Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

An Optimal Ranking Approach for Cluster based of Clicked URLsusing Firefly Algorithm for Efficient Personalized Web Search

Vikesh Shakya
Department of Computer Science/Information Technology
Madhav Institute of Technology & Science, Gwalior
Madhya Pradesh, India

Abhilash Sonker
Department of Computer Science/Information Technology
Madhav Institute of Technology & Science, Gwalior
Madhya Pradesh, India

Abstract:Research has been finished based on query development to link the gap between terminology of user search query and documents related to user in order to recover appropriate documents before time in search results. In this paper novel approach is proposed which utilizes firefly algorithm for creating the cluster based optimal ranked clicked URLs for efficient personalized web search. The key input query issued for search on the web is used to choose the cluster for recommending the set of terms for query expansion. The process of retrieval of web search results using input query expanded with the selected terms along with the recommendations of next set of query terms for query expansion continues till the search is personalized by the data need of the user. The outcomes exhibit that our proposed work is greatly improved than the existing work.

Keywords:Web page, URL, Web log mining, Genetic Algorithm, Firefly Algorithm, Web Mining, Data Mining.

I. INTRODUCTION

These days all web search tools utilize some sort of Web page link-based ranking in their ranking algorithms. Without uncertainty, this has been the consequence of the success of Google, and its PageRank link algorithm. In all published link ranking algorithms, all links have the same importance. However, web page developers give more importance to some links using different HTML tags, because some Web resources are more vital than others. Thus, a link ranking strategy that gives diverse weights to links may enhance over uniform weight links.

We put forward a variation of PageRank that offers weights to link in view of three attributes: relative position in the page, tag where the link is contained, and span of the anchor words. Our outcomes show that our algorithm, WLRank, enhances over PageRank.

The idea at the back of PageRank is that good pages reference good pages. Hence, pages that are referenced by good pages have higher PageRank. Regardless of the way that there are a few definitions of PageRank, we use the random surf metaphor. Suppose that you are a user surfing the Web in a random fashion, such that, if you are in a page, with certain probability you get bored and leave the page, or you choose uniformly at random to follow one of the links on the page where you are (removing self-links). In this manner, the possibility of being in page p is

$$PR(p) = \frac{q}{T} + (1-q)\sum_{i}\frac{PR(r_i)}{L(r_i)}$$

Where T is the total number of pages, q is the probability of leaving page p (in the original work q = 0.15 is suggested), r_i are the pages that point to page p, and L(r_i) is the number of links in page ri. These values can then be utilized as page ranking, and can be processed by an iterative algorithm converging very fast, as we are interested in the ranking order. The term q is known as damping factor as reducesexponentially link spamming based in sequences of links that return to a page [1].

II. URL

Our concentrate in this paper is on proficient and extensive-scale duplication of documents on the WWW. Web pages which have the similar content yet are referenced by different URLs, are known to cause a host of problems. Crawler assets are wasted in bringing duplicate pages, indexing requires bigger storage and relevance of results are weakened for a query.

Duplicate URLs are available because of many causes, for example:

- Making URLs web search engine well disposed, e.g., http://en. wikipedia.org/wiki/Casino_Royale and http://en. wikipedia.org/?title=Casino_Royale.
- Session-id or cookie information present in URLs, e.g., sid in http://cs.stanford.edu/ degrees/mscs/faq/index.php?sid=67873&cat=8 and http://cs.stanford.edu/degrees/mscs/faq/index. php?sid=78813&cat=8.
- Irrelevant or superfluous components in URLs, e.g., http://www.amazon.com/Lord-Rings/dp/B000634DCW and http://www.amazon.com/dp/B000634DCW
- Removing/adding index files for instance index.html and default.html by web servers.
- Webmasters at times, construct URL representations with custom delimiters, e.g., http://catalog.ebay. com/The-Grudge_UPC_043396062603_W0QQ_fclsZ1QQ_ pcatidZ1QQ_pidZ43973351QQ_tabZ2 and http:// catalog.ebay.com/The-Grudge_UPC_043396062603_ W0?_fcls=1&_pcatid=1&_pid=43973351&_tab=2.

An estimate by demonstrate that roughly 29 percent of pages in the WWW are duplicates and the magnitude is expanding. Unmistakably, this prompts for an efficient solution that can execute de-duplication without getting the content of the page. As duplicate URLs have particular patterns which can be used

for de-duplication, in this paper we concentrate on the issue of de-duplication of web-pages utilizing just URLs without getting the content [2].

III. URL MINING

Web log mining is study of web log files with web pages sequences. Web mining is broadly classified as web content mining, web usage mining and web structure mining. Web content mining is a procedure of finding information from huge number of sources across the WWW. User communication on the web is recorded on a web logs. As each user interaction corresponds to a mouse click & it is often referred as click stream. Click stream is a grouping of URLs scan by a user within a specific website in one session.

URL Mining is one of the subsequences of Web Mining in which we can recognize the actual developer of URL, the date on which this URL was made, the reason of URL et cetera.

The need of URL Mining resides in its past work which incorporated some shortcomings as follows:

- Query Log Analysis
- Clustering URL‘s which include.

1. Text Free Pages: A distance function computed based on the connection between two web pages, for e.g. the connection between two pages can’t be found out if a webpage contains just a picture of Emu and the other webpage contains just appearance and behavior of Emu.

2. Pages with limited access: URL‘s may be password secured or temporarily inaccessible, making its cluster rarely or totally unusable[3].

IV. GENETIC ALGORITHM

Genetic algorithm is stochastic search techniquewhich have been processby the procedure of biological development. Due to gas strength and their uniform ways to deal with vast number of various classes of issues, they have been utilized as a part of numerous applications. Genetic algorithm can be utilized either to enhance criterion for other sort of data mining algorithms or to find knowledge by itself. The benefit of Genetic algorithm becomes more understandablewhen the search space of a task is extensive. Genetic algorithms are a pursuit and transformativeadvancement approach which is motivated by Darwin's theory of evolution. And this technique used in computing to discover correct or estimated solution to optimization and search problems [4].

V. OPERATORS FOR GENETIC ALGORITHM

A. Crossover

Crossover probability crossover the parents to frame another offspring (children). We utilize a Subset Size-Oriented Common Feature Crossover Operator (SSOCF) which keeps useful pieces and creates offspring’s which have the similar distribution than the parents. Offspring’s are kept, only if they fit better than the least good individual of the population.

B. Mutation

During the mutation stage, a chromosome has a likelihood to mutate. If a chromosome is chosen to mutate, we pick randomly a number n of bits to be flipped then n bits are chosen randomly and flipped. It is utilized to maintain and introduce diversity in the genetic population and is usually applied with a low probability. If the likelihood is very high, the GA gets reduced to a random search.

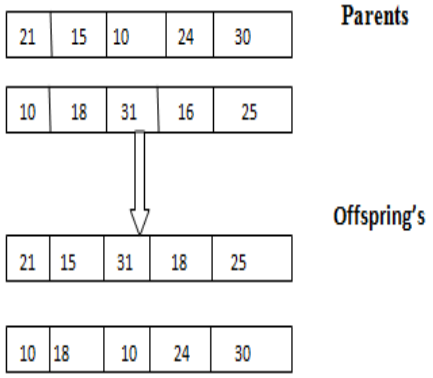

Fig.1 mutation

C. Selection

Selection is the phase of a GA in which discrete genomes are selected from a populace for later reproducing (utilizing the crossover operator).

Fig. 2 selection

VI. LITERATURE SURVEY

Fayyaz Ali et. al (2016) in this paper, Web Information Retrieval (IR) has been successful with page-ranking algorithms that order web pages in view of their rankings and relevance. These ranking algorithms are one of the success factors behind today's popular web search engines including Ask, Bing, Google, and Yahoo! etc., with Google on the top since long. Besides other ranking signals, Google uses PageRank algorithm in ranking the search results, which makes Google successful and superior to others. Since its origin in 1998, it has been at the heart of Google's ranking system and considered a serious breakthrough in ranking web pages on the Web. Researchers and scientists followed this linkbased strategy and came up with similar ranking algorithms including Weighted PageRank, which together with PageRank, have been the focus of research articles covering several aspects and properties. In this analysis, we report an exact investigation of PageRank algorithm and Weighted PageRank algorithm with respect to the property of convergence. Results of the study show that both these algorithms are limited especially with respect to convergence. Based on these outcomes, we present a new flavor of PageRank called Ratio-based Weighted PageRank that performs better than PageRank and Weighted PageRank algorithms [5].

Andrea Morichettaet. al (2016) In this paper, we present CLUE, Clustering for URL investigation, an approach that uses clustering algorithms, i.e., unsupervised methods created in the data mining field to extract knowledge from inactive perception of URLs conveyed by the network. This is a challenging issue given the unstructured arrangement of URLs, which, being strings, call for particular methodologies. Inspired by text-mining algorithms, we present the idea of URL-distance and utilize it to make clusters of URLs utilizing the well-known DBSCAN algorithm. Probes real datasets

indicate empowering outcomes. Well-isolated and predictable clusters rise and enable us to recognize, e.g., malicious traffic, advertising services, and third party tracking systems. More or less, our clustering algorithm offers the way to get bits of knowledge on the data conveyed by the network, with applications in the security or privacy protection fields [6].

Dr. Daya Gupta et. al (2016) in this paper, WWW has become the major source of information dissemination . Due to its vast expansion and heterogeneity, users face difficulty in finding relevant results quickly. Ranking is an important application of web mining which is based on the structure, content and usage. Many algorithms exist for web page ranking and these algorithms are based upon one or more parameters such as forward links, backward links, contents and user interaction time. The efficiency of an algorithm may be based upon the parameters that are applied to determine the ranking of the page. In this paper some important page ranking algorithms are discussed and a new page ranking algorithm is proposed named as User Preference Based Page Ranking. The proposed algorithm is efficient in terms of relevancy because it uses agents to determine pages content relevancy and user behavior is also considered while ranking the web pages. Hence User Preference Based Page Ranking makes users search result navigation easier and more satisfactory to find the desired information [7].

PatiñoGalván, et. al (2016) in this paper, The educational evaluation requires of analysis and continue strategies to adapted the current context, so that this research presents the need to define models of educational evaluation with adaptive characteristics to the area of knowledge and the students to predict behaviors of academic performance. This need is based on the theoretical framework and with the reconciliation of state of the art to outline and define alternative solutions based in Intelligent Computing [8].

MaziyarGramiet. al (2016) in this paper, Today, development of internet causes a fast growth of internet shops and retailers and makes them as a main marketing channel. This type of marketing generates a numerous transaction and data which are potentially valuable. Utilizing data mining is an alternative to discover frequent patterns and association rules from datasets. In this paper, we utilize data mining strategies for finding frequent customers' buying patterns from a Customer Relationship Management database. There are various algorithms for this reason, for instance, Apriori and FPGrowth. However, they may not have efficient performance when the data is big, therefore various meta-heuristic techniques can be an alternative. In this paper we first excerpt loyal customers by using RFM criterion to face more reliable answers and create relevant dataset. Then association rules are found using proposed genetic algorithm. The outcomes showed that our proposed approach is more proficient and have some distinction in compare with other techniques mentioned in this research [9].

Yitong Lu et. al (2016) in this paper, The group structure as a indispensable property for complex networks contributes a great deal for understanding and detecting inherent functions of real networks. However, existing algorithms which are running from the optimization-based to model-based techniques still need to be strengthened further as far as their robustness and accuracy. In this paper, a sort of multiheaded slime molds, Physarum, is utilized for optimizing genetic algorithm (GA), because to its intelligence of creating scavenging networks. In this way, a Physarum-based Network Model (PNM) is proposed in light of the Physarum-based Model, which demonstrate an ability of recognizing inter-community edges. Combining PNM with a genetic algorithm, a new genetic algorithm, called PNGACD, is putting forward to improve the GA's efficiency, in which a priori edge recognition of PNM is included into the period of initialization. Outcomes demonstrate that there is a remarkable enhancement in term of the robustness and accuracy, which demonstrates that PNGACD has a better performance, compared with the current algorithms [10].

Lissa Rodrigues et. al (2015) in this paper, Nowadays, web has become widespread in terms of availability of contents related to every field. Also a large repository of web contents is turned up as a most challenging tool for searching and recovering information. For scientists and researchers, resource or content searching has been very important. Todays, market is brimming with variant search tools over web having discrepancy in terms of working and the end search results. Search tools normally return a huge number of applicable web pages, for given query. In order to give more productive result, this paper presents an advanced approach which considers web content mining and web structure mining towards ranking of web pages to give relevancy of user's query [11].

Lissa Rodrigues et. al (2015) in this paper, World is brimming with information. The World Wide Web serves fills in as real getting such information. Due to the changing nature of web plenty of web pages are deleted and added newly. Every time a surfer looks web utilizing the search engine, data ought to be new and relevant. Recovering efficient, important and significant information from these large sources of information is very challenging job. Because of the large size of web, relating to any query made by a user number of pages are being retrieved so the outcome should be ordered in the manner that most significant webpages is on top of the list. With a specific end goal to get most significant pages at top and reduce the issue of theme drift we propose a hybrid approach of Enhanced-Ratio Rank and Page level keyword algorithms [12].

OuessaiAbdessamed (2015) in this paper, Web page classification based on topic or sentiments is a common application of web content mining techniques. In this paper we will show a new application intended to identify the nation targeted by a specific web page. The aim is to have the capacity to recognize websites focusing on a particular nation, utilizing both the URL and the content of a web page. In this paper we will discourse the issue of distinguishing Algerian-interest web pages utilizing a machine learning technique. We will present the process of gaining data for the supervised learning stage and adjusting it into a usable dataset, as well as using it to construct three distinct classifiers utilizing different parts of the data. The subsequent classifiers have demonstrated outstanding performances (up to F-score = 0.93) for such implementation [13].

VII. PROPOSED WORK

In the existing work, genetic algorithm is used for the personalization of clicked URLs by generating the clusters. In genetic algorithm, there is a large production of population which is not efficient method for the clicked URLs. So we can overcome this problem by utilizing firefly algorithm in place of GA.

The principle reason for a FA is to perform as a signal system to draw in other fireflies. Any firefly can attract the other firefly individually for the movement from one place then onto the next. They have to obey these two conditions: (i) Attractiveness is relative to their brightness, and for any two fireflies, the less bright one will be pulled in by the brighter one; however, the intensity (apparent brightness) diminish as their common distance increases, (ii) If nobody is brighter than a specific firefly, it moves arbitrarily. The brightness ought to be related with the objective function.

Proposed algorithm
Step:1 Get the dataset from the database
Step:2 Apply K-means clustering for cluster query
Step:3 Fetch clustered query with the list of clicked URLs
Step:4 Now for each cluster we apply firefly algorithm
Step:5 Evaluate light intensity(clicked URLs) I_i
Step:6 Set iteration=1
Step:7 While(iterations<=max)

```
For (i=1; i<=cluster_size; i++)
{
    For (j=1; j<=cluster_size; j++)
    {
            If (Ij>Ii) firefly I move towards j
    Else
    Evaluate new solutions
    }
    }
  Ranked clicked URLs and Update best solution
  Iteration++End
```

VIII. RESULT ANALYSIS

Rank Table

S. No.	URL	PWS	PWS with GA	PWS with FA
1.	http://www.tutorialspoint.com/dip/	0.016173	0.004902	0.000277
2.	http://www.google.co.in/ur	0.013947	0.013947	0.000500
3.	http://www.texec.org/contact.html	0.042408	0.016173	0.000513
4.	https://teamtreehouse.com/tracks/web-design	0.013719	0.016423	0.027277
5.	http://www.scaleunlimited.com/about/web-mining/	0.000860	0.030101	0.034158
6.	http://www.google.com/search	0.004097	0.095415	0.036557

The graph below show the elapsed time between the base and proposed work, which demonstrate that the proposed run in less time than the base work. PWSGA is our base technique in which GA is used and PWSFA is the proposed technique requires less time for the execution.

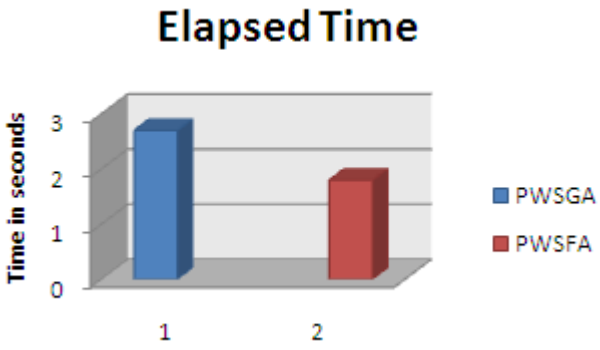

Fig 3 Elapsed Time between PWSGA and PWSFA

IX. CONCLUSION

Web search is customized utilizing high scent clicked URLs in view of clustered user query sessions. The clustered query sessions contain High Scent clicked URLs which satisfy related details require of users on the web where Information Scent is an evaluation of significance of clicked URLs regarding the details require of the user query session. In this paper an algorithm is offered for expansion of query in which firefly algorithm is applied on clustered query sessions keeping in mind the end goal to produce the gathering of most good terms in a precise field for recommendations. We get the improve ranking of clicked URLs and provide the better results for the web search query.

X. REFERENCES

[1] Ricardo Baeza-Yates Emilio Davis, "Web Page Ranking using Link Attributes", ACM 1-58113-912-8/04/0005.
[2] Hema Swetha Koppula, Krishna P. Leela, Amit Agarwal, "Learning URL Patterns for Webpage De-duplication", WSDM'10, February 4–6, 2010, New York City, New York, USA. Copyright 2010 ACM 978-1-60558-889-6/10/02.
[3] Chinmay R. Deshmukh, Prof. R .R. Shelke, "URL Mining Using Agglomerative Clustering Algorithm", ISSN: 2277 128X/Volume 5, Issue 1, January 2015.
[4] Ranno Agarwal, "Genetic Algorithm in Data Mining", ISSN: 2277 128X/Volume 5, Issue 9, September 2015.
[5] Fayyaz Ali, Irfan Ullah, Shah Khusro, "An Empirical Investigation of PageRank and Its Variants in Ranking Pages on the Web", 978-1-5090-5300-1/16 $31.00 © 2016 IEEE.
[6] Andrea Morichetta, Enrico Bocchi, Hassan Metwalley, Marco Mellia, "CLUE: Clustering for Mining Web URLs", 978-0-9883045-1-2/16/2016 ITC.
[7] Dr. Daya Gupta, Devika Singh, "User Preference Based Page Ranking Algorithm" ISBN: 978-1-5090-1666-2/16/2016 IEEE.
[8] Patiño Galván, "Educational Evaluation and Prediction of School Performance through Data Mining and Genetic Algorithms", 978-1-5090-4171-8/16/2016 IEEE.
[9] Maziyar Grami , Reza Gheibi , Fakhereh Rahimi, "A Novel Association Rule Mining Using Genetic Algorithm", 978-1-5090-4335-4/16/2016 IEEE.
[10] Yitong Lu, Mingxin Liang, Chao Gao,Yuxin Liua, Xianghua Li, "A Bio-inspired Genetic Algorithm for Community Mining", 978-1-5090-4093-3/16/2016 IEEE.

[11] Lissa Rodrigues, Shree Jaswal, "Hybrid Model for Improvised Page Ranking Algorithm", Hybrid Model for Improvised Page Ranking Algorithm", 978-1-4673-9825-1/15/20 15 IEEE.

[12] Lissa Rodrigues, Shree Jaswal, "An Efficient Page Ranking Approach Based On Hybrid Model", 978-1-4799-1734-1/15/2015 IEEE.

[13] Ouessai Abdessamed, Elberrichi Zakaria, "Web Site Classification based on URL and Content: Algerian Vs. non-Algerian Case", 10.1109/ISPS.2015.7244974/ 2015IEEE.