IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WEB REORANIZATION USING METAHEURISTIC OPTIMIZATION

[1]Surabhi Singh
Software Engineer

[2]Dr. Daya Shankar Pandey
Software Engineer

Abstract
Today we live in a developing era, where we all have access to the internet and depend on it to collect various information in our daily life. As we know the Internet is a very vast topic, which is a collection of various websites, and its size is increasing day by day that it becomes hard for users to differentiate between real and duplicate content present on the websites. It is a challenging task for users to search for relevant and required information from such a huge source of information present over the internet. When a person searches for any information or site the information that is on the top are those which are most viewed or top rated, but some sites are underrated but the content area is worth viewing and using. This paper aims to reorganize the website using Metaheuristics optimization in a way that the top searches will be those whose content is more relevant to the searches. Metaheuristic algorithms will be used to extract useful information from the literature. It will help in recommending sites from different web organizations. It will help in improving the user's experience on websites.

Keywords: Web Mining, Genetic Algorithm, Metaheuristics Optimization, Metaheuristic algorithms

1. Introduction

To arrange the material that is available on the internet for the clients, the web needs to be reorganized. Providing the users with experiences that are tailored to their needs, also helps in gathering information about their preferences and goals. Web rearrangement is gathering user interactions and website navigation to provide users with more precise and ordered results online. Making the website more aesthetically pleasing, offering correct search results, and simply giving users the information they need are all parts of the reorganization process. This website could be any type of general website, from an e-commerce site to one for stock trading. Web rearrangement enables businesses to give their users something special and unique based on their needs and preferences rather than delivering a unit or broader experience.

Finding relevant information from the vast amount of data kept on numerous servers across the world that each user needs to use the internet is their main challenge. The user must invest their time and effort into finding what they need. It would be quicker for users to find the needed web link if these websites could be prioritized according to user relevance. Another issue is that many people do not click through to the next page of the results displayed.

Finding the minimal and maximum points depending on some optimal solution is the process of optimization [9]. There are numerous varieties of algorithms that are used to solve optimization problems. Numerous significant applications have found a place for optimization, and various optimization techniques have been effectively used in these applications. Right now all the results we got from the searches we made in the Search engines are through two algorithms that are Paper Rank(PR) and SOE algorithm.

Metaheuristic optimization algorithms are algorithms designed to solve optimization problems with large spaces of candidate solutions quickly [9]. There is no guarantee that these algorithms will provide a definite optimal solution, and they may produce different results in each problem. However, researchers prefer these algorithms because they provide solutions quickly and require little mathematical knowledge to use. Existing metaheuristic algorithms include genetic algorithms, tabu search algorithms, ant colony algorithms, artificial bee colony algorithms, and artificial immune system algorithms, with new additions being developed on a daily basis [9]. The effectiveness of an algorithm is the measurement pertaining to the solutions that are being given by an algorithm. The efficiency of an algorithm is primarily connected to time

and space, which includes speed and the rate of convergence in the direction of the global optimum. The number of optimal solutions and the statistical analysis of the importance of the result shows the algorithm's competence primarily.

Modern metaheuristics are becoming more and more popular among the many optimization methods available today, which is leading to the emergence of a new field of optimization known as metaheuristic optimization. The majority of metaheuristic algorithms, including simulated annealing, ant colony optimization, and particle swarm optimization, are based on natural phenomena. The metaheuristic algorithm is mostly utilized in the disciplines of optimization, data mining, machine learning, design, schedule mining, and several more applications of this nature. The primary issues with metaheuristic algorithms are that they are extremely complex, non-linear, and stochastic [11].

The purpose of this paper is to provide web link ranking algorithms. The algorithms are tested on various dataset values and then evaluated based on their results. These findings will be useful in examining different applications of the algorithms and solving other problems.

1.1 Methods Used by Current Search Engines for Web Organization

As we have mentioned above most the Search Engines like Google, Bing, etc. use PR and SOE for the search result.

Method 1: Page Rank

Every search result is assigned a rank or score by PageRank. The higher the page's score, the higher it will appear in the search results list.

The number of other Web pages that link to the target page influences the score. Each link represents a vote for the target. The logic behind this is that pages with high-quality content will be linked to more frequently than pages with low-quality content.

Not all votes are created equal. Votes from high-ranking Web pages are weighted more heavily than votes from low-ranking sites. You can't really boost the rank of one Web page by creating a bunch of empty Web sites that link back to the target page [14].

The more links a website sends out, the less powerful its voting power becomes. In other words, if a high-ranking page links to hundreds of other pages, each individual vote will be worth less than if the page is only linked to a few sites. Other factors that may influence scoring include the age of the links leading to and from the site, the strength of the domain name, how and where the keywords appear on the site, and the age of the links leading to and from the site. Google gives more weight to sites that have been around for a while. Some claim that Google employs human testers to evaluate search results, manually sorting through results to handpick the best links.

Google denies this, claiming that, while it does use a network of people to test new search formulas, it does not use humans to sort and rank search results [13].

Method 2: SEO(search engine optimization)

SEO is an abbreviation for "search engine optimization." In layman's terms, it refers to the process of improving your website so that it appears more prominently when people search for products or services related to your company on Google, Bing, and other search engines [14] . The higher your pages appear in search results, the more likely you are to attract attention and attract new and existing customers to your business. Bots are used by search engines such as Google and Bing to crawl web pages, going from site to site, collecting information about those pages, and indexing them. Following that, algorithms analyze pages in the index, taking hundreds of ranking factors or signals into account, to determine the order in which pages should appear in search results for a given query [14].

2. Related work

In [1] authors demonstrate how the Genetic Algorithm is combined with various other methods and techniques to derive optimal solutions, improve retrieval system computation time, and demonstrate the applications of genetic algorithms in various fields. To obtain the best solutions, we need a metric that distinguishes the best solutions from the worst solutions. The measure could be objective, such as a statistical model or simulation, or subjective, in which we prefer better solutions over worse ones.

In [2] the authors propose a scheduling framework based on genetic algorithms for real-world scheduling problems that take into account job release times, job due dates, and different assembly levels. The foundation of this framework is a decomposition of the job-shop scheduling problem into a series of deterministic single-machine scheduling problems. A mechanism for coordination is proposed. In [3] authors' comprehensive review of the firefly algorithm is presented in this paper, and various characteristics are discussed. The various variants of FA are discussed, including binary, multiobjective, and hybrid with other meta-heuristics. The applications and metrics for performance evolution are presented. This paper discusses FA's potential future research directions.

In [4] authors describe and compare three categories of Web Mining that are Web Structure Mining, Web Usage Mining, and Web content Mining. It also provides comparative statements of various page ranking algorithms with link editing, General Utility Mining, and Topological frequency Utility Mining Model by taking constraints such as Web Mining activity, topology, Process, Weighting factor, Time complexity, and Limitations into account, among other things.

In [5] authors examine the applications of the Firefly Algorithm (FA) in a variety of optimization problems. According to the literature, FA was mostly used by researchers to solve optimization problems in the Computer Science and Engineering domains. Some are enhanced or hybridized with other techniques to improve performance. Furthermore, studies found that the majority of cases where the FA technique was used outperformed other metaheuristic algorithms.

In [6] authors explain the Genetic Algorithm and Traveling Salesman Problem. And then explain that for the traveling salesman problem, many different crossover and mutation operators have been developed, each producing a different set of results. And compare these results and discover that operators that use heuristic information or a graph matrix representation produce the best results.

In [7] authors propose that metaheuristic calculations are methods developed to successfully handle computational testing enhancement concerns. He adds that this section focuses on metaheuristic calculations shown on non-straight actual peculiarities with a significant streamlining worldview,

having demonstrated impressive investigation and double-dealing capacities for such improvement issues, and centering on a few well-known materials science-based metaheuristics as well as depicting the hidden unique actual cycles related with each calculation.

In [8] authors propose a complete analysis of Nature-Inspired MetaHeuristic Techniques for Feature Selection Problems that provide the best solution by iteratively exploring and exploiting the entire query space. One hundred and 76 articles have been combined and investigated using a precise survey process. The list of nature-driven meta-heuristic techniques and their modifications, along with datasets and execution, are also shown. Also provided is the itemized distribution pattern of metaheuristic element determination methods.

In [9] authors discussed how meta-heuristic algorithms are used to solve various optimization challenges. By modifying the input parameters, these algorithms are being tested on various test scenarios after being developed by studying diverse natural phenomena. By altering the input parameters for each optimization, the outputs of the algorithms can be observed to change. In [10] authors discuss the use of the genetic algorithm in important areas of optimization, including fuzzy, combinatorial, and multi-objective optimizations, as a clear-cut introductory optimization tool for quickly tackling a variety of optimization problems.

In [11] authors' aim is to rank the online links that might be used to restructure websites for business intelligence. Different factors have been taken into consideration when prioritizing topT web links, including keyword frequency, and user navigation behavior, such as unique visitors, duration of stays, access frequency, hub information, and authority information using metaheuristic optimization.

In [12] authors proposes a study based on a comprehensive survey of existing works and practice considerations, we propose a new formulation for the website structure optimization (WSO) problem in this study. An enhanced tabu search (ETS) algorithm with advanced search features such as multiple neighborhoods, adaptive tabu lists, dynamic tabu tenure, and multi-level aspiration criteria is proposed.

In [13] authors analyses and compare the web page ranking algorithms based on various parameters and results to determine their advantages and limitations for web page ranking We worked on Panda and Penguin, two algorithms used by the popular search engine Google to list a web page. Panda is a content filtering algorithm that searches for duplicate content on websites and penalises them by listing them last in response to similar query parameters. Along with Penguin, there is a linking-based algorithm that filters websites with the highest and highest quality links with relative information.

In [14] authors describes the tweaks of getting a page to the top of Google by increasing the Page rank, which may result in increased visibility and a profitable deal for a company. Google is the most user-friendly search engine proven for Indian users, providing user-oriented results. Furthermore, because most other search engines use Google search patterns, we focused on it. As a result, if a page is optimized for Google, it is optimized for the majority of search engines.

3. Formulation of Problem

We all use search engines in our day-to-day life, for various things like searching for a recipe, or e-commerce websites, to get some help in our work, or to get any other information. Web Browsers like chrome have become an important part of our daily life. When we searched we got a minimum of hundreds of responses and finding the required and relevant information is very time-consuming work. And even after giving our time to this, we are not able to get the most desired and relevant content regarding the search. As we all know that people are busier and busier day by day due to the fast-moving world we live in. So, it's quite impossible to provide this much time searching for something. We all want a web browser where the top recommendation is those websites that have the most desired and relevant content present in them. Not only it takes a lot of time in searching for something relevant, but those authors that are not well known or acquitted on the internet even though their content is worth reviewing are left behind this not only affect them but us also as we won't be able to get the desired content as the while searching the top recommended websites are those which are top rated and the websites which contain more number of words searched in it. So, to enhance users' experience and to make available what they want we are using metaheuristic optimization. In this, the two main algorithms that we have mainly used are the Genetic Algorithm and Firefly

Algorithm. These algorithms are based on the survival instinct of an animal.

4. Findings

There will be many aspects that will be used by us in this project of web reorganization. That will be the User Count- the number of users who visited a website, Unique User Count- Sometimes an author can create fake user's to increase their UserCount. So, it's the number of unique users who visited a website, Stay Time- for how many seconds the user stayed on that website, Keywords Frequency- how many times the keyword provided in the search box is used in a website, Hubs - How many websites refer to a particular link in their content, and Authority- How many websites refer to a particular website in their content for preference. For this we are using two metaheuristics algorithms that are Genetic Algorithm and FireFly Algorithm.

5. Cost function

Cost function depends on various parameters

Table1: Fitness function Parameters

Parameters	Description
KF	Frequency of keywords
MK	Maximum keyword in web page
MA	Maximum authority value
MH	Maximum hub value
MD	Maximum duration spent by visitor on web page
MV	Maximum visitors
MAC	Maximum access count

$$cost(webP_i) = cost_{mac} + cost_{mv} + cost_{md} + cost_{mh} + cost_{ma} + cost_{kf} \ldots\ldots$$

6. Proposed Model

We are using the Genetic Algorithm for web reorganization. In this, we will first initiate a random population. And this population will be of URLs and a unique Id will be given to them in the range of 1 to 1000 and is saved in a multidimensional array named as initialPop which will contain arrays of size five with randomly selected unique ids in each array. Then we will assign the value to different IDs according to the aspects that are mentioned in the findings above. Then will evaluate these arrays according to the maximum value. After setting a crossover percentage we will take arrays depending on the percentage from the initialPop and perform the cyclic crossover on those arrays though there are many methods for crossover one of them being single crossover after it the child population that will be created have a lot of chance of repeated URL or unique id in the same array. So, the cyclic crossover is the optimal option here. After this, all the new arrays or child populations will be stored in the newPopulation array. The next step after crossover is a mutation in which a few elements of the array will be replaced by random elements. After mutation, all the steps from the start will repeat themselves till we get a convergence point that will either get the same result again or again after repeating the process we will get a single array in the whole multidimensional array.

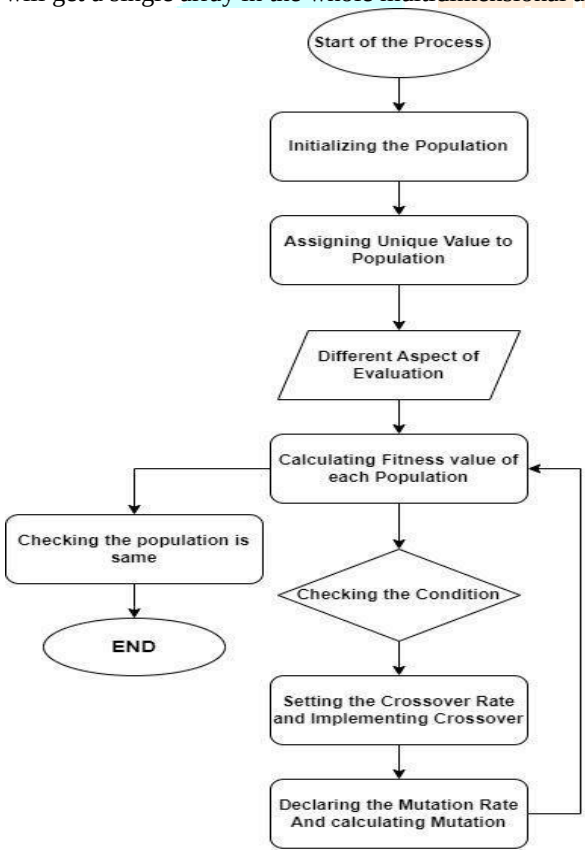

Fig 1. Flowchart of Proposed Model

7. Algorithms Used

7.1 Genetic Algorithm

Genetic algo. Is composed of two process selection and mutation. Belongs to evolutionary algorithm. It is a heuristic search algo. Genetic algo. is based on genetic and natural selection. It is used to generate high quality solution from optimization problem.

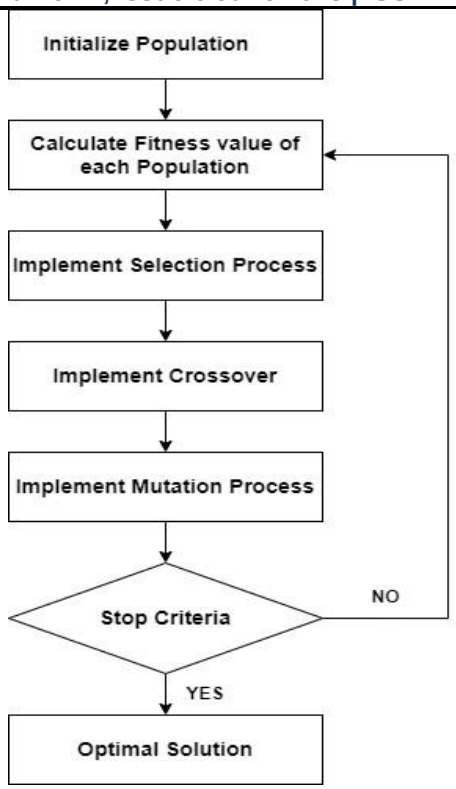

Fig 2. Flowchart of Genetic algorithm

7.2 Firefly Algorithm

Firefly Algorithm is a metaheuristic optimization algorithm which is inspired by the flashing behavior of fireflies that is used to solve big optimization problem.

In this algo, FF attract towards each other, less brighter ff is attracted towards brighter ff and as the distance between 2ff increases attractiveness decreases. Algorithm:

Let there are N initial solutions and i denotes the number of iteration, with 'n' maximum iteration number. 'I' is the intensity or brightness of flash. *While(t< maxgeneration)*

```
{
For(i=1;i<=n;i++)
{
  For (j=1;j<=I;j++) // shifting firefly i towards j
   {
     If(Ij > Ii) // vary attractiveness with distance r via exp(-ϒ,r)
   }
}
}
```

Objective function :

1. f(x), x=(x1,x2,x3,x4…..x_d)
2. generate an initial population of fireflies x_i(i=1,2,3….n).
3. Formulate light intensity I so that it is associated with f(x)
4. define absorption coefficient ϒ

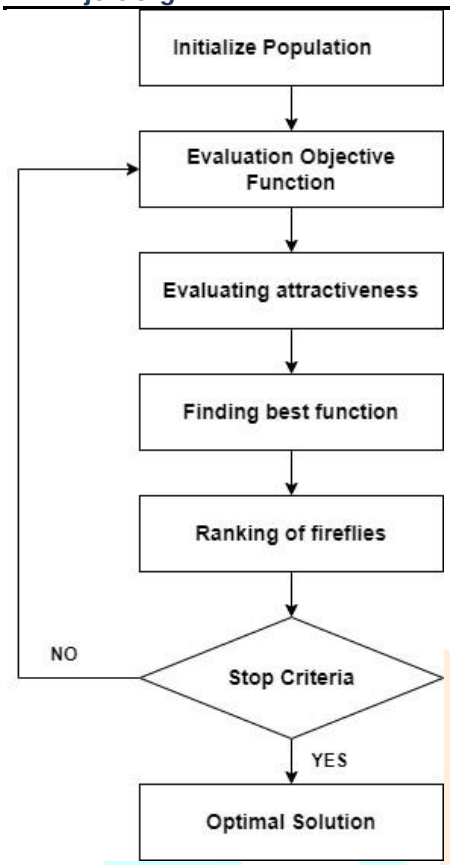

Fig3. Flowchart of firefly algorithm

8. Combined flowchart for Genetic and firefly algorithm

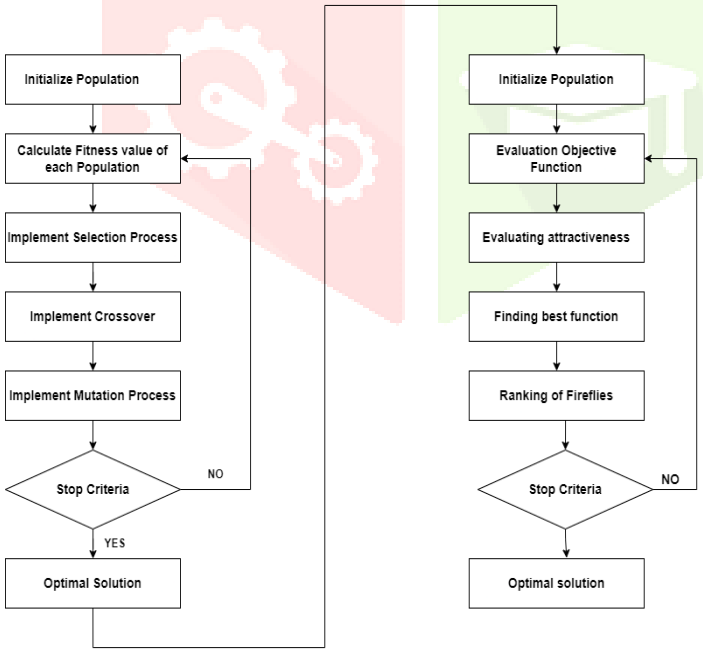

Fig 4. Hybrid flowchart for Genetic and firefly Algorithm

In this firstly we have find out the optimal solution for genetic algorithm and the final solution will be the input for firefly algorithm as initialize population and calculate the best optimal solution for firefly algorithm.

9. Result

Following are the Tables and Graphs observed during the project.

In tab1 we have calculated fitness for different crossover_rate from 60 to 90 and topT at 10, and opted for the best crossover_rate which is 70 here.

In tab2 we have calculated fitness for different mutation_rate which is 5% and 10% and topT at 10, and opted for the best mutation _rate which is 5% here.

CR	for_top10
60	2956
70	3175
80	3025
90	3147

Table 2.Crossover Rates against topT=10 fitness

MR
5
10

Table 3. Mutation Rates against topT

topT	GA(CR =60, MR0.05)	FF(CR=60, MR0.05)	GA+FF(CR=60, MR0.05)
5	1646	1665	1715
6	2020	2097	2167
7	2113	2278	2346
8	2762	2799	2805
9	2669	2790	2845
10	3185	3367	3389

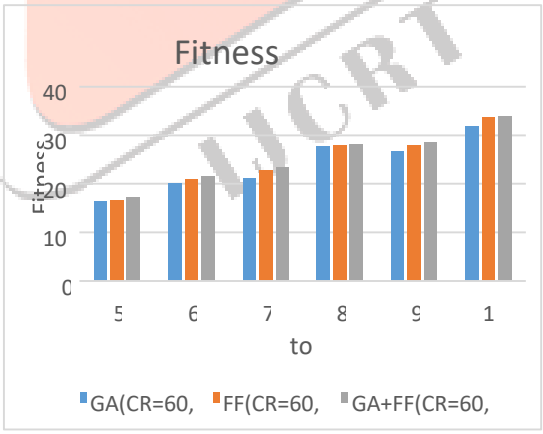

Fig: 1 Fitness (CR=60,MR=0.05)

topT	GA(CR=70, MR0.05)	FF(CR=70, MR0.05)	GA+FF(CR=70, MR0.05)
5	1629	1697	1789
6	1987	1999	2090
7	2326	2422	2578
8	2655	2705	2768
9	2776	2796	2813
10	3185	3190	3278

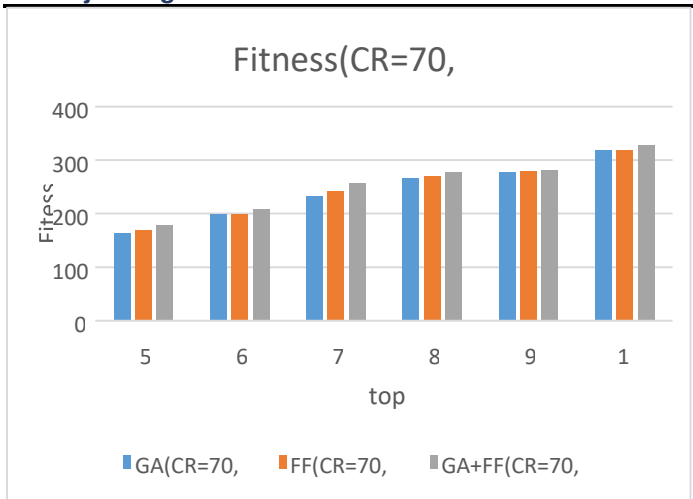

Fig: 2 Fitness(CR=70, MR=0.05)

topT	GA(CR=80, MR0.05)	FF(CR=80, MR0.05)	GA+FF(CR=80, MR0.05)
5	1558	1692	1789
6	1936	2027	2095
7	2256	2374	2456
8	2584	2599	2616
9	2628	2746	2876
10	3029	3189	3245

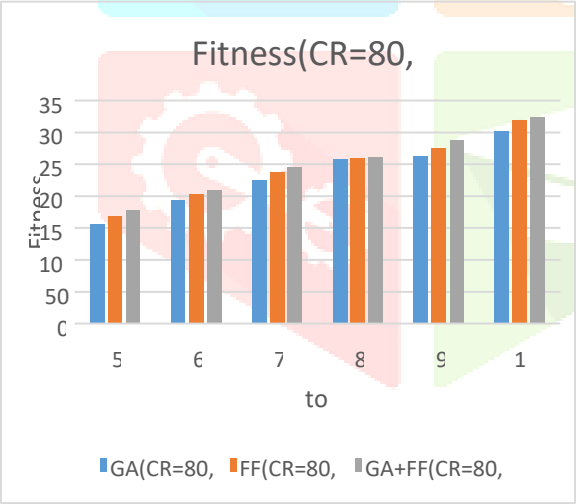

Fig: 3 Fitness(CR=80, MR=0.05)

In Fig1: We have drawn graph between genetic, firefly and genetic + firefly algorithm for
CR=60 and MR = 0.05 for different topT whose ranges between 5 to 10. In Fig2: We have drawn graph between genetic, firefly and genetic + firefly algorithm for CR=70 and MR = 0.05 for different topT whose ranges between 5 to 10. In Fig3: We have drawn graph between genetic, firefly and genetic + firefly algorithm for CR=80 and MR = 0.05 for different topT whose ranges between 5 to 10.

10. Conclusion

In this paper we have addressed the problem of adaptive websites by finding the top-T web documents with the help of hybrid approach using genetic and Firefly algorithm. First, population of top the web documents has been defined with the help of genetic algorithm and it optimises result it has been passed as input to the Firefly algorithm face further refine the set of top T web documents. Booth algorithm has been implemented in Java and is it has been experimentally shown that the hybrid approach is able to to give better quality top T web documents which can be used to reorganize the website. These top ranked web documents can be kept higher in the hierarchy that is near to home page.

11. Reference

[1] Lingaraj, Haldurai. (2016). A Study on Genetic Algorithm and its Applications. International Journal of Computer Sciences and Engineering. 4. 139-143.

[2] .[MAD02] Madureira, A., Ramos, C., & Silva, S. Carmo, (2002). A Coordination Mechanism for Real World Scheduling Problems using Genetic Algorithms. Evolutionary Computation, in Proceedings of the 2002 CEC, vol. 1, pp. 175 –180.

[3] Verma, Himanshu. (2020). A Systematic Review on Firefly Algorithm: Past, Present, and Future. 10.36227/techrxiv.12122748.

[4] Bharamagoudar, Geeta. (2012). Literature Survey on Web Mining. IOSR Journal of Computer Engineering. 5. 31-36. 10.9790/0661-0543136.

[5] Johari, Nur & Zain, Azlan & Mustaffa, Noorfa & Udin, Amirmudin. (2013). Firefly Algorithm for Optimization Problem. Applied Mechanics and Materials. 421.10.4028/www.scientific.net/AMM.421.512.

[6] [BRY00] Bryant, Kylie (2000). Genetic Algorithms and the Traveling Salesman Problem, in Proceedings of 1st GNT Regional Conference on Mathematics, Statistics and Applications.

[7] Chattopadhyay, S., Marik, A. and Pramanik, R., 2022. A Brief Overview of Physicsinspired Metaheuristic Optimization Techniques. arXiv preprint arXiv:2201.12810.

[8] Manik Sharma, Prableen Kaur. A Comprehensive Analysis of Nature-Inspired MetaHeuristic Techniques for Feature Selection Problem. Arch Computat Methods Eng 28, 1103–1127 (2021).

[9] Sariman, Guncel & Kucuksille, Ecir Ugur. (2014). Web Based Educational Tool for Metaheuristic Algorithms. Pamukkale University Journal of Engineering Sciences.

[10] 46-53. 10.5505/pajes.2014.15870. 12. Sivanandam, S.N. and Deepa, S.N., 2008. Genetic algorithm optimization problems. In Introduction to ge netic algorithms (pp. 165-209). Springer, Berlin, Heidelberg.

[11] Optimization of multi-criteria website structure based on enhanced tabu search and web usage mining Applied Mathematics and Computation15 August 2013Volume 219, Issue 24Pages 11082-11095. Peng-Yeng Yin, Yi-Ming Guo

[12] WDPMA: An MA-Based Model for Web Documents Prioritization Santosh Kumar, ABES Engineering College, Ghaziabad, India

[13] Goyal, Deepak & Goyal, Drdinesh. (2013). Comparative Analysis of Various Google Algorithms and Their Effects on Performance of a Web Portal.

[14] Gunjan, Vinit & Pooja, & Kumari, Monika & Kumar, Amit & Allam, Col & Rao, Appa. (2012). Search engine optimization with Google. 10.13140/2.1.1634.5288.