more information are getting accumulated in the form of web sites in the internet. As the size of the internet is increasing, the search engine fails to provide good results to the users. They fail to index all the information available on the web. The search engine becomes helpless to display the un-indexed pages which are relevant. With high usability of internet the amount of information present on the web has taken a galactic structure. Web allows us to receive or disseminate information as and when we require and at any place. According to various surveys it has found that with so much of information present on the web we are drowning in the blind alley. It has become difficult for the web user to extract relevant information. Web mining came as a rescue to this problem. It acts as technique to provide the user desired information by making use of several techniques. It helps in fighting with the trouble of information overload. It has emerged as an interdisciplinary research area as it is at the intersection level of several established research fields such as database, information retrieval, machine learning, artificial intelligence and data mining.

Web mining is categorized into three areas of interest: Web Content Mining, Web Structure Mining and Web Usage Mining [1].Web Content mining deals with discovery of useful information from unstructured, semi structured or structured contents of web documents. Text, images, audio, video comprised by unstructured document, semi structured data includes HTML documents and lists and tables represent structured documents. Web Structure Mining mines the information by utilizing the link structure of the web documents. It works on inter document level and discovers hyperlink structure. It helps in describing the similarities and relationships between sites. Web structure mining categorizes the pages into authorities and hubs. Authority pages are considered as high quality pages which are related to particular query and hub pages provide pointers to authority pages. Web structure mining is further divided into two categories hyperlinks and documents. Web Usage Mining is a data mining technique that mines the information by analysing the log files that contains the user access patterns. Web Usage Mining mines the secondary data which is present in log files and derived from the interactions of the users with the web. Web usage Mining techniques are applied on the data present in web server logs, browser logs, cookies, user profiles, bookmarks, mouse clicks etc.[2]

Search engines have to focus on to return relevant result for a user query and to achieve this by using web mining, a large variety of optimization approaches have been recommended. Here in this paper we present two meta- heuristic algorithm: memetic algorithm and simulated annealing for finding the top T web documents relevant to a search query based on web usage and web structure mining. The paper is organized in the following manner: A background of paper in the form of literature review is given in section 2. Then methodology is presented in section 3. Proposed memetic algorithm for web document prioritization is given in section 4. Simulated

annealing based algorithm is presented in part 5 of paper. Then section 6 deals with experimental evaluation and paper concluded with the conclusion and future scope along with references.

II. RELATED WORK

The World Wide Web contain huge amount of information. To extract information from the internet different techniques should be used. Web Mining should be decomposed to few subtasks such as resource finding, information selection and pre-processing, generalization and analysis [3]. Resource finding is to retrieve the information from the net available in different forms such as HTML, text, electronic newsletter, newsgroups etc. Pre-processing is the transformation of the retrieved information by treating the raw data with certain rules. For Generalization machine learning or data mining techniques are used. Analysis is done by human intervention. As a result web mining can be said as the process of discovering useful information.

Connection between web mining categories and related paradigm, R. Kosla and H. Blockeel presented a survey of research done in the area of web mining [4]. They have described various problems faced by users while they interacts with the web and then explains the categories of web mining. Resource finding, information selection and pre-processing, generalization and analysis are described as basic steps used for performing web mining. The relation between information retrieval and web mining is also depicted in the survey. Web mining mainly consist of three types of web agents named as user interface agents, distributed agents and mobile agents and content based and collaborative filtering are the two approaches used for developing these agents. It is also shown that mining of multimedia data is difficult than mining of text documents. But nowadays most of the web documents are equipped with multimedia data so multimedia web mining has also emerged as an important research area.

Application of web usage mining for online learning, in [5] author focused on developing web usage based tool which is useful in providing distance learning education to those students who cannot go to educational institution on daily basis. A web usage suggestion system is proposed which will provide the suggestion to students who are studying online on the basis of their navigational behaviour and content of web. The system will suggest that part of content which might be of interest to users but still has not focused by him. Collaboration filtering and content based filtering are the two approaches used for making suggestions. The architecture of the proposed system consists of six components: student assistant agent, student identification component, suggestion generation component, student behaviour component, suggestion delivery component and data warehouse component.

A system based on deep web content mining, as large number of online databases has been created, web has also turned into deep web. Deep web is formed by integrating several online databases. It becomes difficult to extract information from deep web as online databases provides query based access. This problem was solved in which a system is proposed for extracting and matching the information. Correlation mining techniques were used for matching the attributes retrieved from query interface and then extract the information by using Web Content Mining. Jaccard measures were used for finding synonyms and grouping terms accurately [6].

Page relevance ranking based on page content exploration, another area where web content mining has been explored is in the designing of a Page content algorithm. This algorithm is advancement over Page rank algorithm which is unable to determine the proper criteria on which ranking is provided to pages. So the proposed algorithm included a no. of important heuristics like occurrence frequency, distance of key terms and incidence of pages etc. for determining the importance of pages .Neural network is used as inner classification structure and java is chosen as programming language. It is desired that Page Content Ranking (PCR) should be present at the site of web machine chosen. The experiment performed proves that PCR is better than Page rank as it explains a topic more clearly [7].

Signed approach to produce best result of query, authors gives a signed algorithm in [8] and explains the importance of outliers in providing the relevant information as required by the user. The search result produced by search engine need not to be best because they include only the relevant document whereas much important information may also be present in the outliers. The proposed approach uses both relevant and irrelevant document along with organized domain dictionary representation. The web document is pre-processed and full word profile is generated. Word is searched if found then positive count is incremented else negative count. The result of algorithm shows that it takes less run time and proves the efficiency of the algorithm.

Website information filter system, nowadays business transactions are possible over internet because of presence of web usage mining. It analyses the usage pattern of the web user in order to perform mining. Content type, usage type or structure type data can be used for performing usage mining. There are three main phases present in web usage mining: pre-processing of data, discovery of patterns and analyses of patterns. WebSIFT system is devised in order to perform web usage mining. This system makes use of server log in

order to perform mining. It covert sessions into episodes by performing content and structure preprocessing. Knowledge discovery algorithms such as OLAP are applied and information filters automatically filters the result of these algorithms [9].

Ontology based query refinement, authors have found that answer of the search query produced by search engine need not to be the one that is needed by user. There may be much more relevant result present but is not produced because of large information base. To deal with this issue several techniques have been used. Mining is performed by creating a knowledge base of the domain. This can be achieved by define the ontology of the information. Defining ontology of knowledge base comes across several problems such as it is difficult to identify vocabulary for describing relevant concepts as well as it is also not easy to find the relation between vocabulary and definition. Another technique is to perform the refinement of user queries and extracting the interesting results which were used for performing personalization of web search. This method makes use of interactive query refining techniques. Another solution of above mentioned problem is given by using graph based algorithm for developing web crawler. Focused web crawlers have been developed which are based on graphical structure. Graph has been drawn by finding the association between web pages and their keywords. Information in Web Content mining can also be filtered by using captions but searching of web caption has not been easy as their presence was not clearly defined [10].

Advancements in web usage mining, all the recent developments in the field of Web Usage Mining have been described in the form of a survey in [11]. For performing usage mining there are various data sources present. Web servers are one of the data sources that store all the browsing behaviour of user in the form of CLF format. Usage data can also be collected at proxy side which provides advancement in navigation speed through caching. Java scripts, java applets and modified browsers can also be used for tracking usage data. Main techniques which have been used for performing the usage mining are association rules, sequential patterns and clustering. Association rules have been used for finding the web pages that are associated with other web pages. Sequential patterns have been used for finding those parts of navigation data which are following certain specific sequences. PSP+, FreeSpan and PrefixSpan are techniques used for finding sequential patterns. Clustering focuses on gather together those contents which are similar in nature. Main idea behind clustering is distance function. Multimodal clustering technique forms the clusters by using multiple parameters. Various applications of Web Usage Mining is also given in the survey such as web content personalization, pre fetching and caching for improving the server performance, e-commerce to bring together customer and vendor at the same platform.

Heuristic approach to reduce information overload and computation time, the relevance of web Usage mining and its relationship with web structure mining has been described in [12]. The knowledge obtained after performing web usage mining process is used for structuring the web document. Web pages have been divided into three main categories: Eminent web pages which are those with highest number of hit counts, average web page includes average number of hit counts and web pages with lest number of hit counts is classified as delicate pages. The pages present with more number of hit counts present higher relevance so eminent pages are placed near the home page and then average page and at last delicate pages. So Web Usage Mining information is used for structuring the web site.

Automatic term selection and web usage distinction, in [13]web usage and content mining have been integrated for recommendation of hyperlinks. The hyperlinks recommendation model utilizes the navigational information and content of the web site. This model helps in quick and easy access of the web. Overall method has been based on two independent steps. In first step extraction of text takes place whereas in next step navigational pattern is discovered. A new process has been introduced which automatically selects the terms used in clustering. Introduction of time factor will remove the old session and makes use of new session. All the session vector coordinates that are present in at least n user session have been set to 0. Experimental evaluation promises a better performance in contrast to existing solutions.

III. METHODOLOGY

Two meta- heuristic algorithms named as memetic algorithm and simulated annealing algorithm have been utilized to prioritize the web document links. The fitness value of randomly generated individual have been computed on the basis of following criterion: access frequency (AF), time duration (DUR) for which user stayed on a web page, unique visitors (UNQV), hubs (HUB) and authorities (AUTH).

$$\text{Fitness Value} = \text{Cost} = AF_i + DUR_i + UNQV_i + HUB_i + AUTH_i$$

IV. MEMETIC ALGORITHM

According to the theory given by Richard Dawkins, Memetic Algorithm [14] is inspired by memes, which are simple units of imitations of human behavior. Basically a meme is considered as an element of knowledge which has property of duplication, modification as well as combination. When a meme is added to other meme it results in forming a new meme and in this way works towards human culture evolution. Memes lifespan depends on their interestingness

frequency. Those who are less interesting lose their importance and die in short span of time and those which are strongly interested keeps on propagating through the community. One of the important characteristic of meme which turns out to be an inspiration for memetic algorithm is their ability to evolve themselves throughout their entire lifespan. The MA is population dependent stochastic global search meta-heuristics which amalgamates the local search heuristics in evolutionary computing framework. It mainly combines the advantages of local search heuristics with the evolutionary genetic algorithm approach. In this the population of individuals represents the candidate solution and the elements of candidate solution composed to form a chromosome are called memes not genes. This algorithm is sometimes also referred as hybrid GA as it applies a local search refinement process to the population in every subsequent generation. The basic concept behind MA is to bring together both the local search and global search heuristics. This concept has been widely implemented to a vast range of NP hard combinatorial optimization problem scheduling problem, cell formation problem, travelling sales man problem etc.[15] A basic memetic algorithm is given in figure 1.

```
Input: randomly generated web doc
Output: Set of top priority web doc
Begin
Randomly initialize population
while (solution is not found or predefined no of generations is not reached) do
   Perform crossover between two parents individual and generate the children
Children=crossover(parent1(webdoc),parent2( web doc))
Apply local-search iterative improvement (children(web doc))
Add children to current population
    for (all the web doc)
          Apply mutation operator
     if (fitness improved) then
          Replace the current population with mutated version)
     end if
   end for
   for (all the web doc )
          Apply iterative improvement heuristic
   end for
Until terminate=true
end while
end
```

Figure 1. Memetic algorithm

V. SIMULATED ANNEALING

Simulated Annealing (SA) originated in 1983 is a random search methodology that has its root in the theory of statistical mechanics. It has been used unbeatably to solve a number of combinatorial optimization problems [16]. The inspiration behind SA is the process of annealing of solids which is initially used in metallurgy. During the process of annealing the substance is first heated at high temperature until it reaches its fusion point and liquefies and then is cooled down slowly until it solidifies. Simulated Annealing works by deciding to a neighbourhood solution from a current solution. The acceptance probability of moving to a neighbourhood structure depends upon the quality of neighbour and the temperature value. Initially temperature is placed at a high value then there is a possibility to take a worse move with the expectation that its neighbours will represent a better solution. Then slowly and gradually temperature is lowered so that it stabilise the search and move forward and near to optimal solution. It has been found that SA procedure exhibits a lot of similarity with iterative improvement procedure but with the difference that it also performs uphill moves comprising of some probability. The point behind accepting uphill moves is that it may be possible that some local minima lie in the proximity of each other and are separated by a small number of uphill moves. So SA performs uphill moves so that a better solution is not missed if the algorithm stops at the first local minimum visited. SA begins by randomly generating an initial state S_I and initial high temperature T_0 which is being decreased gradually. Then a neighbouring solution of initial solution is generated at a particular temperature and less efficient solution with certain acceptance probability is chosen for next move. The process continues with the reduction in temperature in every next move till the equilibrium is reached. The SA algorithm reaches to termination when the frozen state is achieved which is when the temperature t becomes less than the previously mentioned temperature [17]. The simulated annealing algorithm is given in the figure 2.

Step1: Generate an initial random solution S_I
Step2: Choose initial temperature T_0
Step3: Repeat until temp<predefined min temp
Step4: Repeat until equilibrium not attained
1) Generate S_N as neighbour solution of S_I
2) Compute ΔE= value(S_N) – value(S_I)
3) IF $\Delta E \leq 0$ then assign $S_I = S_N$ as downhill move
4) ELSE assign $S_I = S_N$ with P_{accept} $e^{-\Delta E/T}$ as uphill move

Step5: Assign T= reduce(T_0)// T_0 is pre-specified value ranges between 0 and 1
Step6: Stop when stopping criteria is met

Figure 2. Simulated annealing algorithm

VI. EXPERIMENTAL RESULTS

Both memetic and simulated annealing algorithm were implemented using JDK 1.6 in Windows 10 environment. They were compared by performing experiment on an Intel based 2.13 GHz PC having 3 GB RAM. The comparisons were carried out on fitness cost due to top document selected by two algorithms. Graphs were plotted to compare simulated annealing and memetic algorithm on fitness cost against top T web document. On the x-axis top T web documents are plotted while y-axis represents the fitness cost or quality of web documents. For simulated annealing following parameters were used: initial temperature=610, frozen temperature T<1, cooling rate = 0.75 and equilibrium point= 5. The parameters were set according to predefined standards. For memetic algorithm following set of parameters were used: population size=5, number of generations =100, crossover probability= 0.7 and mutation rate= 0.05. The quality of web documents were calculated on the basis of factors such as access frequency, time duration, unique visitors, hubs and authorities. In the case of memetic algorithm results shows that higher the value of generation number better results are obtained even if the population is small in number. All the graphs are shown in figure 3 to 6.

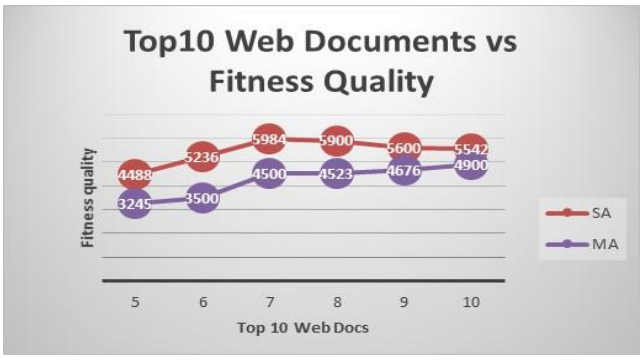

Figure 3. Comparison among SA and MA for iteration=100

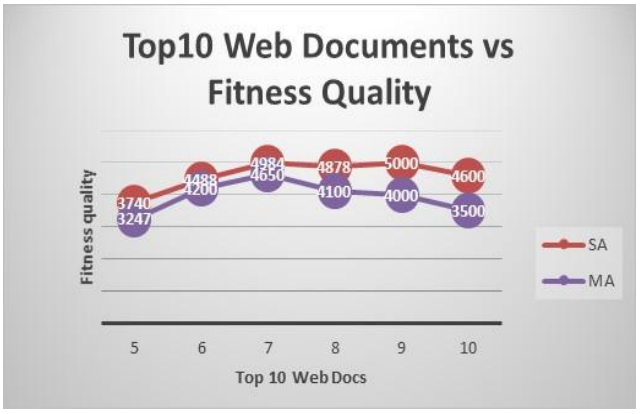

Figure 4. Comparison among SA and MA for iteration=200

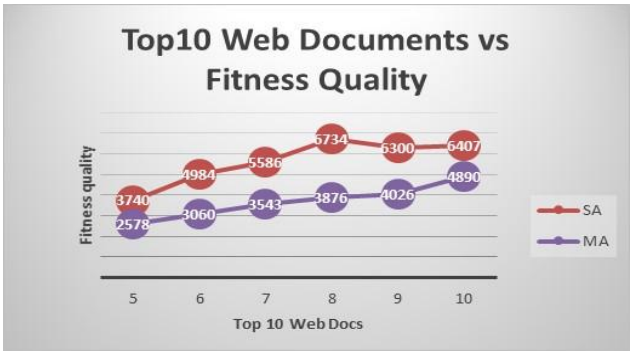

Figure 5.Comparison among SA and MA for iteration=300

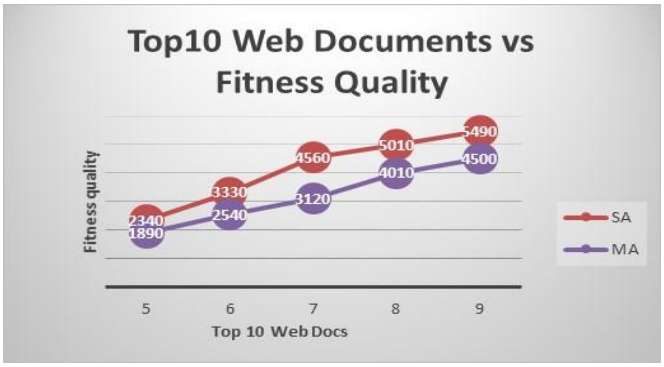

Figure 6. Comparison among SA and MA for iteration=400

Above graphs shows the comparison of both the algorithms for different number of iterations. It can be analyzed from above graphs that with the increase in number of iterations memetic algorithm fitness cost is reducing which indicates increase in quality. So memetic algorithm is considered as better in performance in comparison to simulated annealing as it provides top 10 web documents in prioritized order by utilizing lesser resources.

Conclusion and Future Scope

In this paper an overview of web mining has been stated along with the description of all the categories of web mining including content, structure and usage. It also focuses on the issues and challenges faced at the time of accessing the World Wide Web. To solve the problem of information extraction memetic algorithm as well as simulated annealing has been applied to find the top T web document according to their prioritization order. Both the algorithms were implemented experimentally and a comparison has been shown which results in the superiority of memetic algorithm.

REFERENCES

[1] S. Sharma, and M. Rai, "Customer Behavior Analysis using Web Usage Mining," International Journal of Scientific Research in Computer Science and Engineering, vol. 5, issue 6, pp. 47-50,

2017.
[2] A. Kashyap, I. Naseem, and D. Mandloi, "Web Mining an Approach to Evaluate Web", International Journal of Scientific Research in Computer Science and Engineering, vol. 5, issue 3, pp. 79-85, 2017.
[3] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," pp. 558–567, 2002.
[4] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," vol. 2, no. 1, 2000.
[5] O. R. Za, "10.1.1.21.799.Pdf."
[6] S. Ajoudanian and M. D. Jazi, "Deep Web Content Mining," vol. 3, no. 1, pp. 501–505, 2009.
[7] U. shi and M. R. Singh, "Page Content Rank: An Approach to the Web Content Mining," *Int. J. Eng. Trends Technol.*, vol. 22, no. 2, pp. 74–78, 2015.
[8] G. Poonkuzhali, K. Thiagarajan, K. Sarukesi, and G. V Uma, "Signed Approach for Mining Web Content Outliers," vol. 3, no. 8, pp. 820–824, 2009.
[9] Cooley, R., Tan, P. N., & Srivastava, J. (1999, August). Websift: the web site information filter system. In *Proceedings of the Web Usage Analysis and User Profiling Workshop* (Vol. 8).
[10] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann, "Ontology refinement for improved information retrieval," *Inf. Process. Manag.*, vol. 46, no. 4, pp. 426–435, 2010.
[11] F. M. Facca and P. L. Lanzi, "Recent Developments in Web Usage Mining Research," pp. 140–150, 2003.
[12] C. C. Lin, "Optimal Web site reorganization considering information overload and search depth," *Eur. J. Oper. Res.*, vol. 173, no. 3, pp. 839–848, 2006.
[13] R. Fuller, R. John, R. B. Eds, P. Sincak, J. Vascak, and V. Kvasnicka, *and Web Mining Advances in Soft Computing*. .
[14] F. Neri and C. Cotta, "Memetic algorithms and memetic computing optimization: A literature review," *Swarm Evol. Comput.*, vol. 2, no. February, pp. 1–14, 2012.
[15] P. Moscato and C. Cotta, "A Modern Introduction to Memetic Algorithms," no. January 2003, pp. 141–183, 2010.
[16] A. Orman, E. Aarts, and J. K. Lenstra, "Local Search in Combinatorial Optimisation.," *J. Oper. Res. Soc.*, vol. 50, no. 2, p. 191, 2006.
[17] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization of Simulated Annealing," *Ann. Phys. (N. Y).*, vol. 54, no. 2, pp. 671–680, 1969.

.

Authors Profile

Ms. Kamika Chaudhary, is pursuing Ph. D. in computer science from Gurukul Kangri Vishvavidyalaya Haridwar, Uttarakhand. She is a master in technology and is UGC National Elgibility Test qualified. She has research interest in the field of web mining, information retrieval and recommendation systems..

Dr. Neena Gupta, is working as an assistant professor in Kanya Gurukul Campus **(**Gurukul Kangri Vishvavidyalaya) Dehradun, Uttarakhand. She has more than 11 years of experience and has guided several Ph.D. research scholars. Her research interest includes distributed systems, data mining, and web database. She has published a large number of research papers in international journal and has attended many international conferences ,workshps and seminars.

