Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 13 (2020) Pp-258-264

BBRC
Bioscience Biotechnology Research Communications

An Efficient path Completion and Construction with Candidate key Constrained Clustering Algorithm in web Mining

J. Umarani[1], S. Manikandan[2] K. Balasubramanian[3] and G.Thangaraju[4]

[1]*Research and Development Centre, Bharathiyar University, Coimbatore*
[2]*Department of Information Technology, K. Ramakrishnan College of Engineering, Trichy*
[3]*Department of Computer Applications, PSNA College Of Engineering And Technology., Dindigul*
[4]*Department of Computer Science, Government Arts and Science College, Veppanthattai-621116, Perambalur-dt, Tamil Nadu, India*

ABSTRACT

Web usage Mining is a kind of the web analysis, pre-processing stage in WUM consists the following Data Cleaning, user identification, session identification and path completion (path added). In this research article focuses the path completion part. Web log is the most valuable input of the web analysis with WUM. Web log information is collected from the server, client and proxy server; in these missed some of the information's. It results missing access references, user access patterns are not clearly identified by incomplete access log. To rectify these issues path completion takes the role to acquire the missing reference. Different approaches are available to find the missing reference like Url, IP address, reference length but no one can be efficient. But our proposed candidate key constrained clustering algorithm for path completion can utilize the both type web log file client and server log file and also provide the better results in efficiency rather than the other path completion and construction techniques.

KEY WORDS: WEB USAGE MINING, PRE-PROCESSING, PATH COMPLETION, CANDIDATE KEY, CLUSTERING.

INTRODUCTION

In current arena of World Wide Web, have a numerous types of applications invented by the software industry and others for their specific needs and common to all (Prabu et al. 2019). The applications are utilizes by consumer are end user depends upon their requirements. A platform needs to communicate the industry provided application and end user [2]. Nowadays all types of communications either Government to citizens, C2G, B2C, B2B, like all necessity is transmitted via the WWW. Research and development of computer sciences are focus their research in different aspects, mining a species of information in a huge volume of data by the use data mining technology (Dixit and Dwivedi 2017). Data mining technology covers the different sub categories, Web mining is one among them, and web mining is a technology to find the piece of data from the WWW for the requestor based query. et al., 2020; Mathew, Roopa and Soni, 2020;

A). Web Mining And Web Usage Mining: Web mining consist the three sub categories they are; 1.Web structure mining 2. Web content mining and 3.Web usage mining. Web structure mining is the study of web page schema of collection hyperlinks (Rooba and ValliMayil 2015). Web content mining is the process to extract data from the content of a web page. Example: image, audio,

ARTICLE INFORMATION

*Corresponding Author: smk76dgl@gmail.com
Received 11th Oct 2020 Accepted after revision 27th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

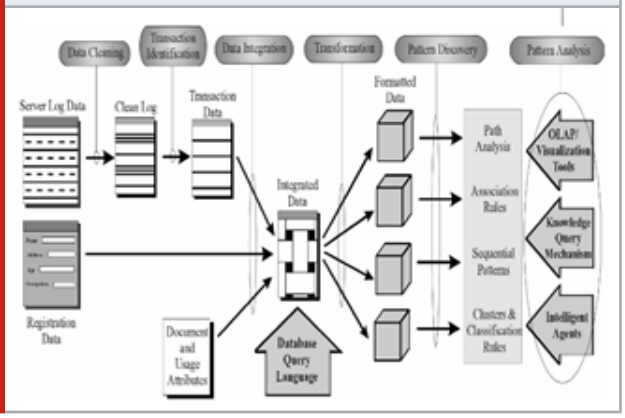
Clarivate Analytics

crossref

NAAS Journal Score 2020 (4.31)
A Society of Science and Nature Publication, Bhopal India 2020. All rights reserved.
Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.13/38

video, metadata and hyperlinks. Web usage mining process divided into three phases they are; 1. Data-pre-processing 2.Pattern discovery 3. Pattern analysis. Figure.1: represents the overall process involved in the web usage mining. The first phase of WUM is the Data-pre-processing, the source for this phase is the collection of web logs, and it is collected from web server, proxy server, and client or web browser. The second phase is pattern discovery, in this phase take the pre-processed information for their discovery purpose and produce the results. The third phase is pattern analysis; discovered patterns are the inputs for this phase, pattern analysis done by the On Line Analytical Processing tools.

Figure 1: Overall Architecture of Web Usage Mining



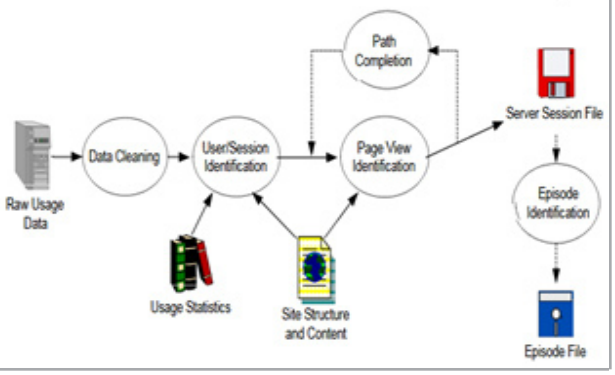
B).Data Pre-Processing In Web Usage Mining: Pre-processing of data is the first stage in the web usage mining, it is accomplished through the different phases, and the first one is the Data cleaning. Itis the primary role of the pre-processing (Sidana and Aggrwal 2017). Web log file contains the lot of information's some of them are no needed for analysis they are removed in this stage. The following figure.2: represents the overall architecture of the Data-pre-processing. In second stage of the data cleaning process is User Identification; the cleaned web log file is on input for this, form the web log which user can access the web pages to be found with the different heuristics. Like the third one also Session Identification, this stage also take the cleaned web log and find sessions with different heuristics. The final stage of the pre-processing is the Path Completion. In this research article focus the path completion. The following sections are describing the detailed research aspects of the path completion.

C).Web Log Files: Web page access history is stored in file that is called web log files. It is automatically generated if the user clicks or requests a page. Each time of a page access log file should be updated automatically. A web log file is located in the following locations; 1. Web server 2. Web proxy server 3. Client browser.

Web server logs– It provides more accurate and complete usage of data to web server.

Web proxy serve logs - It takes HTTP request from user, gives them to web server, then result passed to web server and return to user (Bhavani et al. 2017). Client send request to web server via proxy server.

Figure 2: Over all Architecture of Data-Pre-processing



Client Browser logs: It can reside in client browser alone. In the form HTTP cookies.

C. Web Log Files: Web page access history is stored in file that is called web log files. It is automatically generated if the user clicks or requests a page. Each time of a page access log file should be updated automatically. A web log file is located in the following locations; 1. Web server 2. Web proxy server 3. Client browser.

Web server logs: It provides more accurate and complete usage of data to web server.

Web proxy serve logs: It takes HTTP request from user, gives them to web server, then result passed to web server and return to user (Durgadevi 2018). Client send request to web server via proxy server.

Client Browser logs: It can reside in client browser alone. In the form HTTP cookies.

A). Type of Web log files: There are four types web log files are there; Access Log file, Error Log file, Agent Log file, Referrer Log file.

Access Log file: All incoming request data's and information about client of server.

Error Log file: Internal Error generated by the server. The page is being requested by the client to the web server.

Agent Log file: Information about user browser name and the other details.

Referrer Log file: It contains the information about link and redirects visitors to site.

```
"%h %l %u %t \"%r\" %>s %b"
THRC/access_log_common
eg: 127.0.0.1 RFC 1413 frank
[20/Jan/2018:17:35:33 -0700] "GET
/apache_pb.gif HTTP/1.0" 200 2326
```

B). Web Log File Formats: It is the standardized text file format that is used by most ofthe web servers to generate the log files. The configuration of common log file format is given below in the box.

Literature Survey: (Dixit and Dwivedi 2017) conducted a survey on path completion and other techniques of web usage mining, they studied nearly sixteen research article related to the web usage mining and other techniques, out of their study they represents the web usage mining process in detail, data cleaning, user identification, session identification and pattern discovery. In addition to that they also depicts the some of web usage mining application and also represents performance of web usage mining, what are the requirements to the web usage mining , represents the functionality of web usage mining with neat diagram. Finally they represent about the path completion techniques, in web log contains detailed log information, but some of the pages have back catch of page and link. This type of information not available in the web log file, it is available only the client machine alone, the path completion process to identify the missing catch pages and links, and added to the web log file.

(Rooba and ValliMayil 2015) presents a review for server log data processing in web usage mining, referred nearly fifteen papers and presents the detailed description of the server log information, and also give the necessary steps involved in preprocessing like data cleaning, user identification, session identification, path completion and transaction identification. In path completion stage, the suggests the three approaches for path completion namely; 1. Reference length approach 2. Maximal forward reference 3.Time window these approaches are used to find out the missing reference and links. Finally depicts the techniques in transaction identification which deals with two kinds of transactions travel path transactions and content only transactions and also presents detailed comparison table preprocessing methods.

(Sidana and Aggrwal 2017) reviewed various web mining algorithms and techniques that have been used by the previous researchers. In detailed analysis of the web mining help to identify the benefits and limitations of these techniques. Along with this, we have provided a proper process of web usage mining with three phases including the preprocessing, pattern discovery and pattern analysis. The reviewed research will help us in the further research on the topic. B.Bhavani et al. [6] reviewed the ten papers five among them are represents the web usage mining techniques and remaining of them are web usage mining application related. Web usage mining consist the different techniques such as the data pre-processing, it includes the data-preprocessing, user, session identification and path completion. The another technique pattern discovery, it includes the data mining techniques such as Association, clustering, sequential pattern and classification. Pattern Analysis includes the OLAP, data and knowledge querying, usability analysis and visualizations technique. The author analysis the web usage mining applications such as the personalization of web content, perfecting and caching, support to the design and E-commerce applications.

(Durgadevi 2018) presents a web mining and web usage mining essentials, provides the necessary steps to collect web log data and also discussed the web log data in two formats CLF, ECLF. Referred nearly eight papers gather the techniques implemented for web usage mining and finally came to the conclusion to implement what type techniques for their proposed methods. (Babu et al. 2011) suggest more about WUM, web usage mining model is a kind of mining to server logs. WUM plays an important role in enhancing the usability of website design, the improving the requirements of system performance and improvement of customer's relations. It also cover the another concepts like the personalization of server and other business making decisions. They discussed the all the activities of the web usage mining growth. They also proposed a new frame work Online Miner seems to work well for developing predication models to analyze the web traffic volume.

(Faizan and Kankale 2016) review the web usage mining related papers and presents the essentials of web mining, web usage mining concepts, they provides their own framework for preprocessing steps. Data cleaning process consists the five steps they are elimination of local and global noise, removal records of graphics, videos and the format information, removal of records with failed HTTP status code, method field and Robots Cleaning. (Mary and Baburaj 2013) presents a web mining content in detail, elaborately discusses about the data collections consists of the server logs, neatly presents a data pre-processing phases with block diagram, each phase of the pre-processing elaborately discussed by the authors.

(Padmapriya and Maheswari2017) presented various details regarding data pre-processing activities that are necessary to perform web usage mining. In each phase of the pre-processing, they give some rules to design and implement them simply and efficiently. Their proposed method is used to reduce the size of the log file but also increases the worth of the data available. The path completion process which is used to find out the missing pages and append lostpages and construction of transactions in pre-processing stage. (Prabha and Suganya 2017) studied eleven papers presented a basics of web mining and web usage mining concepts, also presents a few web usage mining algorithm concepts like association rules, clustering, classification and sequential patterns. Finally discussed about the web usage mining techniques implemented by the previous authors.

METRIAL AND METHODS

In this research article the path completion stage of the pre-processing is focused. In this part of the article how the data's collected for experiment analysis and what are the major parameters used for analysis are carried out.

Data Collection: The Web logs are collected from the Internet Server which is located at Thanthai Roever

Group of Educational Institutions, In that institutions have a one centralized server and four proxy servers, these are located in the respective institutions among the group. And also few of client logs are also collected from the client machines connected to the proxy servers. The following are the server's details;

1. RMS01-RoeverMainServer(198.162.1.100)
2. RMS02-RoeverMainServer(198.162.1.101)
3. RECPROXY-(198.162.2.100)
4. THRCPROXY-(198.162.3.100)

The academic year of college is started from June month of the every year. The above mentioned records are collected from these servers on the period of January 2018to April 2018. Table 1. Provides the details of log servers and number of records collected from the respective servers.

Table 1. Log files Server description with Number of records

S.No	Server Description	No.of.records
01	RMS01-RoeverMain Server(198.162.1.100)	10000
02	RMS02-RoeverMain Server(198.162.1.101)	15000
03	RECPROXY-(198.162.2.100)	35000
04	THRCPROXY-(198.162.3.100)	40000
05	Client log file from client machines	25000
	Total Number of records	125000

Figure 3: Web Browser History on March 2018

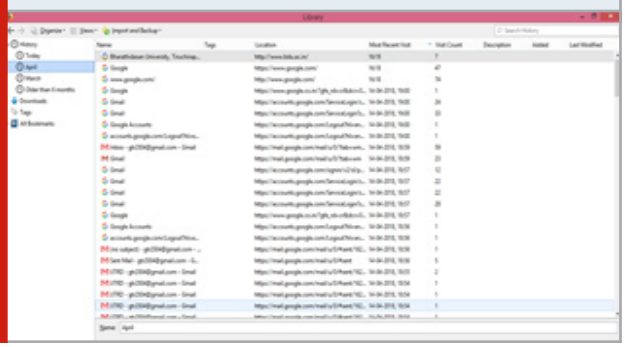


Figure 4: Web Browser History on April 2018

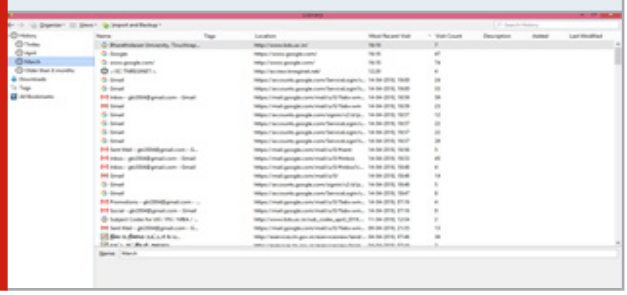


Figure 5: Web Log content March 2018

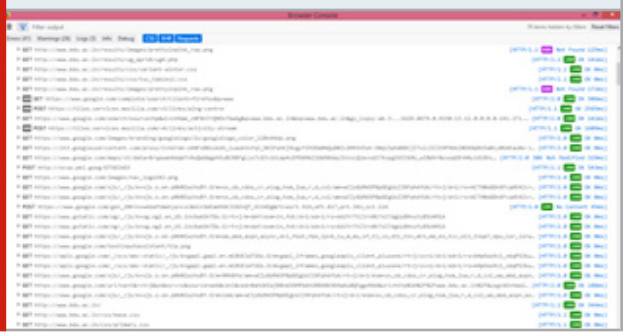


Figure 6: Web Log content April 2018

Sample Web Log and Cookies Collection:

Figure 7: Sample web page access list

```
http://www.bdu.ac.in/
http://www.bdu.ac.in/about-bharathidasan-university.php
http://www.bdu.ac.in/campuses.php
http://www.bdu.ac.in/docs/BDU_Act_1981_June_1992.pdf
http://www.bdu.ac.in/docs/BDU_Statutes.pdf
http://www.bdu.ac.in/iqac/
http://www.bdu.ac.in/nirf/
http://www.bdu.ac.in/administration/chancellor.php
http://www.bdu.ac.in/results/
http://www.bdu.ac.in/results/ug_apr18/
http://www.bdu.ac.in/results/ug_apr18/ug8.php
http://www.bdu.ac.in/results/pg_apr18/index.php
http://www.bdu.ac.in/results/pg_apr18/pg9.php
http://www.bdu.ac.in/results/index_nov_2017.php
http://www.bdu.ac.in/results/mphil/
http://www.bdu.ac.in/timetables/apr2017/ug/BSc_Semester_CBCS_2008_V1.pdf
```

Sample web page access with URL

Reference Length Computation

Ref_length=RLtdf - bysr/c
Where Ref_length- Reference Length time,

RLtdf-Reference Length time difference access time current record and other record
Bysr-byte sent or receives from the server to client, c-data transfer rate.
Companied Session

Proposed Work: In pre-processing stage of web usage mining includes the data cleaning, user identification and session identification and final step is Path Completion. It is one of the critical steps in pre-processing; it finds out the missing pages/references in web log file and adds the missing page into the URL. There are some of the access are not recorded in the access log like agent cache, local cache, post techniques and reset, back button clicks. With this sequence the number of URL (Uniform Resource Locator) is less than the real one. The ultimate aim is to discover the exact travel pattern of the user and the missing pages/references in the user access path is appended. For this mentioned above purpose different authors are proposed their own view. Here we saw the two of them; the first one is path analysis that is finding the missing pages.

Figure 8: Overall Architecture of Proposed Path Completion technique

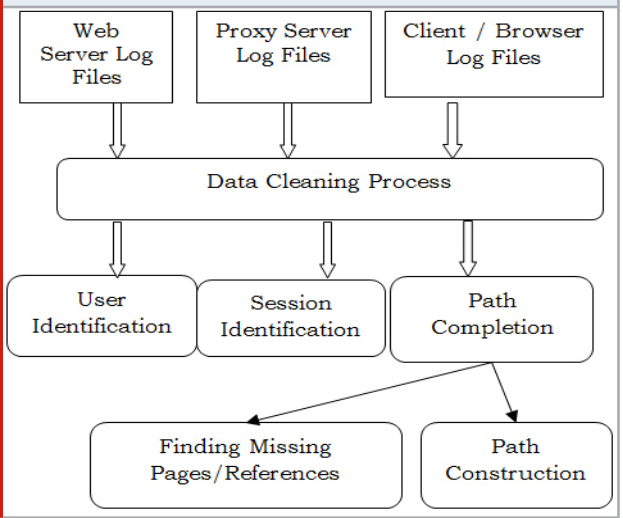


Figure 9: Session list combined algorithm

Input: Session Construction Output- (IPAH, TMOH, RFRH)
Output: Cumulative_list_of user and session
Step 1: create the file named SCF
Step 2: Set the file in open mode with write enable
Step 3: Read the session construction result from the file of IPAH
Step 4: Write the results to the file SCF
Step 5: Read the session construction result from the file of TMOH
Step 6: Append the results to the file SCF
Step 7: Read the session construction result from the file of RFRH
Step 8: Append the results to the file SCF //Session Construction File
Step 9: Stop.

Session List Companied Algorithm (Slca)

Figure 10: Proposed CKCPCACCA

Input: Selected user and session list output of SLCA algorithm
Output: Path Combination and Completion
Step 1: Begin
Step 2: Read the list from the file SCF
Step 3: Assign RS=Number of records in a file // RS-Record Set
Step 4: do
Step 5: while RS>=1
Step 6: read the records in a file
Step 7: if (Url_i=Url_j) then // to check Consecutive Url is same
Step 8: PS = {Uid_1, Sid_1, $Date_1$, $Page_1$,.....,Uid_n,Sid_n,$Date_n$, $Page_n$,}
Step 9:elsif(Url=ReferrerUrl)then // Url is referred by other Url
Step 10: PS= {Uid_1, Sid_1, $Date_1$, Ref_length_1,.....,Uid_n,Sid_n,$Date_n$, Ref_length_n)
Step 11: elsif(PageAccess=BAP) then //BackAccessPage
Step 12: PS= {Uid_1, Sid_1, $Date_1$, BAP_1,.....,Uid_n,Sid_n,$Date_n$, BAP_n)
Step 13:elsif(PAT>PATL) then //PageAccessTime&PageAccessTimeLimit
Step 14: PS={Uid_1, Sid_1, $Date_1$, $PATP_1$,.....,Uid_n,Sid_n,$Date_n$, $PATP_n$)
// PageAccessTimePath
Step 15: else, Display the message "Invalid Record"
Step 16: endif
Step 17: endif
Step 18: endif
Step 19: endif
Step 20: RS=RS-1
Step 21: Display the resultant PS // Path Set based on the group or cluster
Step 21: end

Table 2. The Access path of on TWO user sessions

Date	Session	Access Page Number	Referrer Number
05.04.2018 10:15:07	S1	20	--
05.04.2018 10:15:09	S1	22	20
05.04.2018 10:15:10	S1	23	22
05.04.2018 10:15:12	S1	25	23
05.04.2018 10:15:13	S1	27	26
05.04.2018 10:15:15	S1	37	26
05.04.2018 10:15:16	S1	25	24
05.04.2018 11:35:40	S2	35	--
05.04.2018 11:35:42	S2	37	35
05.04.2018 11:35:43	S2	41	37
05.04.2018 11:35:45	S2	43	41
05.04.2018 11:35:46	S2	50	43

Candidate Key Constrained Path Completion And Construction Clustering Algorithm [Ckcpcacca]

In this proposed algorithm we use the candidate key to find access level and access paths of pages, candidate

key construction in four types they are as follows;

1. Userid+SessionId+Date+Page 2. Userid+SessionId+Date+ReerenceLength 3.Userid+SessionId+Date+BAP 4.Userid+SessionId+Date+PATP

In this algorithm Figure 10. Represents to get the inputs from the SCF file; it contains the combined session construction records which include the fields, userid and data of access etc., the number of records available in a file is assigned to RS variable. Repeatedly executes or verify the each record in file until the end of the records reached. In the verification of candidate key sequence the first type verify the records which is in the forward page access sequence then construct the page set as PS = {Uid1, Sid1, Date1, Page1,.....,Uidn,Sidn,Daten, Pagen,}, if it is not in the sequence to check the next key condition, if it is referred by other Url or link then construct the page set as follows; PS= {Uid1, Sid1, Date1, Ref_length1,.....,Uidn,Sidn,Daten, Ref_lengthn). if the Url or Page link not referred but backward click then the page set is constructed as follows; PS= {Uid1, Sid1, Date1, BAP1,.....,Uidn,Sidn,Daten, BAPn). if the page reference not backward click but it will have more time taken that is exceeds their time limit then page access constructed as follows; PS={Uid1, Sid1, Date1, PATP1,.....,Uidn,Sidn,Daten, PATPn). Finally display the resultant PS which clustered order.

Table 3. The Path Completion Result

	Session 1	Session 2
Page Sequence	20-22-23-25-27-37-25	35-37-41-43-50
Combination	20-22-23-25-37	35-37-41-43-50
Path Completion	20-22-23-25-37	35-37-41-43-50

Figure 11: The Path Completion Category with No.of. Records

RESULTS AND DISCUSSION

The following table 2 represents the referrer URL based experiment results which consists the date, session, access page number and referrer page number field details with two sessions on same date. These data's are applied to our prosed algorithm and got the resultant ant path completion is depicted in the table 3.

Table 4. The Path Completion Category with No.of.Records and Paths constructed

S.No	Path Construction Category	No.of. records	No.of. Paths Constructed
1	Url and Page Sequence	45000	5000
2	Reference Length	35000	3000
3	Back Access	15000	1500
4	Page Access Time	15000	1500

Figure 12: The Path Completion Category with No. of. Paths constructed

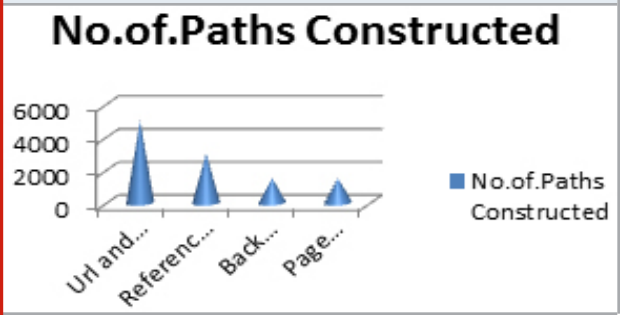
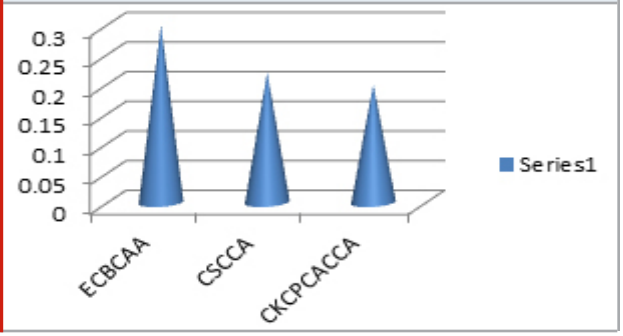


Table 5. Time Comparison of Three Algorithms used for the Path Completion

S. No.	Algorithms	Size of the File	Time Complexity
1	ECBCAA	120 KB	0.3 Seconds
2	CSCCA	120 KB	0.22 Seconds
3	CKCPCACCA	120 KB	0.20 Seconds

Figure 13: Time Comparison of Three Algorithms used for the Path Completion



CONCLUSION

The proposed candidate key constrained path completion and construction clustering algorithm [CKCPCACCA] is executed with the four clustering category and find the missing reference then build path in appropriate

path sequence. The time consumption is minimum for this proposed algorithm compared with the previous algorithms namely ECBCAA, CSCCA. So this technique is best among the three algorithms used for path Completion and Path Construction Process.

REFERENCES

Prabu, S., Lakshmanan, M. and Mohammed, V.N., 2019. A multimodal authentication for biometric recognition system using intelligent hybrid fusion techniques. Journal of medical systems, 43(8), pp.1-9

Kumar, M.K., Parameshachari, B.D., Prabu, S. and liberata Ullo, S., 2020, September. Comparative Analysis to Identify Efficient Technique for Interfacing BCI System. In IOP Conference Series: Materials Science and Engineering (Vol. 925, No. 1, p. 012062). IOP Publishing.

Varun Dixit and Abisshek Dwivedi (2017), "A Survey on path completion and various techniques in web usage mining", International Journal of LNCT, Vol 1(1) pp:16-21.

Rooba R, Dr.ValliMayil V (2015)," Review the Steps of Server Log Data Processing for Web Usage Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol.3, Issue 11, pp:11819-11825.

Arjun Sidana and Dr. Himanshu Aggrwal (2017), "Review of web usage of data mining in web mining", International Journal of Advanced Research in Computer Science, Volume 8, No.5.

Bhavani B, Dr. Sucharita V and Dr. Satyanarana K.V.V. (2017), "Review on Techniques and Applications Involved in Web Usage Mining", International Journal of Applied Engineering Research, Volume 12, Number 24, pp.15994-15998.

Durgadevi D, "A Discovery on Web usage mining using Preprocessing", International Journals of Computer Trends and Technology (IJCIT), Volume 49 Number 1

Dr SureshBabu D, Abdul Nabi SK, Mohd Anwar Ali and Raju Y (2011), "Web Usage Mining: A Research Concept of Web Mining", International Journal of Computer Science and Information Technologies, Vol. 2(5), pp: 2390-2393.

Faizan I Khandwani and Ashok P Kankale (2016), "Preprocessing Techniques for Web Usage Mining", International Journal Scientific Development and Research (IJSDR), Volume 1, Issue 4, pp: 330-334.

Prince Mary S, Baburaj E (2013)," An efficient Approach to Perform Pre-Processing", Indian Journal of Computer Science Engineering, Vol. 4 No.5, pp: 404-410.

Padmapriya R and Maheswari D (2017), "A Novel Technique for Path Completion in Web Usage Mining", International Journal of Advance Research, Ideas, and Innovations in Technology, Volume 3, Issue2, pp:1076-1079.

Prabha k and Suganya T (2017), "A Guesstimate on Web Usage Mining Algorithms and Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 6, pp: 518-521.