I J C T A, 9(26) 2016, pp. 293-299
© International Science Press

Analysis and Design of Outlier Detection Method to Improve Web Page Surfing

S. Vasuki [1] and K. Subramanian[2]

ABSTRACT

Web page surfing and extracting results from the global space is a trendiest feature in the current fast moving world. The most important challenge we need to deliberate in this era is unwanted page navigation and page region hanging. This issue arises because of loading the unwanted tags and navigation links presented into the website, which is literally created by the intruders or it may cause due to multiple page visits over same time. In this paper, we need to prove the effect and causes of site page surfing problems and eliminate the outliers along with the pages to cause these kinds of navigation problems. An innovative algorithm is proposed to solve these issues occurring while surfing with web pages, called Efficient Page Surfing Algorithm [EPSA]. The EPSA scheme enables the page content analysis scheme at each moment whenever the user surf for the page and produce the analysis result at every time of surfing. It will be more perfect for the users who surf for the respective details in web medium. For all with this system we need to experimentally prove our proposed algorithm is perfectly suitable to detect and eliminate the outliers presented into the web page and provide efficient surfing over large resource availability medium, which is nothing but web and web services.

Keywords: Page Surfing, Content Analysis, Outlier Detection, Page Estimation, Navigation.

I. INTRODUCTION

The web based outlier detection and efficient surfing methodology concerns the issue of anticipating conduct of web clients, in view of genuine authentic information which is of commonsense significance for some Internet-related organizations and applications. For concentrating the outlier detection scheme in E-Surfing/ Web Surfing, the following things need to be concentrated:

The following figure illustrates the preliminary tasks in web data preprocessing for web usage mining [2] and surfing.

The Page View identification usually involving these tasks, which ultimately result in a set of *n* pageviews

$$P = \{p_1, p_2, ..., pn\}$$

A group of v user transactions mentioned below:

$$T = \{t_1, t_2, ..., tv\}$$

A user transaction captures the activity of a user during a particular session. At last, one or more exchanges or sessions connected with a given client can be totaled to shape the last profile for that client

- In the event that the profile is created from a solitary session, it speaks to fleeting hobbies
- Conglomeration of various sessions results in profiles that catch long haul intrigues

[1] Ph.D Research scholar, Assistant professor, Department of computer Appliccations, J.J College of Arts and Science (Autonomous), Pudukkottai.

[2] Research Guide, Head, Assistant Professor, Department of Computer Science, Government Arts College, Pulankuruchi.

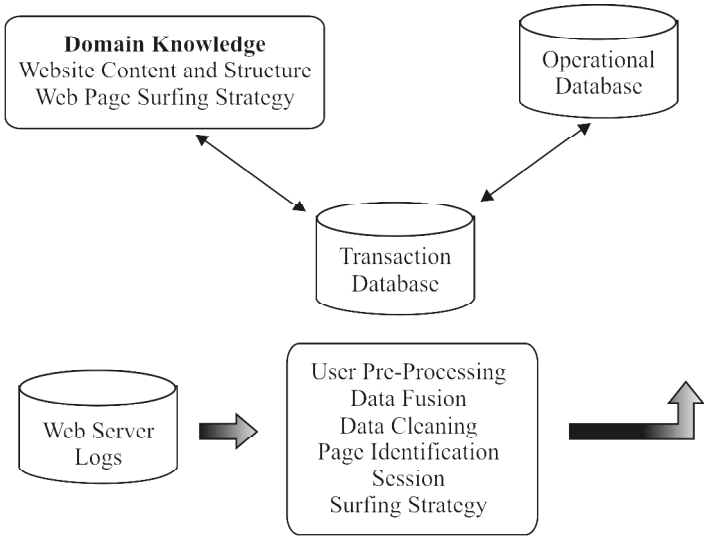

Figure 1: Preliminary tasks in web data preprocessing for web usage mining.

Web Based Data Clustering divides the information/content presented into the page into several groups such as:

- *Inter*-cluster
- *Intra*-cluster

In these above clustering [1] applications the inter-cluster similarities is minimized using the data clustering rules and the intra-cluster similarities are maximized. The web generalization process falls into the following strategies such as:

- *User-based vs. Item-based clustering*
- *Efficiency and scalability improvements*

Algorithm-1: Page Recommendation Algorithm

Input

A live/active session window *wx* = {p1, p2, p3..pn}, Minimum Threshold Value

Output

Recommendation Set SEC

Rec Set = 0;

Init Node = root; depth Value = 0;

Repeat;

depth Value++;

if Node Children < > 0

and x. item set with y. item set < = {p_1, p_2, p_3. , *pn* (depth Val)}

then Node = *X*;

else

if Node.Children = 0 then Rec Set = 1; return;

II. MARKOV WEB PAGE MINING

- Designed to predict the next user web oriented action, which is based on the user's previous web page surfing behaviour. Also the same strategy is used to discover high probability user navigational paths in a website.
- ***User Preferred Trails [UPT] :*** Various web page surfing optimization methods like goal surfing issue analyzer, page mover and scanner etc.
- Other than the Markov model some mixture models are used to attain the strategy of user profile and page surfing strategies.

The main objective of this web page mining strategy is to evaluate the accuracy and effectiveness of web personalization models.

Various measurements have been proposed in writing for assessing the strength and prescient exactness of a recommender framework: this incorporates the following procedures such as:

- Mean Absolute Error [MAE]
- Grouping Metrics [Precision and Recall]
- Beneficiary Operating Characteristic [BOC]
- The utilization of business measurements to quantify the client faithfulness and fulfillment, for example, Recency Frequency Monetary [RFM].
- The utilization of other key measurements a long-side measurements, for example, Accuracy, Coverage, Utility, Explain ability, Robustness, Scalability and User Satisfaction.

III. OUTLIER AND REDUNDANCY ELIMINATION SCHEME

With the tremendous development on the web, clients get effortlessly lost in the rich hyper structure. Subsequently creating easy to use and computerized instruments for giving important data with no redundant and outlier enabled co. This system proposes new calculation for detecting so as to mine the web content the redundant connections from the web reports utilizing set theoretical classical science, for example, subset, union, crossing point and so on. At that point the redundant connections are expelled from the first web substance to get the required data by the client.

Web information mining is the procedure of mining, extraction and integration of helpful information, data and learning from Web page substance. A percentage of the territories of doing examination in web content mining are recorded underneath:

- Formatted Information Retrieval
- Unformatted Text Retrieval
- Web Content Integration and Schema coordinating
- Building web logic State-tree
- Partitioning and Outlier Detection
- Opinion Revealing and Outlier Elimination

Existing web mining [6] algorithms do not consider documents having varying contents within the same category called web content outliers [4][5]. Unlike traditional outlier mining algorithm designed only for numeric data sets, web outliers mining algorithm should be applicable to various types of data including

text, hyper-text, image, video etc. Web pages that have different contents from the category in which they were taken constitute web content outliers. Web content outliers mining concentrates on finding outliers such as noise, irrelevant and redundant pages from the web documents. Also, web content outliers mining can be used to determine pages with entirely different contents from their parent web sites.

Sample Anomalies and Outliers

{“pswam.txt”sssd.gami, irst neuo.rd, neptone.d, neptune., private, S0, 0, normal.http, SF, 295, normal.smtp, SF, 792, normal.http, SF, 321, normal.ftp_data, SF, 3468, normal.http, SF, 227, neptune.private, S0, 0, neptune.private, S0, 0, normal.http, SF, 203, normal.http, SF, 209, normal.http, SF, 242, ipsweep.eco_i, SF, 8, neptune.private, S0, 0, neptune.private, S0, 0, smurf.ecr_i, SF, 1032, smurf.ecr_i, SF, 1032, normal.http, SF, 315, normal.http, SF, 201, normal.http, SF, 234, back.http, SF, 54540, , neptune.private, S0, 0, neptune.private, S0, 0, normal.http, SF, 203, normal.http, SF, 209, normal.http, SF, 242, ipsweep.eco_i, SF, 8, neptune.private, S0, 0, neptune.private, S0, , neptune.private, S0, 0, neptune.private, S0, 0, normal.http, SF, 203, normal.http, SF, 209, normal.http, SF, 242, ipsweep.eco_i, SF, 8, neptune.private, S0, 0, neptune.private, S0”}]}

Algorithm-2: Outlier and Redundancy Elimination

Input

Web Page, Content, Navigation Links

Outputs

Outlier and Redundancy Elimination in Pages

Used Variable

Weights[w[Tk]], penalties[p[Tk]], RedChkFlg[1], OutFlg[1]

1. Read the subject and details of the WebPage[Di] and related Contents
2. Generate n-ranges frequency profile for content learning
3. Generate n-range frequency profile for content
4. For [int i =0; m< No-Of-Doc i ++] {
5. For[int n =0; n< No-Of-Outliers ; n++{
6. IF [N-range exists in content]{ RedChkFlg=1;
7. = [“ “] i k n j e W i p Nj w Tk F N j Tk Dieight [ ] [ ] [, ,] Else
8. = [“]iknj eW i w Tk F N j Tk Dieight [ ] [OutFlg=1]

   End IF
9. } // end of loop

IV. DATASET AND METHODOLOGY

This information contains general demographic data on web clients in 2007-2010. Actually this dataset is collected from the following sources. They are: Graphics, Visualization and Usability Center, College of Computing, Atlantic Georgia Institute of Technology and GA and which is sponsored by the following University called Dr Di Cook Department of Statistics Iowa State University.

0	tcp	private	S0	0	0
0	tcp	http	SF	295	811
2	tcp	smtp	SF	792	330
0	tcp	http	SF	321	2486
0	tcp	ftp_data		SF	3468
0	tcp	http	SF	227	4073
0	tcp	private	S0	0	0
0	tcp	private	S0	0	0
0	tcp	http	SF	203	1046
0	tcp	http	SF	209	1558
0	tcp	http	SF	242	392
0	icmp	eco_i	SF	8	0
0	tcp	private	S0	0	0
0	tcp	private	S0	0	0
0	icmp	ecr_i	SF	1032	0
0	icmp	ecr_i	SF	1032	0
0	tcp	http	SF	315	1102
0	tcp	http	SF	201	7911
0	tcp	http	SF	234	11485
0	tcp	http	SF	54540	8314
0	tcp	http	SF	340	1804
0	tcp	ftp_data		SF	641
0	tcp	http	SF	338	10554
0	tcp	private	S0	0	0
0	tcp	private	S0	0	0
0	tcp	http	SF	223	753
0	tcp	http	SF	54540	8314

Figure 2: Dataset Sample

The data characteristics are defined as follows: this information originates from a study led by the Graphics and Visualization Unit at Georgia Tech. The full points of interest of the overview are accessible over the systematic procedures described above. The specific subset of the study gave here is the “general demographics” of web clients. The information have been recorded as totally numeric, with a list to the codes portrayed in the “Coding” document. The full review is accessible from the site above, alongside rundowns, tables and charts of their examinations. Also there is data on different parts of the overview, including innovation demographics and web trade.

V. EFFICIENT PAGE SURFING ALGORITHM [EPSA]

This Efficient Page Surfing Algorithm operates based on the strategy of analyzing the user who designed the web page and extracting the content of it, then partitioning the portions of the web page into the following categories such as: navigation [7] links presented into the website, title presented into the page, meta tags quoted in it, content descriptions specified into it, paragraph tags and the associated paragraphs presented into the respective webpage, < pre > that is predefined tags associated into the page and any other scripting presented into the page or not. Once the scripting and tags are analyzed, then the system comes into action for comparing the page with the defined pattern presented by the server owner. Once the pattern is exactly satisfying with the present website schema then the corresponding page is launched into the server with proper spacing, otherwise the entire site will be blocked. So that with the help of this EPSA, web user can easily search for the respective page and get the accurate as well as efficient results from the server without any delay.

Website page surfing and looking procedure Eigen value executes the Page Search and Rank calculation and to characterize noticeable status to pages in a system. It portrays the EPSA calculation as a Markov process, site page as condition of Markov chain, Link structure of web as Transitions likelihood framework of Markov chains, the answer for an Eigen vector mathematical statement and Vector emphasis power technique. It for the most part concentrate on the best way to relate the Eigen qualities and Eigen vector of Google lattice to EPSA qualities to ensure that there is a single stationary conveyance vector to which the Page Rank and Page Search calculation unites and proficiently register the positioning for expansive arrangements of website pages.

Table 1
Sample Web Data Log Sheet

Surfing IP Address	Action	Protocol	Browser	Browsed On
192.168.1.2	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.3	GET	HTTP 1.1	Opera	11/2/2015
192.168.1.4	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.6	GET	HTTP 1.1	Internet Explorer	11/2/2015
192.168.1.5	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.7	POST	HTTP 1.1	Internet Explorer	11/2/2015
192.168.1.4	POST	HTTP 1.1	Opera	11/2/2015

VI. EFFICIENT OUTLIER DETECTION

Today, much data is made week by week and the chance to catch the suppositions of the overall population about get-togethers, political developments, organization systems, promoting effort, and item inclinations has brought expanding premium both up in established researchers, for the energizing open difficulties, and in the business world, for the astounding aftermaths in online networking advertising and monetary gauge. Staying aware of the perpetually developing measure of unstructured data on the Web, be that as it may, is an impressive assignment. Not at all like standard measurable methodologies, has sentic figuring depended on a vector space model of full of feeling judgment skills information to work with regular dialect at concept level. In the proposed framework, inquiry given by the client is looked utilizing a web scrapper. Web index, opened in web scrapper, then creates a rundown of related website pages. Every site page is preprocessed by separating every one of the connections in an exceed expectations record. Presently comparing to every page, a different exceed expectations document on the circle is set. After this, each exceeds outliers record is handled by a programming code to wipe out the web structure anomalies/outliers.

Table 2
Differences in Proposed and Existing Outlier Frameworks

Techniques	*Efficiency*	*Accuracy*
Efficient Page Surfing Algorithm	93.5 %	98.1 %
Mining technique used Earlier to detect Outliers and improve web framework efficiency	85.6 %	89.2 %

VII. CONCLUSION AND FUTURE WORK

This paper presents methodology for programmed web log data mining [9] by Web Page Collection and Efficient Page Surfing Algorithm [8], which is been ended up being more viable. It remains above other web mining [3] calculations. With the mined results, the web applications is created and gives versatile client interface. Further the other kind of investigates in web applications will center later on work. It likewise

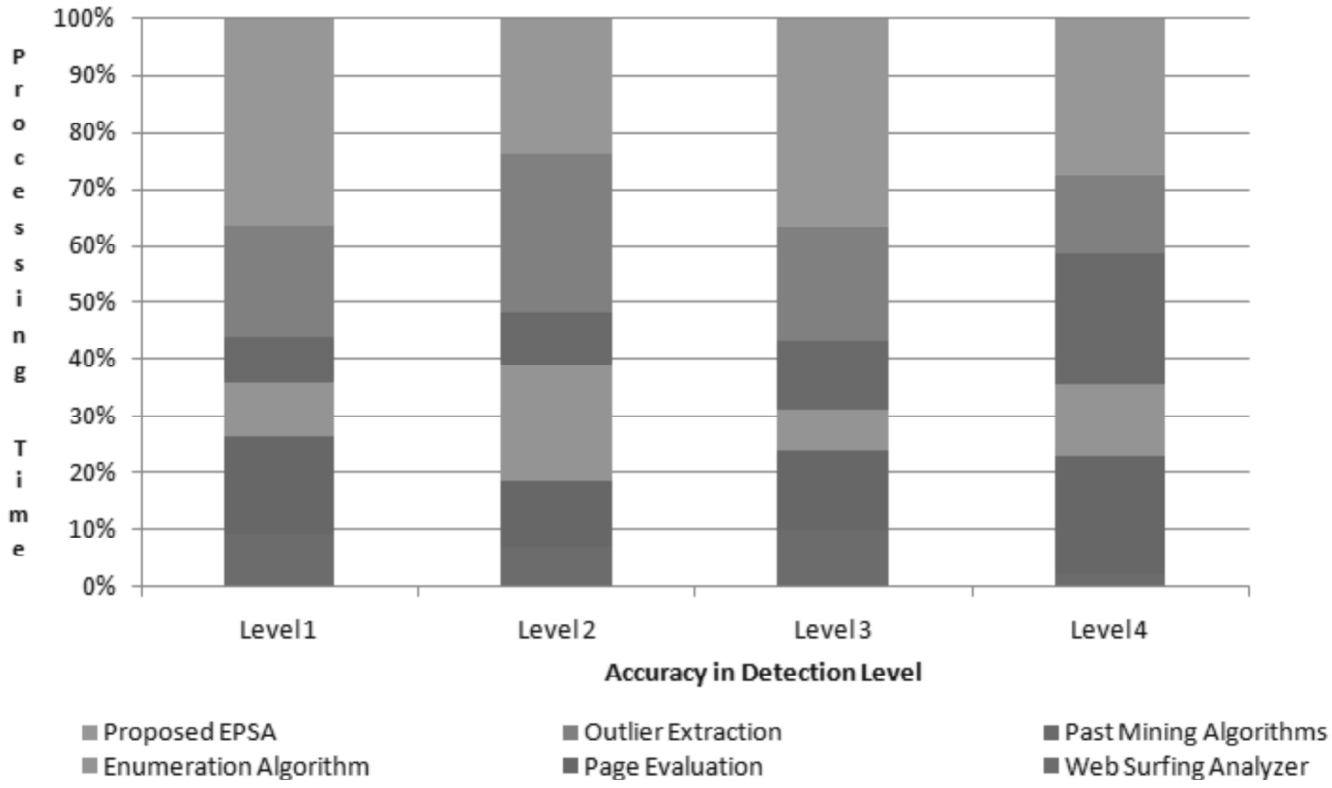

Figure 3: Outlier Detection Time and Accuracy Level for various detection methods

incorporates the data combination of substance learning and information extraction from the different sites. With adjusted EPSA calculation the structure is detailed with the assistance of other continuing calculation. Our configuration device permits trying different things with the ideas of fluffy adjust affiliation rules. It at last we broke down the fresh limit issue in the joined calculation and it is overcome by our altered affiliation Surfing fluffy calculation and the productivity is expanded in it. In future, our work will be improved to build up the upgrading look application framework.

REFERENCES

[1] s.vasuki, "Clustering Based Outlier Detection Using K-Means Strategy" *CiiT International Journal of Software Engineering*, **8**, 226-231, 2014.

[2] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava, "Web Mining: information and Pattern Discovery on the WWW"

[3] s.vasuki Dr.k.subramanian, "A new outlier detection approach to discover low hit web pages using sequential frequent pattern mining to improve website's design, *IJCSI*, **6**, 175-184,2015.

[4] Agyemang, M., Barker, K., & Alhajj R,."Framework for Mining Web Content Outliers", *ACMSAC*, 590-594, 2004.

[5] Agyemang, M. and Ezeife, "C.I. LSC-Mine: Algorithm for Mining Local Outliers". *Proceedings of the 15th Information Resource Management Association (IRMA) International Conference, New Orleans*, 2004.

[6] Lassila, O., "Web Metadata: A Matter of Semantics ", *IEEE Internet Computing* **2(4)**, 30-37, 1998.

[7] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, Umeshwar Dayal, "From user access patterns to dynamic hypertext linking", *Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems*, 1007-1014, May 1996.

[8] Jike Ge Yuhui Qiu Zuqin Chen Shiqun Yin Faculty of Computer and Information Science. Southwest University, Chongqing, "Technology of Information Push Based on Weighted Association Rules Mining".

[9] Jung, J.J., & Jo, G-S, "Semantic Outlier Analysis for Sessionizing Web Logs", *Proceeding of 14th European Conference on Machine Learning/7th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Cavtat–Dubrovnik*, 2004, 13-25.