RESEARCH ARTICLE

DOI: 10.34256/ijcci2112

International Journal of Computer Communication and Informatics

An Efficient Mining Approach for Handling Web Access Sequences

Nandhini. R [a,*], **Evangelin Sonia S.V.** [a,*]

[a] Department of Computer Science and Engineering, Sri Shakthi College of Engineering and Technology, Coimbatore, Tamil Nadu, India.

* Corresponding Author: nanishreesha@gmail.com, evangelinsonia@siet.ac.in

Received: 19-03-2021, Revised: 04-05-2021, Accepted: 05-05-2021, Published: 30-05-2021

Abstract: The World Wide Web (WWW) becomes an important source for collecting, storing, and sharing the information. Based on the users query the traditional web page search approximately retrieves the related link and some of the search engines are Alta, Vista, Google, etc. The process of web mining defines to determine the unknown and useful information from web data. Web mining contains the two approaches such as data-based approach and process-based approach. Now a day the data-based approach is the widely used approach. It is used to extract the knowledge from web data in the form of hyper link, and web log data. In this study, the modern technique is presented for mining web access utility-based tree construction under Modified Genetic Algorithm (MGA). MGA tree are newly created to deploy the tree construction. In the web access sequences tree construction for the most part relies upon internal and external utility values. The performance of the proposed technique provides an efficient Web access sequences for both static and incremental data. Furthermore, this research work is helpful for both forward references and backward references of web access sequences.

Keywords: Genetic Algorithm, Classification and Regression Tree, Hyper Text Transfer Protocol, Internet Protocol, Structured Query Language

1. Introduction

The way toward separating helpful and interesting data from the data storehouses is as called mining. In this modern era, information plays a vital role. Earlier using elegant technologies like computers, satellites, etc., enormous information are collected and stored in mass storage devices. Vast collection of data resulted in a mess and this leads to the structuring and managing of data in a well- organized manner by the usage of databases. Database Management System (DBMS) helps to store and retrieve data from the large repositories efficiently using queries.

Web mining is the derived concept from data mining, which extracts the information directly from web services, web documents, hyperlinks, web contents, and web server logs. It mainly concentrates on the World Wide Web (WWW) that includes its primary source, components, and contents. The data contents are extracted from a website that would be the collection of web pages and it contains structured data. It represents tables, lists, images, audio, and video. Web mining is used to determine the information from web data in the data mining process. In addition, it provides a robust to a web search engine by analyzing web content and web document categorization. It is most useful for e-services and e-commerce applications. Figure 1.1 shows the web mining services.

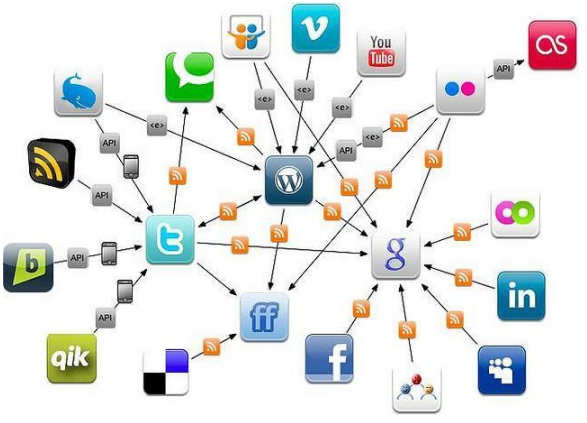

Figure 1.1 Web Mining Services

Also, web mining is used to understand the customer behavior and evaluates the particular web site effectiveness [1]. WWW contains diverse dynamic, massive, and mainly unstructured data that provides a huge amount of information. Web growth gives to some issues such as determining relevant data through the internet and observes user.

Web usage mining is the mining technique, which applied to determine the user access patterns from web repositories. When the user visits the web pages, automatically web servers record the user information such as URL, IP Address, Hits, and weblog file. This file is the input for web usage mining. The proposed novel hybrid approach improves web usability with two attributes such as Hit and Time Spent. The web server logs contain the information of user sessions and user-oriented tasks. The user session provides information about the user spent time inappropriate website. Moreover, it generates the web site ranking accurately with a clustering approach. To motivate the successful web access sequences, we have used the web utility mining system with utility web access. Solutions are offered for the search challenges by the proposed hill-climbing optimization approach. In addition, the genetic algorithm is one of the optimum processes that encompass the extensive issues then by utilizing the local optimums, the complex search space is also solved.

The rest of this paper is formed as follows, related animal classification work in section 2, the proposed approach explained in section 3, material and method described in section 4, results discussion parts presented in section 5 conclusion in section.

2. Litrature Review

Now a day, the internet development was incredible. The huge measures of data from relevant data to users find it extremely difficult. The issues can be solved by web usage mining which includes preprocessing. [2] designed a new technique to identify sessions in Web usage mining (WUM). This was mainly focused on the preprocessing approach. Unnecessary records comprised of graphics files; robots are removed in the data-cleaning phase. In the next phase, identification of sessions, this was derived by forming the user behavior in a matrix format. Matrix comprised of rows and columns in which columns indicate the web pages and rows indicates the users and their sessions are identified. The experimental results showed that the session identification method was effective and accurate.

Pamutha, Chimphlee, Kimpan, & Sanguansat (2012) [3] discussed data preprocessing method for mining user's access patterns on web server log files. WUM is to convert a log into a set of web user sessions. A web log file was gathered from the web server and focused on the preprocessing of the weblog file methods that can be used for the task of session identification. The resulted study produced statistical information on user sessions.

Maheswara Rao & Valli Kumari (2011) [4] implemented an extensive research framework capable of preprocessing web log data. The learning algorithm of the proposed research framework can isolate human user and search engine accessed with less time. The framework reduced the error rate and improved significant learning performance. This framework aided to investigate web user usage behavior effectively. The result showed that the employment of the proposed framework of IPS provided a promising solution in dynamic weblog development.

Pathak, Shah & Almeera (2014) [5] presented an algorithm for pattern discovery based on the association between the users' accessed web pages. This paper discussed a complete preprocessing method to identify distinct users. The association rule-mining algorithm is to find the frequently accessed web pages. The biggest constraint for mining web usage patterns are computation and memory overhead. The experimental result showed that the algorithm was efficient and scalable.

(Huang et al., 2015) [6] presented an AutoODC (Auto Orthogonal defect classification) approach to automate ODC classification by forming it as a supervised text classification issue. ODC is a framework used for software defect analysis and classification, which provides a valuable in-process feedback to system development and maintenance. It is promising approach. This paper trained AutoODC with the support of two machine learning algorithm for support vector machine, Naïve Bayes and text classification and estimated it on both industrial and larger

defect list where the industrial defect was reported from social network domain and larger defect list was extracted from open-source system.

FileZilla. This approach achieved overall accuracy of 83% (NB) and 81% (SVM) on the industrial defect report and accuracy of 77 % (NB) and 75 % (SVM) on the larger defect list. The preprocessing techniques are used to convert the raw data into data abstraction based on the required users, sessions, and page views. The recommendations and ranking techniques are used to assign rank to the web page according to the impact of the webpage. The tree-based approaches are used to construct the Utility based web tree in high utility web access sequences.

3. System Design

The clustering is used to grouping the web session based on similarity and it maximizes the intra-frame similarity [7]. The web session contains hyperlink clicks. Clustering web session topics have the most popular in various applications. In web mining, the log file defines three steps such as data gathering, filtering, and formatting of log entries. Various algorithms are presented for pattern discovery named.

- Clustering
- Sequential pattern analysis
- Rule mining
- Classification

However, the clustering acts to robust for determining the web sequences. For determining the similarity between two web sites, first, it represents the URL as a token. In this similarity computation, we have to compare the corresponding token at the beginning and comparison will stop when the tokens are stopped.

Figure 2 defines the website tree structure in the clustering session based on the user-accessed website. The clustering session is an important factor in web mining and analyses of user access behavior.

The main challenge is to determine both forward and backward web sequences. To recover this issue, the proposed method is presented with tree construction and MGA. This tree construction combines the two trees of SVM tree and IGA tree. This proposed tree construction detects the user access patterns in large database scans. The innovative web access utility is clearly shown in this picture,

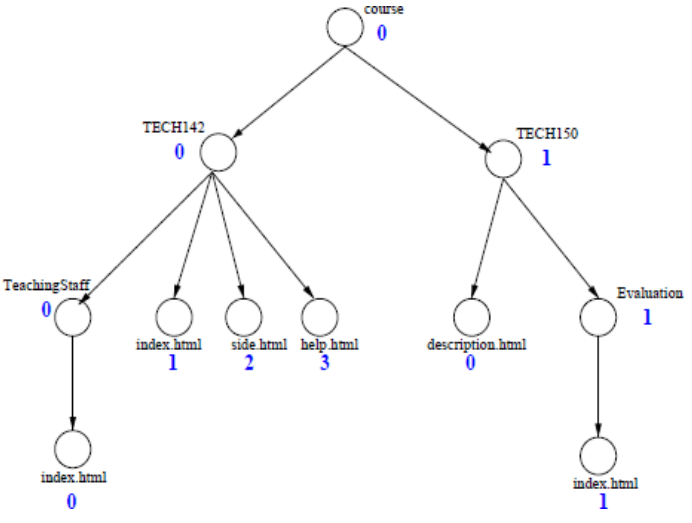

Figure 2. Website tree structure

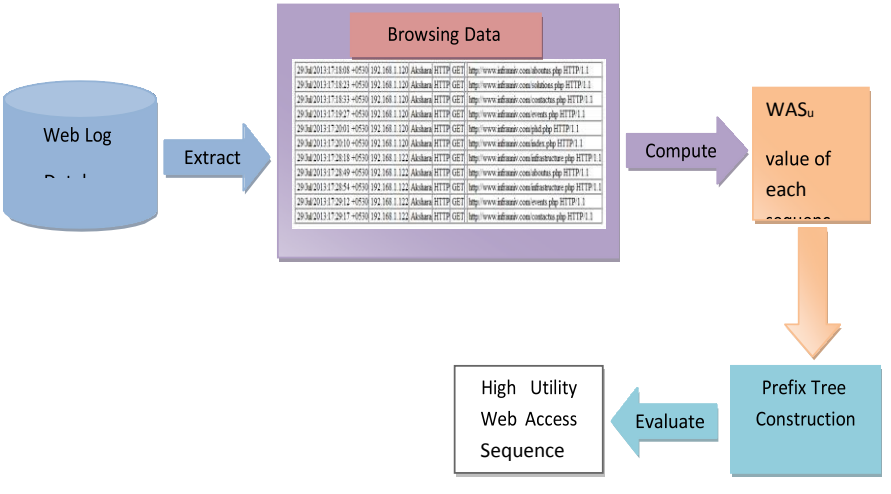

Figure 3. Flow of the proposed method

3.1 Hill Climbing Algorithm

It is oneof the local search algorithms, and it is used to solve the optimization problems in AI. It also called a greedy approach [8]. It continuously moves increasing direction to determine the peak of a mountain or to determine the best solution for a problem. After it reaches the peak value, it terminates when no neighbor has a higher value. It mainly used for optimizing mathematical problems. Traveling salesman problem is the example of a hill-climbing algorithm; it needs to reduce the distance traveled by a salesperson. This algorithm contains two basic components such as state and value.

- Estimate the initial or primary state, or when it is a goal state then return success and stop.
- Loop until the solution is determined or there is no operator left to apply.
- The operator is applied to the current state.
- Identify new state

  i. It the state is goal state, it returns success and stop.

  ii. Else, if it is greater than the current state, then allocate a new state as the current state.

  iii. Else, if it is not better than the current state then back to step 2
- Exit

3.2 Genetic Algorithm

A genetic algorithm is an optimization technique and heuristic search that mimic the natural evolution process. Optimization defines to determine the best set of output values from the set of input values. In web mining, the meta search engine searches the requests by yahoo, vista. The individual search engine results are combined as a single result set. Meta search engine improves the consistent interface and coverage. N number of potential solutions for optimization problems categorizes genetic search.

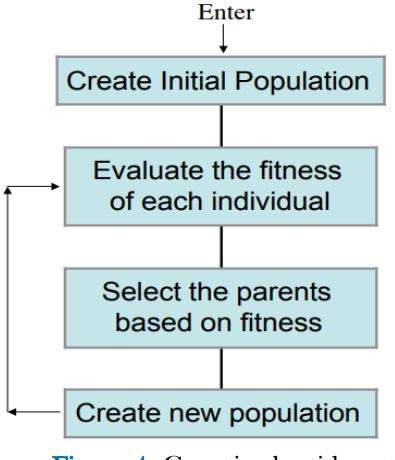

Figure 4. Genetic algorithm steps

- Initially, the GA algorithm initializes the parameters for optimization.
- Then, determine the chromosome representation of parameters.
- Thirdly, generate the individuals of the initial population.
- Then, evaluate the fitness function for each individual.

- Create a new population-based on random behavior or selection rules.

Step 1: Randomly create the initial solution (where, i = 1, 2… n).

Step 2: Evaluate the fitness function

$$Fitness\ function = sum\ of\ the\ total\ weight\ for\ each\ user \quad (4.14)$$

Each parameter of the fitness value is estimated and shortlisted the greatest fitness value as the best chromosome.

Step 3: To achieve the best solution, relate the mutation and crossover

Mutation: According to the probability, the chromosome values are varied.

Figure 5. psudocode for proposed improved Genetic algorithm steps

4. Result and Discussion

The proposed method performances are evaluated from FDR rate, tree construction time, and runtime and memory location by adjusting the threshold value. False Detection Rate (FDR) defines the rate of a false positive and false negative in the null hypothesis when acquiring multiple comparisons.

For threshold value 0.1, the SVM tree contains 0.004 FDR value and the IGA tree has 0.003 FDR value. For threshold value 0.15, the SVM tree contains 0.0056 FDR value and the IGA tree has 0.004 FDR value. For threshold value 0.2, the SVM tree contains 0.009 FDR value and the IGA tree has 0.0084 FDR value. For threshold value 0.25, the SVM tree contains 0.019 FDR value and the IGA tree has 0.012 FDR value. Figure 6 defines the statistical results of FDR value for both tree SVM and IGA.

Table 1. False Detection Rate

Threshold	SVM	IGA
0.1	0.004	0.003
0.15	0.0056	0.004
0.2	0.009	0.0084
0.25	0.019	0.012

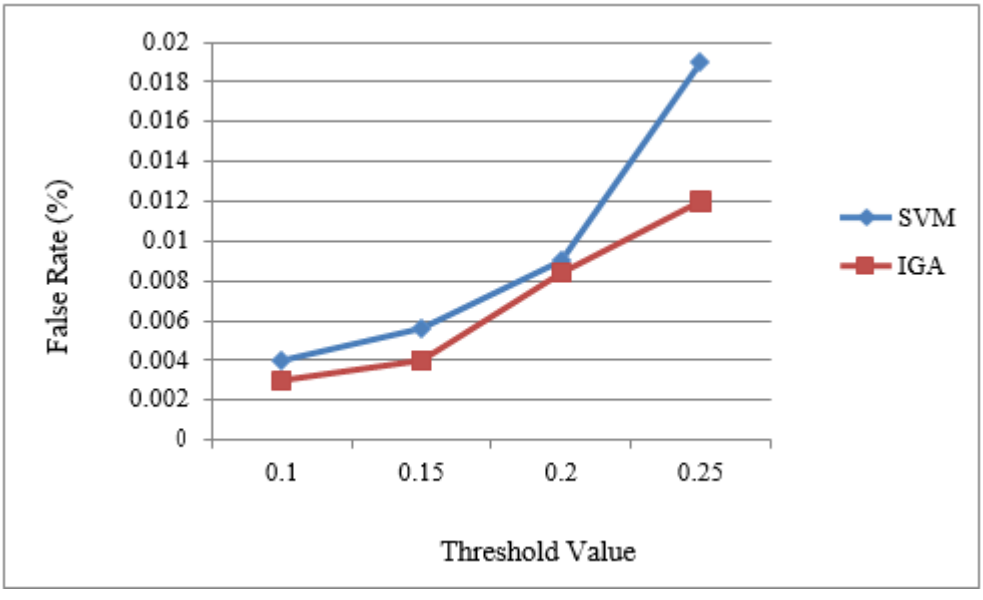

Figure 6. FDR value for both tree SVM and IGA

Table 2 describes the tree construction time for both tree SVM and IGA. Time expended for the construction of the tree is assessed by altering the value of the threshold. In the SVM tree, when the value of the threshold is 0.1, time devoured to the tree is observed to be 12s and furthermore, the IGA. tree is 18s for comparing time. At the point when the value of the threshold is set to 0.15 then the SVM and IGA tree construction time values are observed to be 8s and 9s. At the point when the value of the threshold is 0.2, tree construction time 7s for SVM and IGA is 9s. At the point when the value of the threshold is altered to 0.25, tree construction time 7s for SVM and IGA of the relative time values are observed to be 9s. Figure7 depicts the statistical analysis of Tree construction time based on the threshold value. Based on this results SVM tree has minimum execution time compared to the IGA tree.

Table 2. **Tree** construction time for both tree SVM and IGA

Threshold value	**Tree construction time (sec)**	
	SVM tree	IGA tree
0.1	12	18
0.15	8	9
0.2	7	9
0.25	7	9

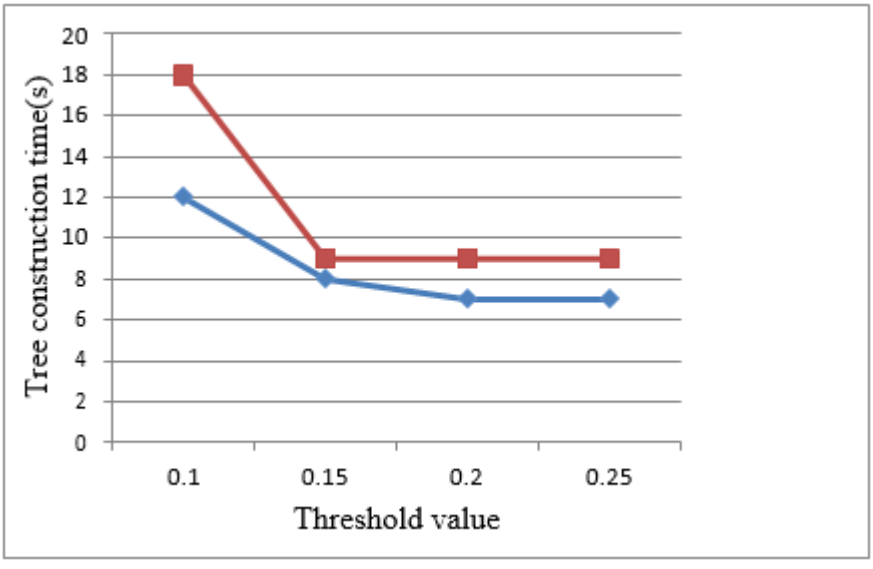

Figure 7. Tree construction time vs. threshold value

Table 3 defines the memory allocation for both tree SVM and IGA. For 0.1 threshold value, the SVM tree has 278808 memory allocation times, and IGA contains 299874. For 0.15 threshold value, the SVM tree has 278896 memory allocation times, and IGA contains 278945. For the threshold value 0.2, the SVM tree has 2778726 memory allocation times, and IGA contains 281451. For the 0.25 threshold value, the SVM tree has 279184 memory allocation times, and IGA contains 277818. Figure 5.4 depicts the statistical analysis of memory allocation for both SVM and IGA tree.

Table 3. Memory allocation for both tree SVM and IGA

Threshold value	Memory allocation (bits)	
	SVM tree	IGA tree
0.1	278808	299874
0.15	278896	278945
0.2	277872	281451
0.25	279184	277818

The proposed method determines both internal and external web access sequences. The results section contains the performance measures of HUWAS and HIUWAS FDR rate, tree construction time, and run time and memory location by adjusting the threshold value. The

comparative analysis compares the proposed method accuracy with various existing methods and it proved the proposed method has the highest accuracy.

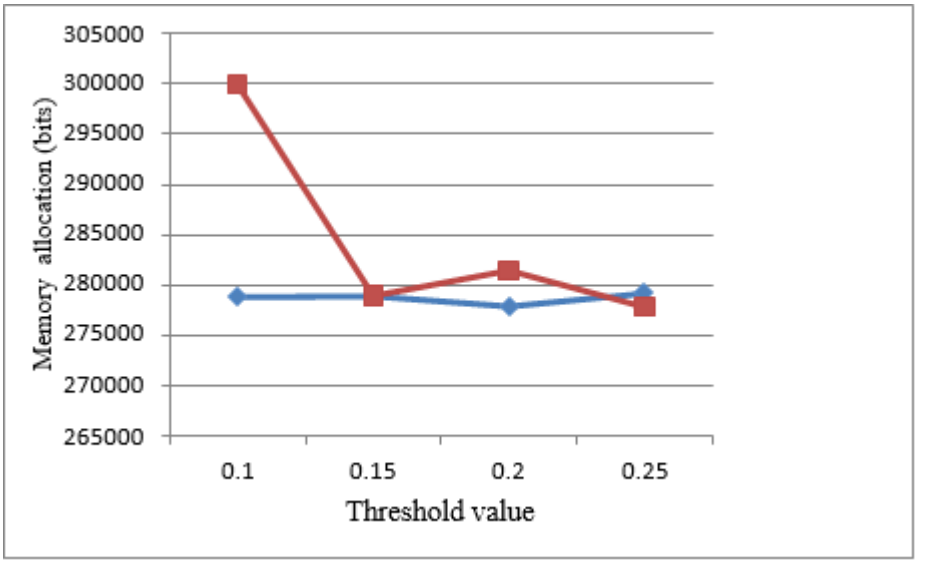

Figure 7. Memory allocations for both SVM and IGA tree

5. Conclusion

In this study, the main research is web usage mining. Web usage mining is the important factor in wide range of applications such as business intelligence, recommendation, web traffic, customer attraction, system improvement and cross sales. proposed the Hybrid Hill Climbing Genetic Algorithm (HHCGA) based on tree construction for extracting the web access sequence. For tree construction, it designed with HUWAS tree (HHCGA and Utility-based Web Access Sequence tree) and the HIUWAS tree (HHCGA and Incremental Utility-based Web Access Sequence tree). This utility-based approach determines both forward and backward references of the web access sequence. In evaluation results, the performance measures of HUWAS and HIUWAS FDR rate, tree construction time, and run time and memory location were evaluated by adjusting the threshold value. From this performance analysis, it is observed that the proposed technique provides an efficient Web access sequences for both static and incremental data.

References

[1] Neelima G., Rodda S. (2016). Predicting user behavior through sessions using the web log mining. *In 2016 International Conference on Advances in Human Machine Interaction (HMI),* IEEE, 1-5.

[2] Chitraa V., Thanamani A.S. (2011) A novel technique for sessions identification in web usage mining preprocessing. *International Journal of Computer Applications*, 34 (9) 23-27.

[3] Pamutha T., Chimphlee S., Kimpan C., Sanguansat P. (2012). Data preprocessing on web server log files for mining users access patterns. *International Journal of Research and Reviews in Wireless Communications (IJRRWC)*, 2

[4] Rao V.M., Kumari V.V. (2011). An Enhanced Pre-Processing Research Framework for Web Log Data Using A Learning Algorithm. *Computer Science and Information Technology,* 10 (5121) 1-15. http://dx.doi.org/10.5121/csit.2011.1101

[5] Pathak N., Shah V., Ajmeera C. (2014). A Memory Efficient Algorithm with Enhance Preprocessing Technique for Web Usage Mining. *In Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*,1-6. https://doi.org/10.1145/2677855.2677902

[6] Huang L., Ng V., Persing I., Chen M., Li Z., Geng R., Tian J. (2015). AutoODC: Automated generation of orthogonal defect classifications. *Automated Software Engineering,* 22 (1) 3-46. https://doi.org/10.1007/s10515-014-0155-1

[7] Vellingiri J., Kaliraj S., Satheeshkumar S., Parthiban T. (2015). A novel approach for user navigation pattern discovery and analysis for web usage mining. *Journal of Computer Science*, 11 (2) 372 -382. http://dx.doi.org/10.3844/jcssp.2015.372.382

[8] Burke E.K., Bykov Y. (2017). The late acceptance hill-climbing heuristic. *European Journal of Operational Research*, 258 (1) 70-78. https://doi.org/10.1016/j.ejor.2016.07.012

Acknowledgements

The authors declare that they have no conflict of interest.

Conflict of interest

The authors declare that they have no conflict of interest.

About The License