

Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests

CH.Mahesh¹, B Bhavyasri², A Akshith Reddy², N Srilatha³, B Siddhardha⁴
^{1,2,3,4}Department of Computer Science and Engineering

^{1,2,3,4}Sree Dattha Institute of Engineering and Science, Sheriguda, Telangana

ABSTRACT:

One of the significant issues facing web users is the amount of noise in web data which hinders the process of finding useful information in relation to their dynamic interests. Current research works consider noise as any data that does not form part of the main web page and propose noise web data reduction tools which mainly focus on eliminating noise in relation to the content and layout of web data. This paper argues that not all data that form part of the main web page is of a user interest and not all noise data is actually noise to a given user. Therefore, learning of noise web data allocated to the user requests ensures not only reduction of noisiness level in a web user profile, but also a decrease in the loss of useful information hence improves the quality of a web user profile. Noise Web Data Learning (NWDL) tool/algorithm capable of learning noise web data in web user profile is proposed. The proposed work considers elimination of noise data in relation to dynamic user interest. In order to validate the performance of the proposed

work, an experimental design setup is presented. The results obtained are compared with the current algorithms applied in noise web data reduction process.

The experimental results show that the proposed work considers the dynamic change of user interest prior to elimination of noise data. The proposed work contributes towards improving the quality of a web user profile by reducing the amount of useful information eliminated as noise.

Index Terms: Noise web data, web user profile, dynamic user interest, noise reduction, NWDL algorithm, web data quality.

1.INTRODUCTION

NOWADAYS the web is widely used in every aspect of day to day life, a daily use of web means that users are searching for useful information [1]-[3]. However, ensuring useful information is available to a specific user has become a challenging issue due to the amount of noise data present on the web [4]. Noise in web data is defined as

any data that is not part of the main content of a web page [5], [6]. For example, advertisements banners, graphics, web page links from external web sites etc. Noise web data elimination is a concept which involves detection of web data that needs to be eliminated because it either does not form part of the main web page content or is not useful to a given user [7]. It is recognised in the current research work [8] that the noise web data reduction process is site-specific, i.e. it involves removal of external web pages that do not form part of the main web page content. However, this work does not focus on the structure and layout of web data to identify and eliminate noise but instead, a key focus is on extracted web log data that defines a web user profile. In view of this research, noise is not necessarily advertisements from external web pages, duplicate links and dead URLs or any data that does not form a part of the main content of a web page, but also useful information that does not reflect dynamic changes in user interests.

Various machine learning tools/algorithms are used to discover useful information from web data, this process is referred to as web usage/data mining process [1], [2]. It finds user interest patterns from web log data. Web log data contains a list of actions that

have occurred on the web based on a user [9]. These log files give an idea about what a user is interested in available web data. Web log data contain basic information such as IP address, user visit duration and visiting path, web page visited by the user, time spent on each web page visit etc. In this work, web log file and web data are used interchangeably because a log file contains web data, therefore elimination of noise web data is based on extracted web user log file.

In a real world, it is practically impossible to extract web log data and create a web user profile free from noise data. A web user profile is defined as a description of user interests, characteristics, and preferences on a given website [10]-[12]. User interests can be implicit or explicit [13]. Explicit interests are where a user tell the system what his/her interests are and what they think about available web data while implicit interest is where the system automatically finds interests of a user through various means such as time and frequency of web page visits [14], [15]. Many users may not be willing to tell the system what their true intentions are on available web data, therefore, this work will focus on implicit user interests.

Current research efforts in noise web data reduction have worked with the assumption

that the web data is static [16]. For example, [17], [18] proposed a mechanism where noise detected from web pages is matched by stored noise data for classification and subsequent elimination. Therefore, it shows that elimination of noise in web data is based on pre-existing noise data patterns. In evolving web data, existing noise data patterns used to identify and eliminate noise from web data may become out of date. For this reason, the dynamic aspects of user interest have recently become important [19], [20]. Moreover, web access patterns are dynamic not only due to evolving web data but also due to changes in user interests [21]. For example, web users are likely to be interested in data derived from events such as Weddings, Christmas, Birthdays etc. Therefore, it is necessary to discover where such dynamic tendencies impact the process of eliminating noise from web data.

research proposes a machine learning algorithm capable of learning noise in web data prior to elimination. The proposed algorithm considers the dynamic change in user interests and evolving web data to identify and learn noise data. The main novelties of this research are:

To demonstrate how dynamic user interests and evolving web data impact noise web

data reduction process. This takes into account contribution made by current research works and their limitations in relation to the current state of the art. □ To propose a machine learning algorithm capable of learning noise in a web user profile prior to elimination. Elimination of noise from a web user profile does not only depend on pre-existing noise data patterns, but it learns noise levels based on dynamic changes in user interest as well as evolving web data. □ The outcome of the practical application of the proposed tool will reduce the amount of useful information eliminated as noise from a web user profile. This may significantly improve the quality of a web user profile.

2.CURRENT RESEARCH WORK

Current tools developed to identify and eliminate noise from web pages are mainly based on the visual layout of web pages. For example, [5] proposed Site Style Tree (SST) to detect and eliminate noise data from web pages. SST is based on an observation that the main web page content usually shares the same presentation style and any other page with different presentation style is considered as noise. To eliminate noise from web pages, SST simply maps the page to the main web page to determine if the page is useful or noise based on its presentation

style. Another noise web data reduction tool that focuses on web page layout is Pattern Tree algorithm [22], it is based on Document Object Model (DOM) tree concept with an assumption that data present on the web can be considered noise if its pattern is dissimilar from the main content of web page. Least Recently Used paging algorithm (LRU) [23] is also used to detect and remove noise from web pages. LRU takes into account visual and non-visual characteristics of a web page and is able to remove noise web data for example news, blogs and discussions. LRU algorithm determines pages that have been frequently visited and those pages that have not been visited over a long period of time. However, this work does not focus on structure and layout of web data to identify and eliminate noise but instead, a key focus is on extracted web log data that defines a web user profile. Based on issues addressed in the previous section, noise in web data should be identified and eliminated taking into account user interest levels on web data. Current research works have applied existing machine learning tools/algorithms to find user interest data and eliminate noise data from extracted web data logs. For example, [17] used Case based reasoning (CBR) and Neural Network to

eliminate noise from web log data. CBR is a machine learning approach which makes use of past experiences to solve future problems, i.e. it detects noise from web pages using existing stored noise data. Different noise patterns in websites are stored in form of DOM tree, the case base is then searched for similar existing noise patterns. Artificial Neural Network is used to match existing noise patterns stored in Case-Based. Even though this approach is based on the idea of case based reasoning to identify noise data by matching existing noise patterns stored in case-based, it is difficult to determine if such information is relevant or noise to a user despite the fact that it matches with existing patterns. This is because web data is dynamic and so is expected user interest, if the usefulness of data is determined using case based approach then the output will be misleading. kNN applied by [24] also used existing noise data to identify and eliminate noise in web pages. Their main focus was on local noise for example advertisements, banners, navigational links etc. Web log data was extracted and surveyed to which server they belong. If the address belongs to a list of already defined advertisement server, then the link is removed.

Due to the dynamic nature of user interests

as well as evolving web data, existing noise data patterns may become out of date and hence difficult to identify and eliminate noise from a web user profile. To determine user interest levels on extracted web log data, [25] used the Naïve Bayesian classification algorithm. Their main objective was to classify extracted web data logs and study its usefulness based on user interests. The initial phase involved removing noise data such as advertisement banners, images and screen savers from extracted web data logs. They used Naïve Bayesian classification model to classify useful and noise data based on a number of pages viewed and time taken on a specific page. However, spending more time on a web page may not necessarily mean a user is interested. If a user is struggling to find information of interest, he/she may spend more time searching. Weighted Association Rule Mining was also used by [26] to extract useful information from web log data. Their objective was to find web pages visited by a user and assign weights based on interest level. The weight of a web page to a user interest is estimated with the frequency of page visit and a number of pages visited. Where pages visited only once by only one user, they will be assigned low weights and subsequently considered noise.

interest in a broad range of information based on time and what is happening around the world. Therefore, user interests can be dynamic as the web evolves. Our justification for this claim is that if noise in web data is not clearly defined and analysed through learning, the purpose and use of data extracted will be compromised. Learning of noise in web data is influenced by the activities of a user on web data which is determined by measures such as time duration, the frequency of visits and the depth of a user visit on a given web page. These measures will influence usefulness of a web page to a user rather than the relationship among web data on a given website.

3.EXISTING SYSTEM

Now-a-days almost all users are using web pages to get various information such as news, sports, technology etc but all web pages will use noise data such as images, video clips or advertisement which makes difficult for the users to get interested information. To remove noise data all existing technologies were using static web matching pattern such as the main page look and feel will be match with rest of the screen and if not match then it will remove unmatched data from the web pages to show only interested data to the user. This static

technique will not work if web pages look and feel changes dynamically.

4. PROPOSED SYSTEM:

To overcome from above issue author is proposing Noise Web Data Learning (NWDL) technique, in this technique server will maintain log for each user access page and will be called as web log dataset. This dataset will have information such as User_id, access_page, date_time, URL. By analyzing such log data we can identify user interested pages in dynamic or static web pages. User interested pages can be found by seeing frequency of web page access by a single user and total time spend on each page.

If user spend more time and access this page more than 2 times then we can consider that user is interested in that page. If user spend less time on seeing that page and visiting that page very rarely then it will consider as uninterested page and will be called as noise page.

5. SYSTEM ARCHITECTURE

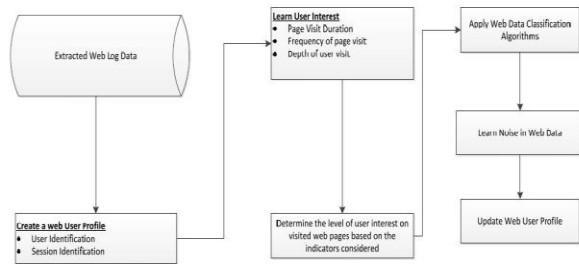


Figure.1 System Architecture

6. IMPLEMENTATION:

MODULES:

1. Add Product Details

To build project I used some sample products image to train product identification models

2. Train Model

In this Module screen train model generated with 100% accuracy and now show product to web cam.

3. Add/Remove Product from basket

To allow application to identify product image and then show in text area and if we again show same product then application will remove from text area

7. RESULTS

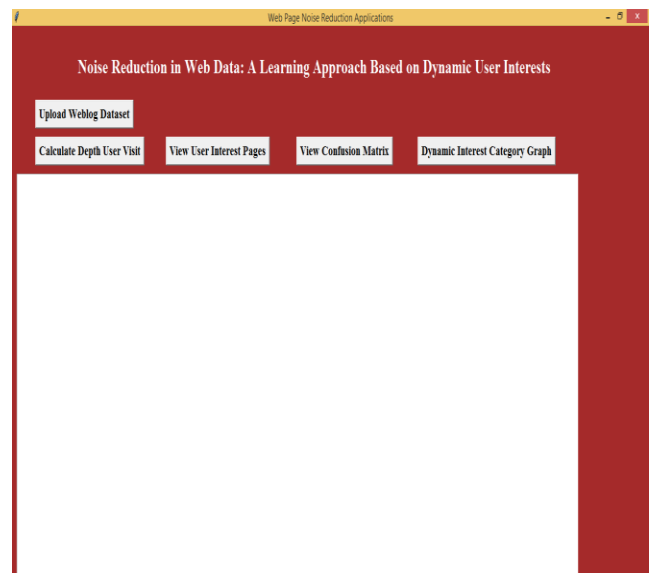


Figure.2 Home Screen

In above screen click on 'Upload Weblog Dataset' button to upload dataset

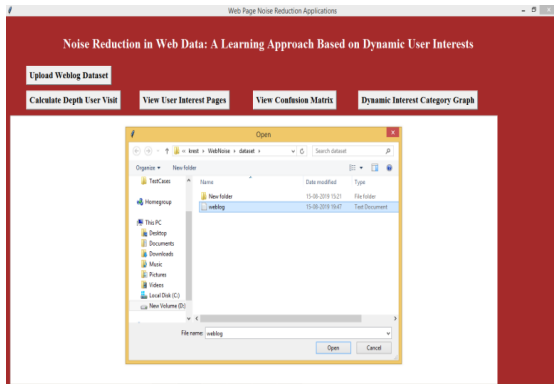


Figure.3 Upload Dataset Screen

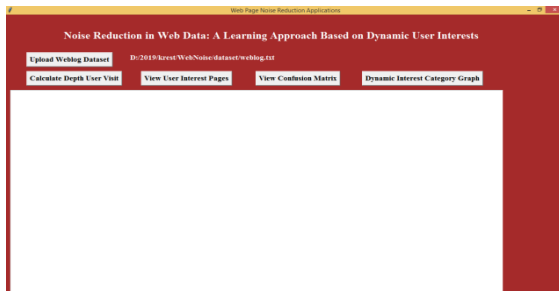


Figure.4 Calculate Depth User Visit Screen

After dataset upload click on ‘Calculate Depth User Visit’ button to calculate frequency and weight of each page visit by single users

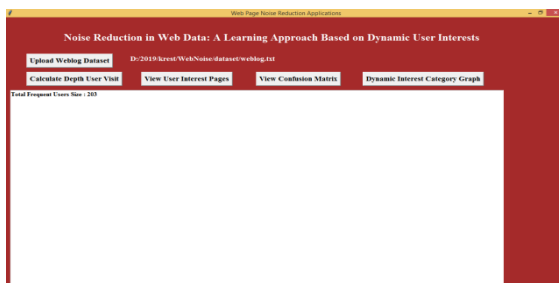


Figure.5 User Frequently Screen

In above screen we can see total 203 web pages which access more frequently. To see frequency of each access page see command prompt console. See below screen

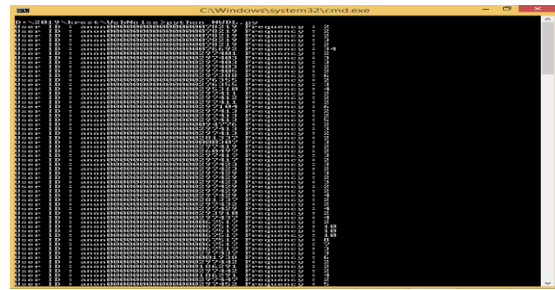


Figure.6 Console Screen

In above screen we can see user id and frequency of each page access by them, if u wants to see page name and weight details then copy one user id from black console. See below screen

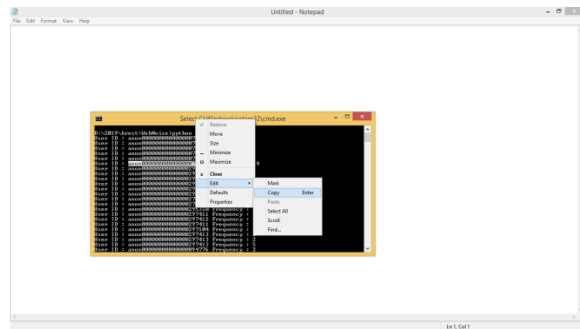


Figure.7 Console Screen

In above screen from command prompt i am copying one user id who access web pages frequently. Now in application click on ‘View User Interest Pages’ button to get below screen

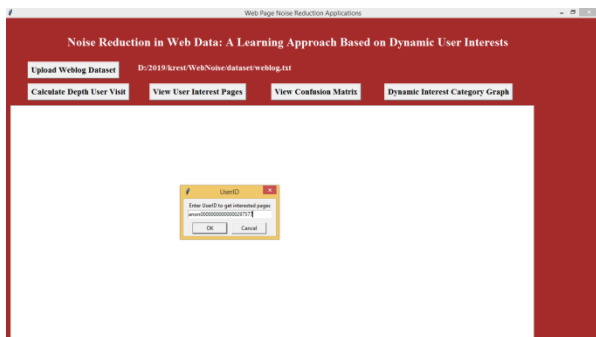


Figure.8 View User Interest Pages Screen

In above screen whatever user id i selected from command prompt i pasted in dialog box. Now click on ok button to get all frequent access pages

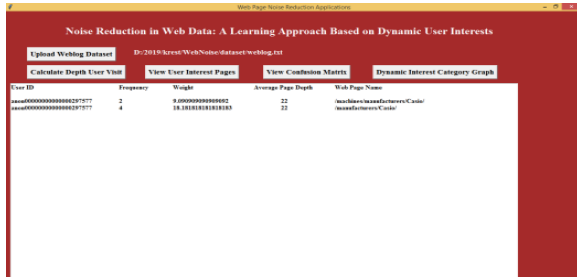


Figure.9 User Id Details Screen

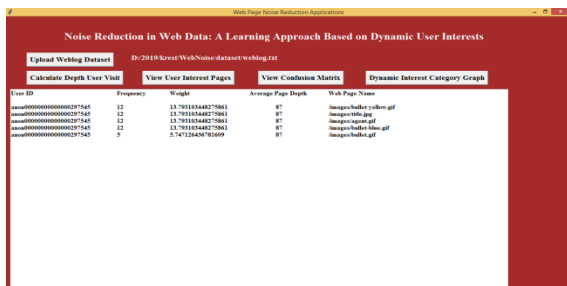


Figure.10 Web Page Name Screen

In above screens we can see user interest pages in the 'Web Page Name' Column. Above are the web pages details which are access by this user more frequently and will be added to user interested list. Now click on 'View Confusion Matrix' button to know no of interested and noise pages obtained by propose NWDL and existing SVM technique

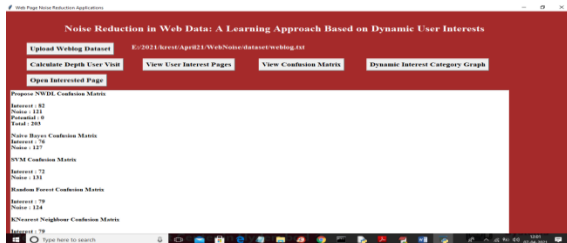


Figure.11 SVM Screen

In above screen we can see propose NWDL predict more interested pages compare to existing algorithm like SVM, Naïve Bayes, Random forest and KNN. In above screen with NWDL we got 82% predicted interested pages and naïve bayes we got 72% predicted interested page. Scroll down above text area to view all details of KNN algorithm. Now click on 'Dynamic Interest Category Graphs' button to know various categories web pages in the form of graph

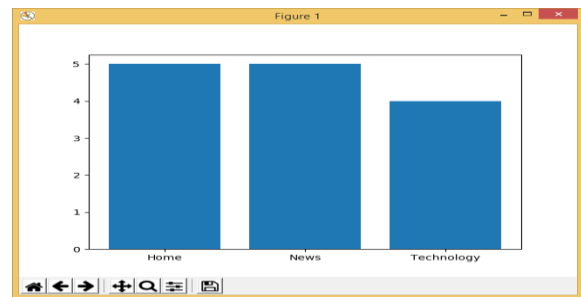


Figure.12 Dynamic Interest Category Graphs

In above screen x-axis represents categories such as Home, News and technology and y-axis represents total count of those categories. Like this in dataset many categories are available but i am showing only 3 categories. Now if user wants to open any interested page then he has to click on 'View User Interest Pages' again to get below screen

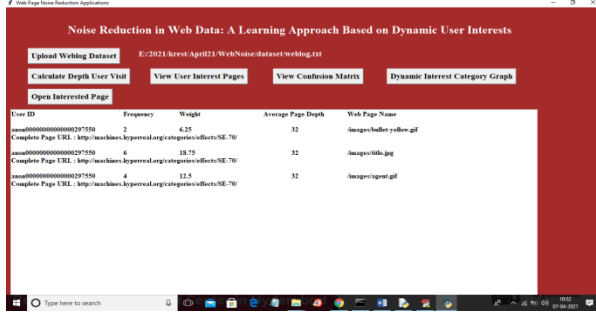


Figure.13 View User Interest Pages

In above screen user can select any URL like below screen and then click on ‘Open Interested Page’ button

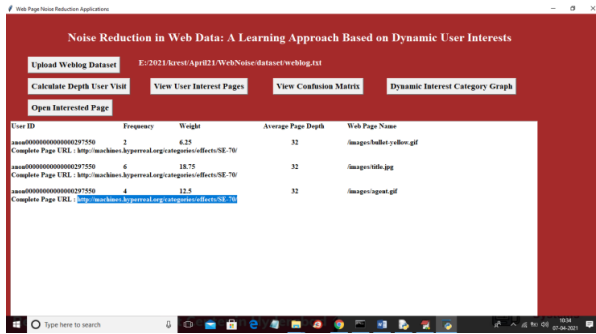


Figure.14 Open Interested Page

In above screen I selected one URL and then click on ‘Open Interested Page’ button to get below screen

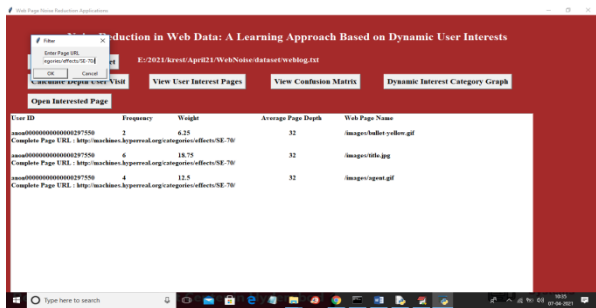


Figure.15 Open Interested Page

In above screen dialog box paste that URL and click OK button to view that page in browser like below screen

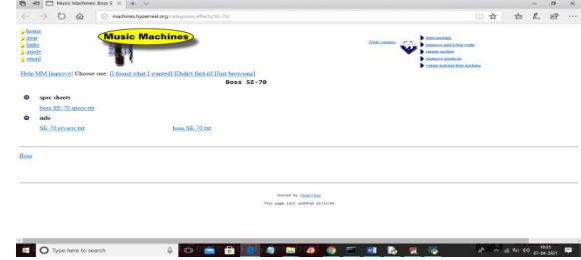


Figure.15 Browser Screen

8.CONCLUSION

A machine learning algorithm capable of learning noise in web data prior to elimination is proposed. The starting point of this paper defines and identifies challenges with current research work in the noise web data reduction process. For example, elimination of noise in web data is based on preexisting noise data patterns and when user interests change, the stored noise data patterns can longer be relied, and hence not relevant. Moreover, current research works consider noise as any data that does not form part of the main web page. Therefore, it is difficult to identify and eliminate noise in web data without taking into dynamic interests of a web user.

This paper undertakes various steps to address the identified problems. Firstly, a machine learning algorithm that considers dynamic changes in user interests by learning the depth of a user visit in a specific web page is presented. Secondly, an algorithm that learns noise web data taking into account changes in user interests and

evolving web data. The proposed algorithm is able to identify what users are interested in a given time, how they are searching and if they are interested in what they searching prior to elimination. Finally, the proposed tool contributes towards improving the quality of a web user profile.

9. REFERENCES

- [1] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, ‘Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data’, SIGKDD Explor Newsl, vol. 1, no. 2, pp. 12–23, Jan. 2000.
- [2] M. Jafari, F. SoleymaniSabzchi, and S. Jamali, ‘Extracting Users’ Navigational Behavior from Web Log Data: a Survey’, J. Comput. Sci. Appl. J. Comput. Sci. Appl., vol. 1, no. 3, pp. 39–45, Jan. 2013.
- [3] N. Soni and P. K. Verma, ‘A Survey On Web Log Mining And Pattern Prediction’, Int. J. Adv. Technol. Eng. Sci.-2348-7550.
- [4] T. R. Ramesh and C. Kavitha, ‘Web user interest prediction framework based on user behavior for dynamic websites’, Life Sci. J., vol. 10, no. 2, pp. 1736–1739, 2013.
- [5] L. Yi, B. Liu, and X. Li, ‘Eliminating Noisy Information in Web Pages for Data Mining’, in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2003, pp. 296–305.
- [6] A. Dutta, S. Paria, T. Golui, and D. K. Kole, ‘Structural analysis and regular expressions based noise elimination from web pages for web content mining’, in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 1445–1451.
- [7] G. D. S. Jayakumar and B. J. Thomas, ‘A new procedure of clustering based on multivariate outlier detection’, J. Data Sci., vol. 11, no. 1, pp. 69–84, 2013.
- [8] V. Chitraa and A. S. Thanamani, ‘Web Log Data Analysis by Enhanced Fuzzy C Means Clustering’, Int. J. Comput. Sci. Appl., vol. 4, no. 2, pp. 81–95, Apr. 2014.
- [9] L. K. Joshila Grace, V. Maheswari, and D. Nagamalai, ‘Analysis of Web Logs And Web User In Web Mining’, Int. J. Netw. Secur. Its Appl., vol. 3, no. 1, pp. 99–110, Jan. 2011.
- [10] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, ‘User profiles for personalized information access’, in The adaptive web, Springer, 2007, pp. 54–89.
- [11] P. Peñas, R. del Hoyo, J. Veá-Murguía, C. González, and S. Mayo, ‘Collective Knowledge Ontology User Profiling for Twitter – Automatic User Profiling’, in 2013 IEEE/WIC/ACM International Joint

Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, vol. 1, pp. 439–444.

[12] S. Kanoje, S. Girase, and D. Mukhopadhyay, ‘User profiling trends, techniques and applications’, ArXiv Prepr. ArXiv150307474, 2015.

[13] H. Kim and P. K. Chan, ‘Implicit indicators for interesting web pages’, 2005.

[14] J. Xiao, Y. Zhang, X. Jia, and T. Li, ‘Measuring similarity of interests for clustering Web-users’, in Proceedings 12th Australasian Database Conference. ADC 2001, 2001, pp. 107–114.

[15] H. Liu and V. Kešelj, ‘Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users’ Future Requests’, *Data Knowl Eng*, vol. 61, no. 2, pp. 304–330, May 2007.

[16] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain, ‘A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites’, *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 2, pp. 202–215, Feb. 2008.

[17] T. Htwe and N. S. M. Kham, ‘Extracting data region in web page by removing noise using DOM and neural network’, in 3rd International Conference on Information and Financial Engineering,

2011.

[18] R. P. Velloso and C. F. Dorneles, ‘Automatic Web Page Segmentation and Noise Removal for Structured Extraction using Tag Path Sequences’, *J. Inf. Data Manag.*, vol. 4, no. 3, p. 173, Sep. 2013.

[19] Y. L. Sulastri, A. B. Ek, and L. L. Hakim, ‘Developing Students’ Interest by Using Weblog Learning’, *GSTF Int. J. Educ.* Vol1 No2, vol. 1, no. 2, Nov. 2013.

[20] A. Nanda, R. Omanwar, and B. Deshpande, ‘Implicitly Learning a User Interest Profile for Personalization of Web Search Using Collaborative Filtering’, in 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014, vol. 2, pp. 54–62.

[21] J. Onyancha, V. Plekhanova, and D. Nelson, ‘Noise Web Data Learning from a Web User Profile: Position Paper’, in Proceedings of the World Congress on Engineering, 2017, vol. 2.

[22] N. Narwal, ‘Improving web data extraction by noise removal’, in Fifth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2013), 2013, pp. 388–395.

[23] A. Garg and B. Kaur, ‘Enhancing Performance of Web Page by