International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

NOISE REDUCTION IN WEB DATA A LEARNING APPROACH BASED ON DYNAMIC USER INTERESTS

[1]**Kushalvenkatchowdary Mikkilineni,** SRM UNIVERSITY AP, kushalchow2222@gmail.com
[2]**Prashanth Pusuluri**, Guru Nanak Institutions Technical Campus,
pusuluriprashanth4239@gmail.com

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.

Keywords *– Web log data, web user profile, user interest, noise web data learning, machine learning.*

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. 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. Noise in web data is defined as any data that is not part of the main content of a web page. 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. 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.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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.

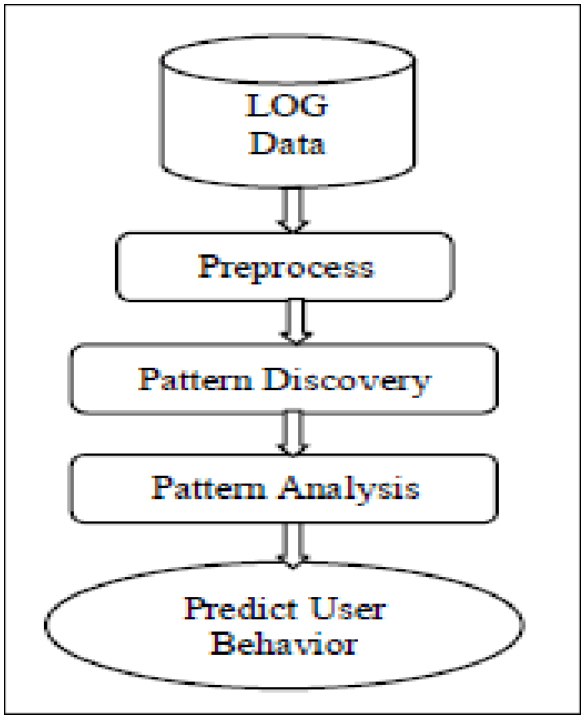

Fig.1: Example figure

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. 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. 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. User interests can be implicit or explicit. 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. 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. For example, 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. Moreover, web access patterns are dynamic not only due to evolving web data but also due to changes in user interests. For example, web users are likely to be interested in data derived from

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

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.

2. LITERATURE REVIEW

Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data:

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. This paper describes each of these phases in detail. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the research and practice communities. This paper provides a detailed taxonomy of the work in this area, including research efforts as well as commercial offerings. An up-to-date survey of the existing work is also provided. Finally, a brief overview of the WebSIFT system as an example of a prototypical Web usage mining system is given.

Extracting Users'Navigational Behavior from Web Log Data: a Survey:

Web Usage Mining (WUM) is a kind of data mining method that can be used to discover user access patterns from Web log data. A lot of research has been done already about this area and the obtained results are used in different applications such as recommending the Web usage patterns, personalization, system improvement and business intelligence. WUM includes three phases that are called preprocessing, pattern discovery and pattern analysis. There are different techniques for WUM that have their own advantages and disadvantages. This paper presents a survey on some of the existing WUM techniques and it is shown that how WUM can be applied to Web server logs.

A Survey On Web Log Mining And Pattern Prediction:

Web sites have abundant web usage log which provides great source of knowledge that can be used for discovery and analysis of user accessibility pattern. The web log mining is the process of identifying browsing patterns by analyzing the user's navigational behaviour. The web log files which store the information about the visitors of web sites is used as input for web log mining and pattern prediction process. First these log files are pre-processed and converted into required formats so web usage mining techniques can apply on these web logs for frequent patterns. The obtained results can be used in different applications like modification of web sites, system improvement, business intelligence, and personalization etc.

Web user interest prediction framework based on user behavior for dynamic websites:

We develop a framework to predict the user interest based on the behavior of user to increase the efficiency of dynamic websites. The content management in the dynamic website is difficult because it varies with the user profiles, i.e. different contents have to

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

be placed for different users according to the user profiles. Various ways have been identified earlier to track the user interest but lacks with the accuracy here we propose a new one which composes both implicit and explicit. We track all behaviors like time of visit, navigation url, web logs, user actions on the web page. Our model uses the web log data of the user and also tracks the implicit behaviors performed by the user. The tracked information are used to identify the user interest and The web users are clusters based on the identified interest which is used by the dynamic websites. The dynamic websites administrator could use the outcome of the cluster for various purposes.

Eliminating Noisy Information in Web Pages for Data Mining:

A commercial Web page typically contains many information blocks. Apart from the main content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements (for business purposes and for easy user access). We call these blocks that are not the main content blocks of the page the noisy blocks. We show that the information contained in these noisy blocks can seriously harm Web data mining. Eliminating these noises is thus of great importance. In this paper, we propose a noise elimination technique based on the following observation: In a given Web site, noisy blocks usually share some common contents and presentation styles, while the main content blocks of the pages are often diverse in their actual contents and/or presentation styles. Based on this observation, we propose a tree structure, called Style Tree, to capture the common presentation styles and the actual contents of the pages in a given Web site. By sampling the pages of the site, a Style Tree can be built for the site, which we call the Site Style Tree (SST). We then introduce an information based measure to determine which parts of the SST represent noises and which parts represent the main contents of the site. The SST is employed to detect and eliminate noises in any Web page of the site by mapping this page to the SST. The proposed technique is evaluated with two data mining tasks, Web page clustering and classification. Experimental results show that our noise elimination technique is able to improve the mining results significantly.

3. METHODOLOGY

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. 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.

Current research efforts in noise web data

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

reduction have worked with the assumption that the web data is static. For example, 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. Moreover, web access patterns are dynamic not only due to evolving web data but also due to changes in user interests. 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.

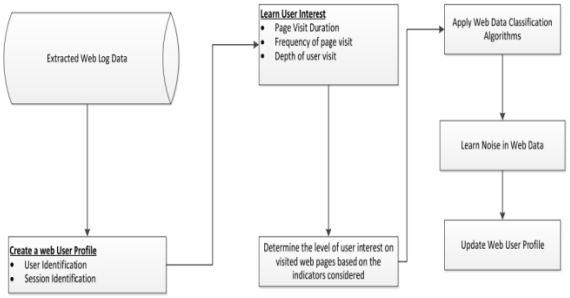

Fig.2: System architecture

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.

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.

Dataset Example

User_id access_page date_time

1 abcd.html 2019-01-22 11:00:12

2 xyz.html 2019-01-22 12:18:23

1 abcd.html 2019-01-22 11:05:18

1 abcd.html 2019-01-22 11:06:12

1 xyz.html 2019-01-22 12:22:23

From above web log dataset we can easily says that user 1 accessing abcd.html more no of time and its frequency is 3 and he spend almost 6 minutes on that page (spend time

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

will be calculated from first same page visit to till last same page visit) and this abcd.html will be consider as interested page from user 1 and xyz.html is rarely access by that user and will be consider as noise page and will not recommend to user.

To perform experiment author has used WEBLOG dataset and i am also using same

4. IMPLEMENTATION

CNN Algorithm:

CNN is a type of deep learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of animal visual cortex and designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. CNN is a mathematical construct that is typically composed of three types of layers (or building blocks): convolution, pooling, and fully connected layers. The first two, convolution and pooling layers, perform feature extraction, whereas the third, a fully connected layer, maps the extracted features into final output, such as classification. A convolution layer plays a key role in CNN, which is composed of a stack of mathematical operations, such as convolution, a specialized type of linear operation. In digital images, pixel values are stored in a two-dimensional (2D) grid, i.e., an array of numbers (Fig. 2), and a small grid of parameters called kernel, an optimizable feature extractor, is applied at each image position, which makes CNNs highly efficient for image processing, since a feature may occur anywhere in the image. As one layer feeds its output into the next layer, extracted features can hierarchically and progressively become more complex. The process of optimizing parameters such as kernels is called training, which is performed so as to minimize the difference between outputs and ground truth labels through an optimization algorithm called backpropagation and gradient descent, among others.

SVM Algorithm

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

- **Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.

You need to remember a thumb rule to identify the right hyper-plane: "Select the hyper-plane which segregates the two classes better". In this scenario, hyper-plane "B" has excellently performed this job.

- **Identify the right hyper-plane (Scenario-2):** Here, we have three hyper-planes (A, B and C) and all are

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

segregating the classes well. Now, How can we identify the right hyper-plane?

Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let's look at the below snapshot:

Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

K-Means Clustering Algorithm:

The K -means algorithm (Lloyd, 1957) is the most popular of partitioning algorithms. It seeks to find K clusters that minimize the sum of squared Euclidean distances between each observation and its respective cluster mean. In its simplest form, the K -means algorithm iteratively alternates between two steps: (1) for a given set of cluster centers, assign each observation to the cluster with the nearest center, and (2) for a given assignment of observations to clusters, update each cluster center as the sample mean of all points in that cluster. Initial center values for Step 1 are often a random sample of K observations. It typically converges to one of the many local optima, rather than the global optimum. Hartigan and Wong (1979) give a more complicated algorithm which is more likely to find a good local optimum. Whatever algorithm is used, it is advisable to repeatedly start the algorithm with different initial values, increasing the chance that a good local optimum is found.

The K-Means algorithm is one of the most widely used techniques for clustering. It starts by initializing the k cluster centers, where k is preliminarily determined. Then, each object (input vector) of the dataset is assigned to the cluster whose center is the nearest. The mean (centroid) of each cluster is then computed so as to update the cluster center . This update occurs as a result of change in the membership of each cluster. The processes of reassigning the objects and the update of the cluster centers is repeated until no more change is the value of any of the cluster centers.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect. K-means is a simple algorithm that has been adapted to many problem domains.

5. EXPERIMENTAL RESULTS

Fig.5: Home screen

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

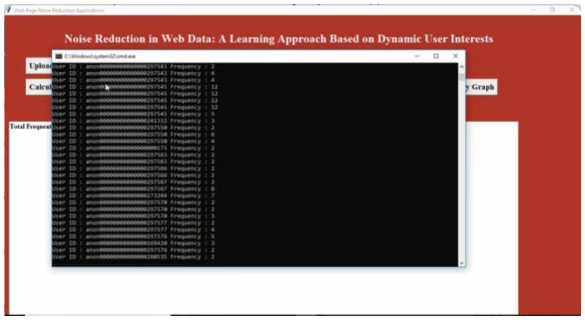

Fig.6: Calculate depth user visit

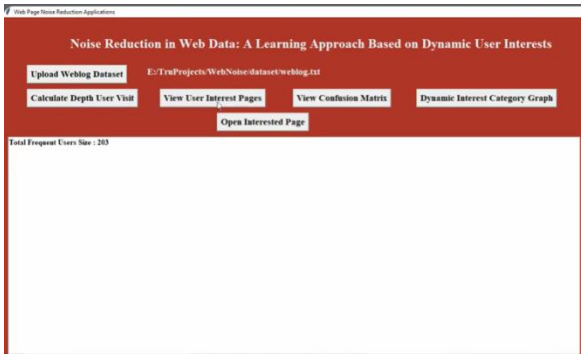

Fig.7: Total user count

Fig.8: User input

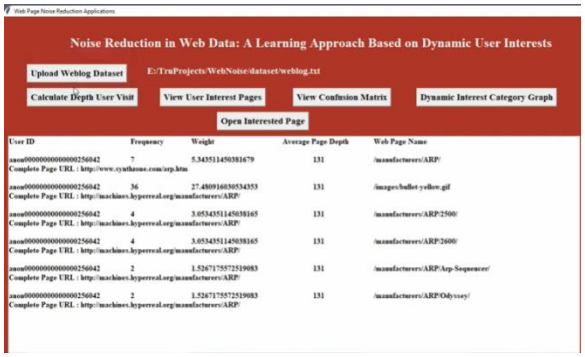

Fig.8: Prediction result

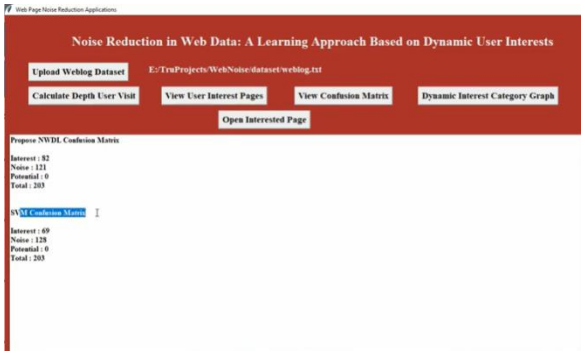

Fig.9: Confusion matrix

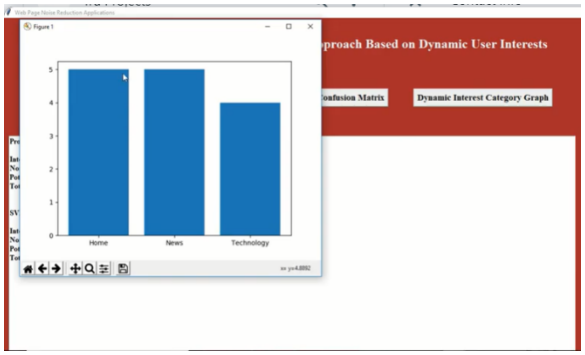

Fig.10: Graph

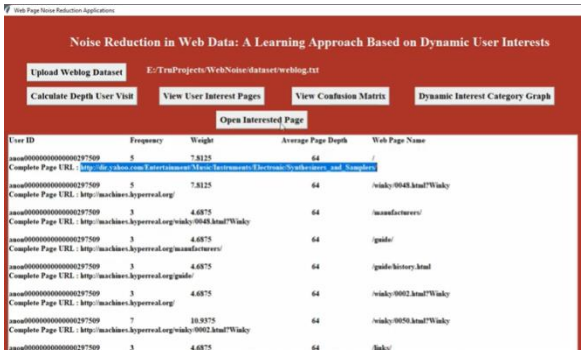

Fig.11: Open interested page

6. 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.

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. Vea-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.