International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

Subterranean Insect based Data Reduction in Web Usage Mining using K-implies Clustering Algorithm

Dushyantsinh B. Rathod, Ramesh T. Prajapati, Harshil Joshi

Abstract*: Information decrease is the way toward limiting the measure of information that should be put away in an information stockpiling condition. Information decrease can build stockpiling effectiveness and lessen costs. Information cleaning act in the Data Preprocessing and Web Usage Mining. The work on information cleaning of web server logs, unessential things and futile information can not totally evacuated and Overlapped information causes trouble during information recovering from database. Right now, we present Ant Based Pattern Clustering Algorithm to get design information for mining .It likewise shows Log Cleaner that can sift through a lot of superfluous, conflicting information dependent on the basic of their URLs. Fundamentally right now are expelling undesirable records . so we are utilizing k-implies bunching calculation . By utilizing this exploration work we can apply this philosophy on web based business stage i.e AMAZON, FLIPKART.

Keywords* : Information Mining, Clustering, Data Reduction, Ant based bunching, Web utilization Mining.

I. INTRODUCTION

Web Mining is system in information mining to separate information from web information, including web archives, hyperlinks between reports, utilization logs of sites, and so on. In Web Mining, information can be gathered at the server side, customer side, intermediary servers, or gotten from an association's database (which contains business information or merged Web information). There are numerous sorts of information that can be utilized in Web Mining. As per information investigation objective, web mining can be partitioned into three unique types, which are web use mining, web content mining and web structure mining. Web utilization mining is the procedure of extricating compelling data from web server logs. Grouping investigation assumes a significant job in information mining field. Information can be assembled into various classes or bunches by grouping investigation. There exists better comparability among the articles in a similar class and more unfortunate likeness among the items in various classes.

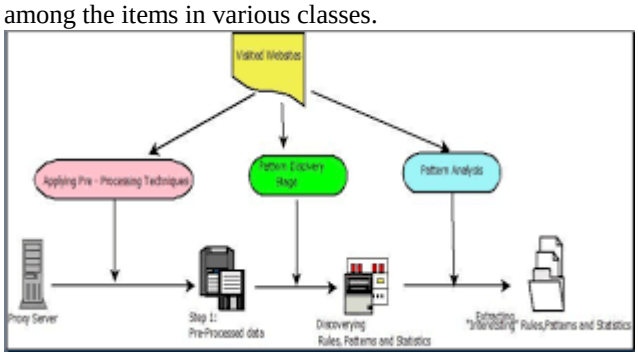

Figure 1: Web usage mining process

II. OBJECTIVE

This paper proposes a grouping strategy dependent on Ant Colony Optimization. For grouping insect based example bunching calculation is applied to pre-prepared logs to extricate visit designs for design disclosure.

III. SORT OF LOG FILE FORMAT

As of late, three arrangements are accessible to catch these records:-

1. W3C (World Wide Web Consortium) Extended Log document Format

2. Microsoft IIS (Internet Information Services) Log File

3. NCSA (National Center for Supercomputing Application) Ordinary Log document Format

All the three are ASCII content formats. Logging information are recorded in four-digit year design in NCSA and W3C Extended designs. The two digit year group is utilized in Microsoft IIS log position before 1999 and after that four-digit design is utilized

A. W3C Log File Format (World Wide Web Consortium)

W3C Extended log is an adaptable ASCII design which has various sorts of fields.

Revised Manuscript Received on March 16, 2020.
* Correspondence Author
Dr. Dushyantsinh B. Rathod*, Associate Professor and HOD, C.omputer Engineering Department, Alpha College of Engineering and Technology Ahmedabd. Gujarat, India Email: dushyantsinh.rathod@gmail.com
Dr. Ramesh T. Prajapati, is currently working as Assistant Professor and HOD, Department of Computer Science Engineering, School of Engineering, Indrashil University, India.Email: rtprajapati1984@gmail.com
Mr. Harshil Joshi, Assistant Professor cum Research Fellow, Depsrtment of Computer Science & Engineering, Devang Patel Institute of Advance Technology and Research, FTE, Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India. Email: mailtoharshil@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

These fields can be separated by spaces. Time can be reported as UTC (Coordinated Universal Time)[7] Following fields are appeared in fig : Client IP-Address, Time-stamp, Strategy, Protocol Status, URI Stem and Protocol Version.

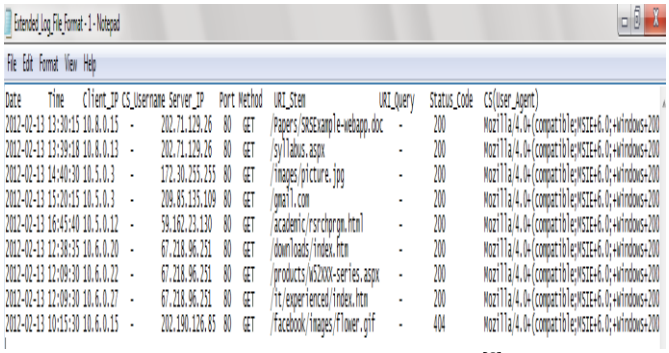

Fig.-2 W3C Log File Format[6]

Prior section shows that on 02-05-2002 at 05:42 P.M., a client with HTTP form 1.0 and the IP address 172.22.255.255 gave an HTTP GET direction for/Default.htm document.
The #Date: field assigns when the main most log section was made and log was made. The #Version: field used to imply the W3C log position. A hyphen (—) appeared in the field shows a placeholder.

B. IIS Log File Format

Microsoft IIS is a non-flexible ASCII group. This organization can record more data than the NCSA design. The IIS group consolidates things like client's IP address, client name, Service status code, demand date-time, and number of bytes got. What's more, it incorporates definite things like the slipped by time, the quantity of bytes sent, the activity and the target document. Commas are utilized to part these things which makes design simple to decipher than the ASCII position, which use spaces for separating. The time is caught as nearby time. On opening a Microsoft IIS design record in the manager, the passages are seen like the accompanying model in Fig.-3

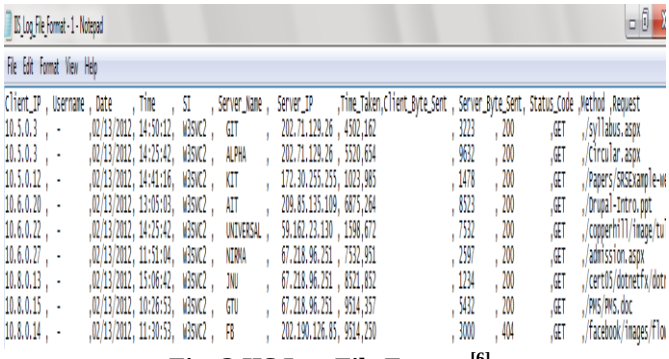

Fig.-3 IIS Log File Format[6]

All the fields are finished with a comma (,). A hyphen(—) fills in as a placeholder for a specific field which has no substantial worth.

C. NCSA Log File Format

NCSA Common arrangement is a non-adaptable ASCII group which is accessible for Web destinations however not for FTP locales. This catches data about client demands like client name, remote host name, time, date, the quantity of bytes sent by the server, HTTP status code and solicitation type. Time can be recorded as nearby time and things can be part by spaces.

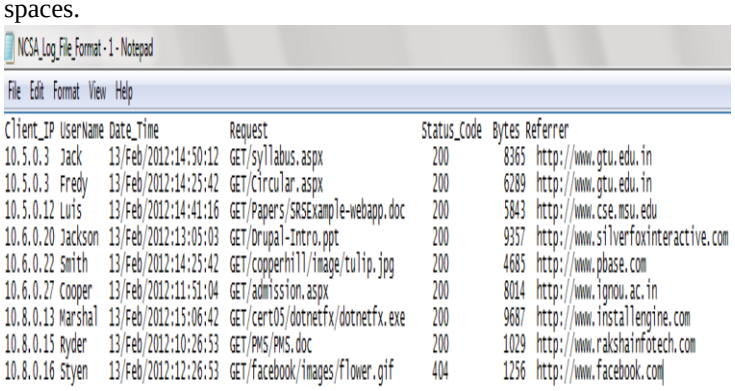

Fig.-4 NCSA Log File Format[6]

IV. NEW METHODOLOGY

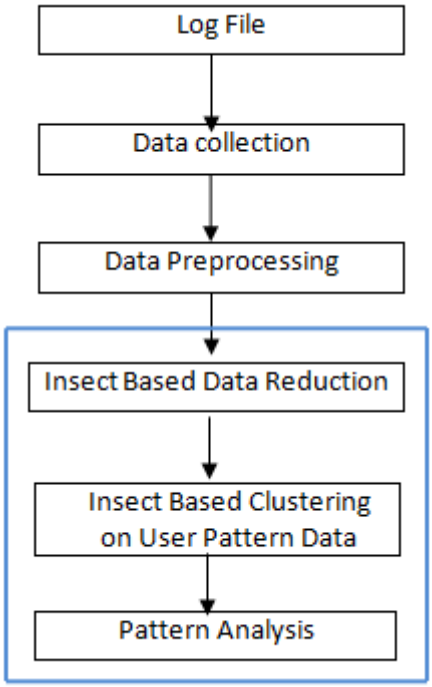

Fig.-4 New Methodology

A. Information assortment

The contribution for the web use mining process is gathered from the web log record. During a client meeting, all route action on the site is recorded in a log document by the web server. It is an enormous vault of site pages and connections, gets to sites are recorded in web logs record. Log record is accessible in two arrangements. The first is the basic log position and broadened log group. _

151.48.123.70 - [08/Dec/2007:00:00:43 - 0800] "GET/img/abull.gif HTTP/1.1" 200 411 "http://www.smsync.com/request/?ref=002" "Mozilla/4.0 (good; MSIE 7.0; Windows NT 5.1)" www.smsync.com
151.48.123.70 - [08/Dec/2007:00:00:43 - 0800] "GET/img/dowld_btn.gif HTTP/1.1" 200 3083 "http://www.smsync.com/request/?ref=002" "Mozilla/4.0 (good; MSIE 7.0; Windows NT 5.1)"

© Copyright: All rights reserved.

www.smsync.com[16]

B. Pre-handling of weblog

By expelling unessential information things for planning log information to investigation called as pre handling. Information cleaning is the initial step of the process. Cleaning of information should be possible by checking the addition of URL name and erasing the sections, for example, JPEG, JPG and GIF. The second step in pre-handling is the User Identification. The necessary fields are extricated from the cleaned log document and put away in the database for additional preparing. Here, IP delivers are considered to distinguish a specific client. After information cleaning and User ID the client meetings are distinguished. A client meeting is viewed as when solicitation of client is inside chosen timeframe . Every client meeting has recognized by the meeting ID [4].

C. Insect Based Data Reduction

Essential subterranean insect bunching calculation was proposed by Deneubourg [3]. As indicated by this model ants have arbitrarily stroll on working region and sense for similitude in close by objects or not. In view of this data, they would pick the component or drop the component. The likelihood of picking and dropping an item relies upon the articles lying in quick condition.

Picking probability of an object i :

$$P_{pick}(i) = \left(\frac{k^+}{k^+ + f(i)}\right)^2$$

Dropping probability of an object i is :

$$P_{drop}(i) = \begin{cases} f(i) \text{ if } f(i) = k^- \\ 1 \text{ otherwise} \end{cases}$$

where, f =estimation of the division of close by focuses which is involved by objects of a similar kind, and K+ known as consistent in this proposed calculation the information which are going to decreased we simply put the banner as opposed to expelling record from informational index . so we can recognize the exhibition and precision bases on Flag records[16].

V. SUBTERRANEAN INSECT BASED PATTERN CLUSTERING ALGORITHM

// it reads N number of record from Frequent Data Set

1. Info Data Set: Read N no of records from clean information source FDS

For i= 1 to i<=N

Next

//it discover each records R from FDS

2. For every record R from information source FDS discover design information

// indicating meaningful information from FDS

3. Peruse design information utilizing indicated address from information source FDS.

//it finds requested records from FDS

4. Whenever mentioned records from visit information source FDS with indicated design at that point

// if records R is same in FDS and PDS then put flag

5. In the event that equivalent record R from FDS = PDS, at that point put FLAG into FDS

// make group or cluster of FDS in PDS

6. Make group in design information source PDS.

// else leave that records

7. Else not select those records.

//end the condition

8. End if

// go for next record

9. Next record.

VI. RESULTS

The information is web log record at that point performing information cleaning to expel superfluous information things. The cleaned web log is utilized for design revelation. The proposed model uses Ant Colony calculation for bunching dependent on client meetings. The clients with pertinent get to examples will go under a similar group.

Index_No	Server_IP	Client_IP	URI_Steam	Status_Code	Page_Request	Flag
0	202.71.129.26	10.8.0.15	/Papers/SRSExample-webapp.doc	200	/alldoc.aspx	0
1	202.71.129.26	10.8.0.15	/syllabus.aspx	200	/os.aspx	0
2	209.85.135.109	10.5.0.54	/starsports.com	200	/cricket.aspx	0
3	59.162.23.130	10.5.0.12	/downloads/index.htm	200	/makemytrip/offer.aspx	0
4	67.218.96.251	10.6.0.20	/downloads/index.htm	200	/admission.aspx	0
5	67.218.96.251	10.6.0.20	/products/W52XXX-series.aspx	200	product/samsung	0
6	67.218.96.251	10.6.0.20	/it/experienced/index.htm	200	/powerbank	0
7	202.71.129.26	10.8.0.15	http://www.flipkart.com/laptops	200	/ac.aspx	0
8	172.30.255.255	10.5.0.20	http://www.flipkart.com/mobiles	200	/mobiles.html	0
9	209.85.135.109	10.5.0.54	http://www.amazon/Electronics	200	/products.aspx	0
10	67.218.96.251	10.6.0.20	http://in.bookmyshow.com	200	moviesinfo.aspx	0
11	202.71.129.26	10.8.0.15	/Papers/SRSExample-webapp.doc	200	/alldoc.aspx	1
12	59.162.23.130	10.5.0.12	/downloads/index.htm	200	/makemytrip/offer.aspx	1
13	202.71.129.26	10.8.0.15	/webapp.doc	200	/laptops.aspx	0
14	202.71.129.26	10.8.0.15	/syllabus.aspx	200	/os.aspx	1
15	209.85.135.109	10.5.0.54	/starsports.com	200	/cricket.aspx	1
16	59.162.23.130	10.5.0.12	/academic/rsrchprgm.html	200	/workshop.aspx	0
17	67.218.96.251	10.6.0.20	/downloads/index.htm	200	/admission.aspx	1

Fig.-5 Mix Clustering

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

Pass No of Cluster: 5 [Cluster Cration]
Cluster No: 202.71.129.26 [Create]

Index_No	Server_IP	Client_IP
0	202.71.129.26	10.8.0.15
1	202.71.129.26	10.8.0.13
7	202.71.129.26	10.5.0.5
11	202.71.129.26	10.8.0.17
13	202.71.129.26	10.8.0.18
14	202.71.129.26	10.8.0.14
20	202.71.129.26	10.5.0.5
24	202.71.129.26	10.8.0.16
26	202.71.129.26	10.8.0.18
27	202.71.129.26	10.8.0.11
33	202.71.129.26	10.5.0.5
37	202.71.129.26	10.8.0.12
39	202.71.129.26	10.8.0.10
40	202.71.129.26	10.8.0.13
46	202.71.129.26	10.5.0.51
50	202.71.129.26	10.8.0.53

Fig.-6 Cluster Creation -1

Pass No of Cluster: 5 [Cluster Cration]
Cluster No: 209.85.135.109 [Create]

Index_No	Server_IP	Client_IP
2	209.85.135.109	10.5.0.54
9	209.85.135.109	10.6.0.26
15	209.85.135.109	10.5.0.51
22	209.85.135.109	10.6.0.28
28	209.85.135.109	10.5.0.55
35	209.85.135.109	10.6.0.29
41	209.85.135.109	10.5.0.12
48	209.85.135.109	10.6.0.21

Fig.-7 Cluster Creation-2

Index_No	Server_IP	Client_IP	URI_Steam	Flag
0	202.71.129.26	10.8.0.15	/Papers/SRSExample-webapp.doc	0
1	202.71.129.26	10.8.0.15	/syllabus.aspx	0
7	202.71.129.26	10.8.0.15	http://www.flipkart.com/laptops	0
11	202.71.129.26	10.8.0.15	/Papers/SRSExample-webapp.doc	1
13	202.71.129.26	10.8.0.15	/webapp.doc	0
14	202.71.129.26	10.8.0.15	/syllabus.aspx	1
20	202.71.129.26	10.8.0.15	www.flipkart.com/laptops	1

Fig.-8 Reduction Data with Flag

Index_No	Server_IP	Client_IP	URI_Steam	Flag
2	209.85.135.109	10.5.0.54	/starsports.com	0
9	209.85.135.109	10.5.0.54	http://www.amazon/Electronics	0
15	209.85.135.109	10.5.0.54	/starsports.com	1
22	209.85.135.109	10.5.0.54	www.amazon/Electronics	1
28	209.85.135.109	10.5.0.54	/gmail.com	0
35	209.85.135.109	10.5.0.54	http://www.amazon/Electronics	1

Fig.-9 Flag Reduction Data with Flag

VII. CONCLUSION

Right now Cleaner sift through approx 60% URL demands with same server IP address which can't be separated by conventional information cleaning techniques for intermediary logs. It make a recurrence get to information and example bunching by executing design grouping systems to produce design bunch for simple access of information from design grouping rather than quality ral database. It additionally improves the future significantly more precise and dependable. It surrenders preferred execution to 60% rather over 30% furthermore, exactness contrast and old calculation . Essentially right now expelled undesirable and copied records. After that we discover the example for visit get to items and make bunch dependent on visit get to items. Which increment the exhibition for bringing the information from database.

ACKNOWLEDGMENT

I have taken efforts in this in this work Preliminaries. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.
I am highly indebted to my friends for valuable guidance and supervision regarding my topic as well as for providing necessary information regarding the dissertation.
I would like to express my gratitude towards my lovely Parents and my wife for their kind co-operation and encouragement which help me in completion of this article.
My thanks and appreciations also go to my friends who helped me out with their abilities.

REFERENCES

1. An Ant-Based Data Reduction Algorithm. Ismail M. Anwar,Khalid M. Salama,Ashraf M. Abdelbar
2. A Hybrid approach for Clustering Weblog Volume 5, Issue 3, March 2015
3. Saroj Bala, S. I. Ahson, R. P. Agarwal ,An Improved Model for Ant based Clustering‖, International Journal of Computer Applications (0975 – 8887) Volume 59– No.20, December 2012
4. Nayana Mariya Varghese, Jomina John ,Cluster Optimization for Enhanced Web Usage Mining using Fuzzy Logic‖, IEEE 2012.
5. Shelokar P S, Jayaraman V K, Kulkarni B D. An Ant Colony Approach for Clustering [J]. Analytica Chimica Acta, 2004, 509: 187-195
6. Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan, Mohamad Mohsin, ―Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm‖, World Academy of Science, Engineering and Technology 48 2008, pp.190-197, DOI: 10.1.1.140.5102
7. Fang Yuan, Li-Juan Wang, Ge Yu, ―Study on Data Pre-processing Algorithm in Web Log Mining‖, IEEE Nov, 2003, pp.28-32 vol.1, ISBN: 0-7803-8131-9

© Copyright: All rights reserved.

8. Nichele C. M. and Becker. K., 2006,―Clustering Web Sessions by Levels of Page Similarity‖, W.K. Ng, M. Kitsuregawa, and J. Li (Eds.): PAKDD 2006, LNAI 3918, pp. 346 – 350, 2006 © Springer-Verlag Berlin Heidelberg 2006
9. Renáta Iváncsy, István Vajk, ―Frequent Pattern Mining in Web Log Data‖ , Acta Polytechnica Hungarica, January 2006
10. Web Usage Mining: A Survey on Preprocessing of Web Log File Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood Department of Computer Science, Muhammad Ali Jinnah University, Islamabad, Pakistan
11. Alphy, S.Prabakaran, ―Cluster Optimization for Improved web Usage Mining using Ant- Nestmate Approach‖, IEEEInternational Conference on Recent Trends in Information Technology, June 3-5, 2011
12. Log files formats‖, http://www.w3c.org, Access Date: [5th of Dec, 2012-10 PM].
13. Kobra Etminani Mohammad-R. Akbarzadeh-T. Noorali Raeeji Yanehsari ,―Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method‖, IFSA-EUSFLAT 2009.
14. Banerjee, A. and J. Ghosh (2001). Clickstream Clustering Chicago (2001)
15. mrs. v. sujatha, dr. punithaval li ,― an approach to user navigation pattern based on ant based clustering and classification using decision trees‖, 2010.
16. Kajal mengar – Ant based data reduction in Web Usage Mining using K-means Clustering Algorithm,2016

AUTHORS PROFILE

Dr. Dushyantsinh Rathod, is currently working as Associate Professor and HOD, Department of Computer Engineering, Alpha College of Engineering & Technology, Ahmedabad, India. His Birth date is 23/09/1983. He has received his Ph.D. Degree in Computer Engineering from Rai University, Ahmedabad, Gujarat, India. He has total 12 Years of experience .His main research interest includes Data Mining, Web Mining and Database Technology. He has been involved in the organization of a number of conferences and workshops. He has been published more than 19 papers in International journals and attended 5+ conferences / workshops. He is also a Ph.D supervisor and Ph.D examiner panel member at GTU since long.

Dr. Ramesh T. Prajapati, is currently working as Assistant Professor and HOD, Department of Computer Science Engineering, School of Engineering, Indrashil University, India. His Birth date is 31/03/1984. He has received his Ph.D. Degree in Computer Engineering from Rai University, Ahmedabad, Gujarat, India. He has total 12 Years of experience .His main research interest includes Grid computing, Cloud computing etc. He has been involved in the organization of a number of conferences and workshops. He has been published more than 19 papers in International journals and attended 5+ conferences / workshops. He is also a Ph.D supervisor and Ph.D examiner panel member at GTU for past 2 years.

Mr. Harshil Joshi, has done Diploma in Computer Engineering from Dharmsinh Desai University. He has done B.E. in Computer Engineering from Veer Narmad South Gujarat University. He has done M.E. in Computer Science & Engineering from Gujarat Technological University. Currently, he is pursuing Ph.D. from Gujarat Technological University. He is interested in various security techniques/algorithms, Internet of Things (IOT) and other prominent areas of computer engineering area. He has more than 9 years of academic experience. He has published various research papers in reputed journals and/or national level conferences. He is currently working as Assistant Professor cum Research Fellow at Department of Computer Science and Engineering (CSE), Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology & Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa (DIST: Anand), Gujarat, India.

© Copyright: All rights reserved.