Journal of Advanced Zoology

ISSN: 0253-7214

*Volume **44** Issue **S-5** **Year 2023** Page **3185-3196***

Noise Reduction In Web Data: A Learning Approach Based On Dynamic User Interests

G.Prathibha Priyadarshini[1*], G.Shaheen Firdous[2], N Lakshmi Aparna[3], Shaik Niyosha Riyazy[4]

[1]Assistant Professor, Department of CSE, Ravindra College of Engineering for Women*
[2]Assistant Professor, Department of CSE, Ravindra College of Engineering for Women
[3]Student, Ravindra College of Engineering for Women
[4]Student, Ravindra College of Engineering for Women

***Corresponding Author:** G.Prathibha Priyadarshini*
**Assistant Professor, Department of CSE, Ravindra College of Engineering for Women*

Article History

Received:
Revised:
Accepted:

Abstract

One of the prominent challenges internet operators encounter is the abundance of extraneous material inside web content, hence impeding the efficient retrieval of relevant information aligned with their evolving interests. The present state of affairs. In academic research, noise is commonly defined as any extraneous data that does not contribute to the intended analysis or study objectives. This study aims to analyse the primary webpage and suggest noise reduction tools for online data. The primary emphasis is on the reduction of noise about the content and its associated factors. The arrangement or organisation of data on the internet. In this paper, including some data inside a dataset may not be universally applicable or appropriate. The web page's primary content pertains to the user's specific interests, while extraneous info is minimised. Noise can be perceived as disruptive or unwanted sound by an individual. Hence, the acquisition of noisy online data and the allocation of resources to user requests ensure not just a decrease in noise levels. There is an observed correlation between the level indicated in a user profile on the web and a reduction in the occurrence of valuable information loss. The inclusion of information consequently enhances the calibre of an online user profile. The phenomenon of noise refers to unwanted or disruptive sounds that can have negative effects on individuals and the Web Data Learning (NWDL) tool/algorithm exhibits the capacity to acquire knowledge. The proposal suggests the use of noise in web user profiles to enhance data privacy. The work that has suggested the removal of noise data in the context of dynamic user behaviour is being considered. The topic of interest is being discussed. To ascertain the efficacy of the proposed study, A presentation of an experimental design arrangement is provided. The results were achieved in contrast to the presently employed algorithms utilised in the context of noisy online data. The reducing process. The experimental findings indicate that the proposed study examines the dynamic evolution of user interest before the removal of extraneous data. The proposed study makes a significant contribution to Enhancing the calibre of an online user profile through

CC License CC-BY-NC-SA 4.0	the reduction of content volume. The elimination of noise results in the removal of beneficial information.

Introduction

The internet is now widely used in all facets of daily life, and users are frequently looking for beneficial information when using it. However, the task of providing relevant and valuable information to a particular user has grown increasingly difficult because of the abundance of irrelevant and low-quality material found on the internet.

The term "noise" in the context of web data refers to any extraneous information that is not considered integral to the primary text on a website. For instance, many forms of visual communication, such as advertisement banners, images, and web page links, originate from external websites. The concept of noise web data elimination entails the identification of online data that needs to be removed because it either does not contribute to the primary content of a website or is not relevant to a particular user.

The present body of research acknowledges that the process of the noise in web data is distinct for each website. This procedure includes removing external websites that are not integral to the core content of the webpage [8]. Nevertheless, this study does not order the analysis by priority of the organisation and presentation of web data to identify and remove irrelevant information. Instead, it places significant emphasis on the examination of extracted web log data, which is used to establish a comprehensive profile of web users.

Based on the findings of this study, noise on web pages includes more than just advertisements from external sources, duplicate links, dead URLs, or any data that is not integral to the primary content. It also includes valuable information that does not accurately represent the dynamic shifts in user interests.
The practice of extracting valuable information from web data is known as web usage or data mining and involves the use of various machine-learning techniques and algorithms. This algorithm identifies and analyses user interest patterns by examining web log data. A weblog, sometimes referred to as the term "weblog" or "blog" digital record that encompasses a compilation of activities that have transpired on the internet, specifically about a particular user. The log files provide insight into the user's interests regarding the accessible online data. A weblog, often known as the term "weblog" or "blog" a digital record that contains fundamental details, such as the Internet Protocol (IP) address of a user, the duration of their visit, the sequence of web pages they accessed, the specific web pages they visited, and the amount of time they spent on each web page. Because a log file comprises web data, the terms web log file and web data are used synonymously in this work. As a result, the extraction of web user log files is used to eliminate noise from web data.

In practical terms, it is highly challenging to extract web log data and generate an online user profile that is completely devoid of extraneous data. An online user outline is a comprehensive depiction of a user's interests, attributes, and preferences within the context of a particular website. User interests might manifest in either implicit or explicit forms. Explicit interests refer to the situation where a user explicitly communicates their preferences and opinions regarding available web data to the system. Contrariwise, implicit interests involve the system automatically inferring a user's interests using methods such as analysing the frequency and duration of their visits to web pages. A significant number of users may exhibit reluctance to disclose their genuine intentions to the system regarding the utilisation of web data. Consequently, this study will primarily concentrate on implicit user interests. Contemporary research endeavours in the field of web data noise reduction have primarily operated under the premise that the web data being analysed remains static. For instance, a proposed approach involves the detection of noise from online pages, which is then compared to pre-existing noise data for classification and subsequent removal.

Hence, it can be observed that the method of reducing noise in online data relies on the identification and utilisation of pre-existing noisy data patterns. In the context of dynamic web data, it is possible for the noisy data patterns that are currently employed to detect and remove data to become outdated over time. As a result, the dynamic elements of user interest have gained significance in recent times. Furthermore, the patterns of web access exhibit dynamism, which can be attributed not only to the continuous evolution of web data but also to fluctuations in user interests. For instance, individuals who use the Internet are inclined to exhibit interest in the information that is obtained from Occasions such as weddings, Christmas, birthdays, and similar occasions. Hence, it is essential to ascertain the particular regions where these dynamic tendencies influence the procedure of noise reduction in online data.

Assuming that web data is static, recent research into noise reduction in web data has been conducted [16]. As an illustration, [17] and [18] described a technique in which noise identified from web pages is matched with stored noise data for categorization and subsequent eradication. Hence, it can be inferred that the method of reducing noise in online data relies on the identification and utilisation of pre-existing patterns of noisy data. existing patterns in the noise data that are accustomed to detecting and removing web data noise could become outdated as the web data changes. Hence, the significance of the dynamic elements of user interest has gained prominence in recent times (19, 20).

Additionally, Noteworthy is the fact that web access patterns exhibit a dynamic nature, which can be attributed not only to the continuous evolution n of web data but also to fluctuations in user interests [21]. Web users demonstrate a propensity to express interest in data that is derived from various events, including but not limited to weddings, Christmas, birthdays, and similar occasions. Hence, it is imperative to ascertain the specific areas in which these dynamic trends influence the process of noise reduction in online data.

This research suggests a machine learning method able to pick up noise in online data before eradication to address dynamic challenges in a decrease of noise in web data. The algorithm under consideration takes into account the dynamic fluctuations in user interests and the continuous evolution of online data intellect and acquires knowledge about noisy data. The primary contributions of this study involve showcasing the influence of users' changing interests and developing online data to reduce noise in web data. This considers the contributions made by recent research studies and their limitations on the state of knowledge.

The objective is to present a machine learning algorithm that can learn and identify noise within an online user profile before it is eliminated.

The removal of noise derived from a user profile on the web is not solely reliant on pre-existing noise data patterns. Instead, it incorporates the acquisition of noise levels through the analysis of dynamic shifts in user interest and the continuous evolution of web data.

The user's text does not contain any information to rewrite. The practical implementation of the proposed technology is expected to result in a decrease in the quantity of valuable information that is discarded as noise from an online user profile. This has the potential to greatly enhance the web quality user's profile.

The subsequent sections of this study are structured as follows: Section II situates the proposed work within the context of existing research. Section III of this paper examines the planned NWDL (Networked Digital Library) methodology. Section IV of the paper presents the experimental data and subsequent analysis. Section V serves as the concluding section of this work.

Current Research Works

In contemporary times, a vast majority of individuals rely on web pages as a primary source for accessing diverse information, encompassing news, sports, technology, and more. However, it is worth noting that these web pages sometimes incorporate extraneous elements, such as photographs, video clips, and advertisements, which can impede users' ability to efficiently obtain the desired information. To eliminate noisy data, previous technologies employed static web matching patterns, wherein the appearance of the main page was compared to the rest of the screen. If a match was not found, the unmatched data was removed from the web pages, resulting in the display of only the relevant data to the user. The efficacy of this static technique may be compromised in situations when web pages undergo dynamic changes in their appearance and user experience. The existing techniques that have been created to identify and remove noise from online pages primarily rely on the presentation of these web pages. As an illustration, the Site Style Tree (SST) was introduced by [5] as a method for identifying and removing extraneous data from web pages. The SST approach is founded upon the fact that the primary content of a web page typically exhibits a consistent presentation style, whereas any pages deviating from this style are regarded as extraneous noise. To mitigate Toe's extraneous elements on online sites, the SST algorithm employs a mapping technique that compares the page in question to the primary web page. This comparison is based on the presentation style used on the page, allowing for the determination of its utility or superfluousness. The Pattern Tree method (22) is an additional tool for reducing noise in web data, specifically focusing on the design of a website. This algorithm operates based on the concept of the Document Object Model (DOM) tree. It assumes that data found On the internet, one classified as noise if its pattern differs significantly from the primary content of the web page. The Least Recently Used (LRU) paging algorithm [23] is employed to identify and eliminate ages. The Least Recently Used algorithm considers both visual and non-visual attributes of a webpage, allowing it to effectively filter out extraneous web content such as news articles, blog posts, and online discussions. The Least Recently Used (LRU) method is used to identify sites that have been accessed frequently and those that have not been accessed for a significant duration.

However, the primary objective of this study is not to analyse the web's architecture and design data to detect and remove irrelevant information the main emphasis is placed on the analysis of extracted web log data, which is used to establish a comprehensive web user profile. The identification and elimination of noise in web data should be conducted with consideration for user interest levels in the data, as discussed in the preceding section. Recent studies have utilised established machine learning techniques to identify user interest data and filter out irrelevant information from web data logs. As an illustration, the authors in reference [17] employed a combination of case-based reasoning (CBR) and neural network techniques to mitigate the noise present in web log data. The CBR methodology is a machine learning technique that leverages prior experiences to address forthcoming challenges. Specifically, it uses a repository of previously recorded noise data to identify and classify noise present in web pages.

Various noise patterns found on websites are recorded in a Document Object Model (DOM) tree. Subsequently, the case base is queried to identify comparable noise patterns that already exist. Artificial neural networks are employed to match pre-existing noise patterns that are stored within a case-based system. Despite being grounded in the concept of case-based reasoning, this technique encounters challenges in determining the relevance of information that matches current noise patterns stored in a case-based system. The reason for this is that web data is subject to constant change, as is the projected user interest. If the determination of data usefulness is based on a case-based approach, the resulting output may be misleading. The k-nearest neighbours (kNN) algorithm, as described by [24], was employed to effectively utilise pre-existing noise data to identify noise present in web pages. The primary emphasis of their study revolved around the impact of local noise, such as advertisements, banners, and navigational links. The extraction and analysis of web log data was conducted to determine the server to which each entry belongs. If the address is found on a predetermined list of advertisement servers, the link will be eliminated.

Current noise data patterns may become out of date and hence challenging to identify and eradicate noise due to the dynamic nature of user interests and evolving online data.
from a web user profile. To assess user interest levels based on extracted web log data, the Nave Bayesian classification technique was employed [25]. The primary goal of the study was to categorise and analyse web data logs that were collected to evaluate and value the user interbout stage of the process entailed the elimination of extraneous data, such as advertisement banners, graphics, and screen savers, from the collected online data logs. The researchers employed a Nave Bayesian classification model to categorise the data into usable and noisy categories. This classification was based on the variables of the number of pages visited and the time spent on a single page.

Nevertheless, it is important to note that an increased duration of time spent on a web page does not necessarily indicate a user's level of interest. If a user encounters difficulty locating material of personal interest, he or she may allocate additional time to the search process. The concept of weighted association rules refers to a statistical technique used to analyse the relationships between items in a dataset, taking into account the relative importance or weight assigned to each item. Mining was also employed by [26] to find useful information in internet log data. The study aims to analyse pages accessed by a user and allocate weights according to their level of interest. The significance of a web page to a user's attention is determined by the frequency of page visits and the quantity of pages visited. Pages that are visited only once by a single user will be assigned low weights and thereafter regarded as noise.

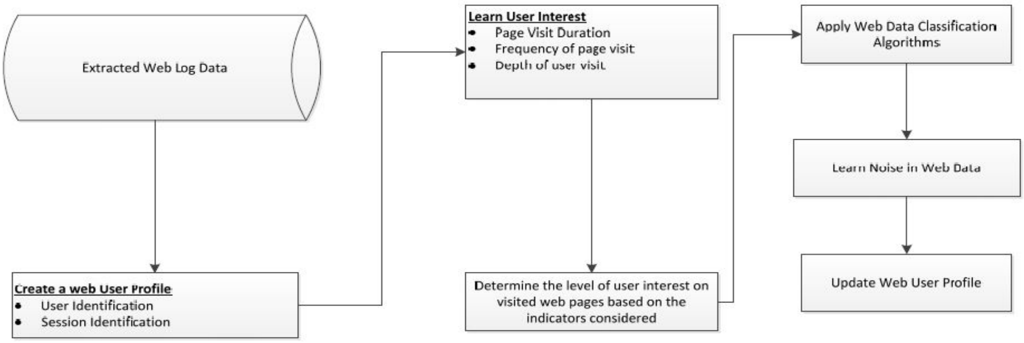
The proposed system

To address the aforementioned issue, the author suggests the use of a technique known as noise web data learning (NWDL). This technique involves the server maintaining a log for each user's accessed page, which is referred to as the weblog dataset. The dataset will contain essential information, including user_id, access_page, date_time, and URL. Through the examination of log data, it is possible to discern the websites that users find interesting, whether they are dynamic or static in nature interest to a user can be identified by examining the frequency of web page access by an individual user and the total time spent on each page.

If a user spends a significant amount of time on a webpage and accesses it more than twice, it might be inferred that the person has a genuine interest in the content of that page. If a person allocates minimal time to viewing and infrequently visits a certain webpage, it can be categorised as an unengaged page and referred to as a noise page. while the authors of the discussion in this section attempted to extract meaningful information from web log data, they focused on removing noise data based on existing noise data patterns and page visit time to assess user interest.

Despite the efforts made by current research to address the issues related to noise in extracted web log data, this study identifies some significant issues that are still unresolved by existing research. For instance, the World Wide Web exhibits a dynamic nature, characterised by the continuous posting and updating of a substantial amount of material on a minute-by-minute basis. The majority of data found on the web has a limited lifespan and loses its usefulness rather quickly. 2) The interests of users in available web data tend to shift when the web data itself undergoes evolution. Web users demonstrate a diverse range of interests in information, which are influenced by temporal factors and global events. Hence, the interests of users have the potential to change as the internet continues to develop. The rationale behind our assertion is that the lack of a clear definition and thorough analysis of noise in online data can undermine the integrity and utility of the retrieved data. The acquisition of knowledge regarding noise in web data is subject to the actions undertaken by a user when interacting with this data. These actions can be quantified using several metrics, including the duration of time spent, the frequency of visits, as well as how much a user delves into the content of a certain web page. These measures will impact the utility of a webpage for a user, as opposed to the interconnections between web data within a specific website.

System Architecture



Feasibility Study

A feasibility study is an assessment conducted to determine the practicality and viability of a proposed project or endeavour. It involves analysing various factors. In this phase, an analysis is conducted to assess the feasibility of the project. Subsequently, a business proposal is presented, outlining a basic plan for the project along with cost estimates. During the process of system analysis, A feasibility study of the suggested system is required. The purpose of this measure is to ascertain that the planned system does not impose any undue strain on the organisation. To conduct a feasibility analysis, it is important to possess a comprehensive understanding of the principal needs of the system.

Economical Feasibility

The purpose of this study is to assess the economic implications that the execution of the system will have on the organisation. The company's financial resources allocated to the research and development of the system are constrained. It is imperative to provide justifications for the expenses. The developed system was successfully implemented under the allocated budget, mostly due to the utilisation of freely accessible technologies. The purchase was limited to only the products that were customised.

The examination of the technical feasibility of a project is an essential aspect in determining its viability and potential success. This assessment involves evaluating the

Technical feasibility

The purpose of this study is to assess the technical viability of the system by examining its technical needs. The development of any system should not impose a significant strain on the existing technical resources. This phenomenon will result in increased pressure on the existing technical resources. This phenomenon will result in the imposition of significant expectations on the client. The system under development should possess a modest demand, as it necessitates only minimal or negligible modifications for its implementation.

Social feasibility

The social feasibility of a project refers to its potential to be accepted and embraced by the community or society in which it will be implemented

The objective of this study is to assess the degree of user acceptability of the system. This encompasses the procedure of instructing the user to effectively utilise the technology. The user should adopt a non-threatening stance towards the system and acknowledge its indispensability. The degree of user acceptance is contingent upon the strategies utilised to educate the user about the system and facilitate their familiarity with it. It is imperative to enhance his level of confidence to enable him to provide constructive criticism, which is highly valued given his role as the ultimate user of the system.

Implementation

The Convolutional Neural Network (CNN) algorithm is a deep learning model that has been designed for processing data with a grid pattern, such as photographs. It draws inspiration from the organisation of the visual cortex in animals and aims to learn spatial hierarchies of information autonomously and adaptively, progressing from low-level to high-level patterns. The convolutional neural network (CNN) is a mathematical framework that commonly consists of three fundamental layers: convolution, pooling, and fully connected layers. The initial two layers, namely the convolution and pooling layers, are responsible for conducting feature extraction. On the other hand, the third layer, known as the fully connected layer, is responsible for mapping the extracted features to generate the final output, which may involve classification tasks. The convolution layer is a crucial component in convolutional neural networks (CNNs), consisting of a series of mathematical operations, including convolution, which is a specialised form of linear operation. In the context of digital images, the values of pixels are often stored in a two-dimensional grid, referred to as a 2D array (Figure 2). Additionally, a kernel, which is a small grid of parameters, is employed at each point within the image. This kernel serves as an optimizable feature extractor. The use of convolutional neural networks (CNNs) in image processing is advantageous due to their efficiency in processing images and the application of kernels at various image positions.

The occurrence of a feature inside a picture is not limited to a certain location. As the output of one layer is passed on to the subsequent layer, the extracted features have the potential to undergo a hierarchical and progressive development towards increased complexity. The act of fine-tuning parameters, such as kernels, is referred to as training. This process aims to minimise the discrepancy between outputs and ground truth labels by utilising optimisation techniques like backpropagation and gradient descent, among others.

The support vector machine (SVM) algorithm is a popular machine-learning technique used for classification and regression tasks. However, its predominant application is in the realm of categorization problems. In the support vector machine (SVM) algorithm, data items are represented as points in an n-dimensional space, where n corresponds to the number of features available. Each feature is assigned a specific coordinate value inside this space. In this context, classification is achieved by identifying a hyperplane that effectively distinguishes between the two classes, as depicted in the provided picture.

The aim in scenario 1 is to choose the proper hyperplane. Three hyperplanes are identified in this context as A, B, and C. Selecting a hyperplane that can reliably categorise objects as stars or circles is the current challenge at hand.

When attempting to choose the proper hyperplane, it is crucial to keep in mind a vital principle: "Choose the hyperplane that effectively separates the two classes. Hyperplane "B" has performed this operation in this instance with outstanding efficiency.

The current aim is to locate the proper hyperplane that successfully divides the classes. The difficulty that now emerges is how we may efficiently choose the proper hyperplane.

By optimising the distances between the closest data points, independent of their class, and the hyperplane, the best hyperplane can be found in this situation. The margin is a popular name for this measurement.

According to the analysis, hyperplane C's margin is more significant than that for A and B. The right hyperplane is given the designation "C" as a result. The robustness of the hyperplane with the wider margin is another

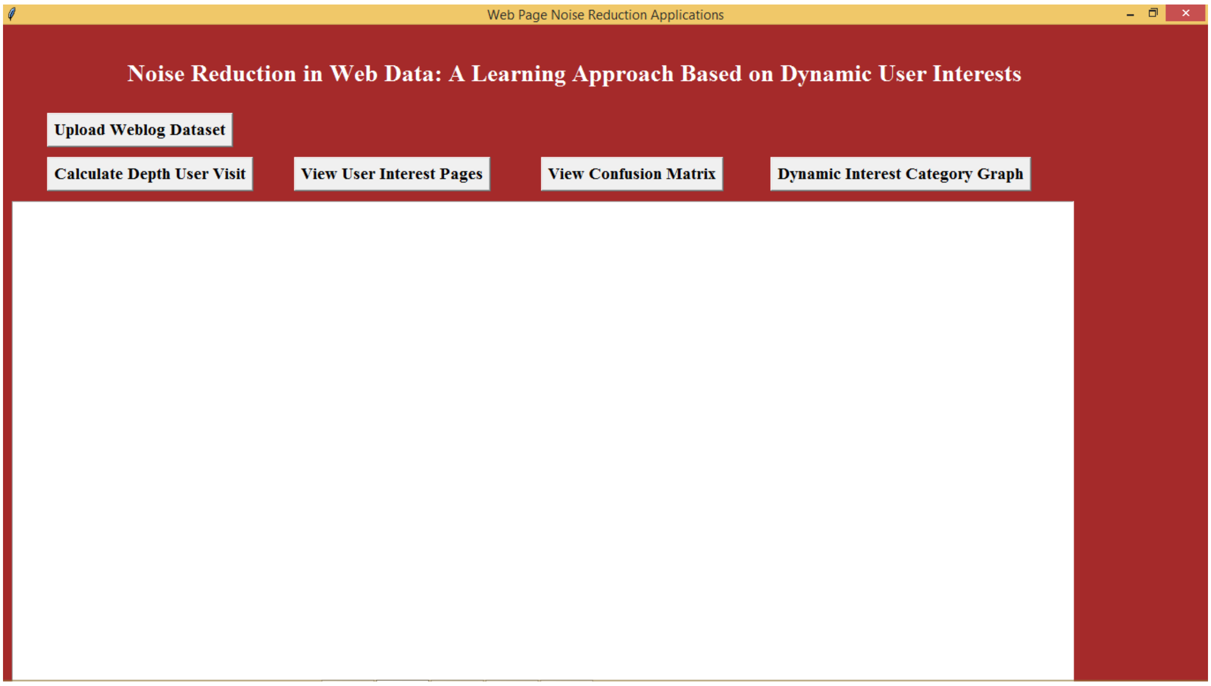
strong argument in its support. Misclassification is more likely to occur if a hyperplane with a slim margin is employed.
A popular unsupervised machine learning method for dividing data into discrete clusters is the K-Means Clustering Algorithm.

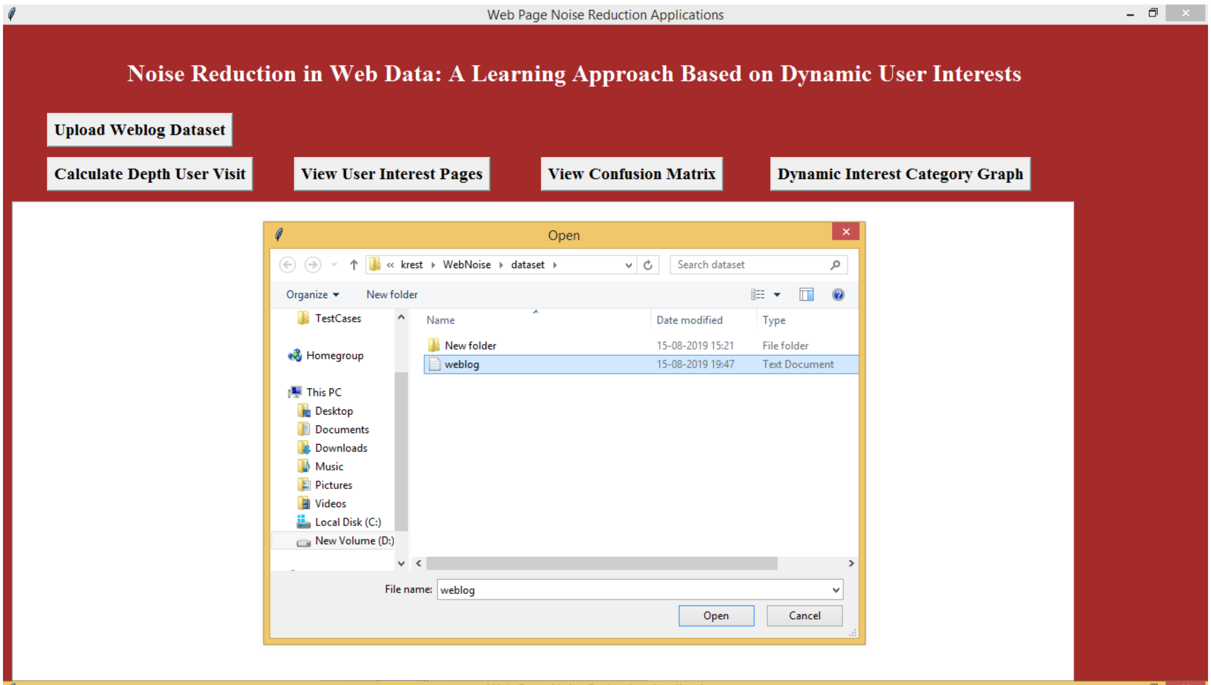
The K-means algorithm, which Lloyd introduced in 1957, is the partitioning method that is most frequently employed. To decrease the sum of squared Euclidean distances between each observation and the mean of its related cluster, this study aims to discover K clusters. The K-means algorithm can be thought of as a two-step process. Each observation is first assigned to the cluster with the closest centre for a given set of cluster centres. Second, each cluster centre is updated by computing the sample mean of all the points contained inside that cluster for a particular assignment of observations to clusters. The first stage usually involves choosing a random sample of K observations to calculate the centre values. The various local optimal points tend to converge, rather than the overall optimum, during the optimisation process. The algorithm offered by Hartigan and Wong (1979) is more complex and has a better chance of finding a beneficial local optimum. No matter the specific algorithm used, it is advised to iteratively start the process with different initial values. This strategy increases the possibility of finding a beneficial local optimum.

In several fields, the K-Means algorithm is a popular clustering technique. Initialising k cluster centres, where k's value has been predetermined, starts the procedure. Each object (also known as an input vector) inside the dataset is then assigned to the cluster with the closest proximity to the cluster's centre. The cluster centre is then updated by computing the mean (centroid) of each cluster. The modification in the membership of each cluster's makeup led to this upgrade.
Until the value of any of the cluster centres no longer changes, the iterative processes of object reassignment and cluster centre updating are carried out.

While it can be demonstrated that the k-means algorithm will always terminate, it does not guarantee the discovery of the most optimal configuration, which corresponds to the minimum of the global objective function. The method is highly influenced by the initial selection of cluster centres, which has a considerable impact on its performance.
The k-means algorithm can be executed several times to mitigate this phenomenon. The K-means algorithm is a straightforward method that has been applied to several problem domains.

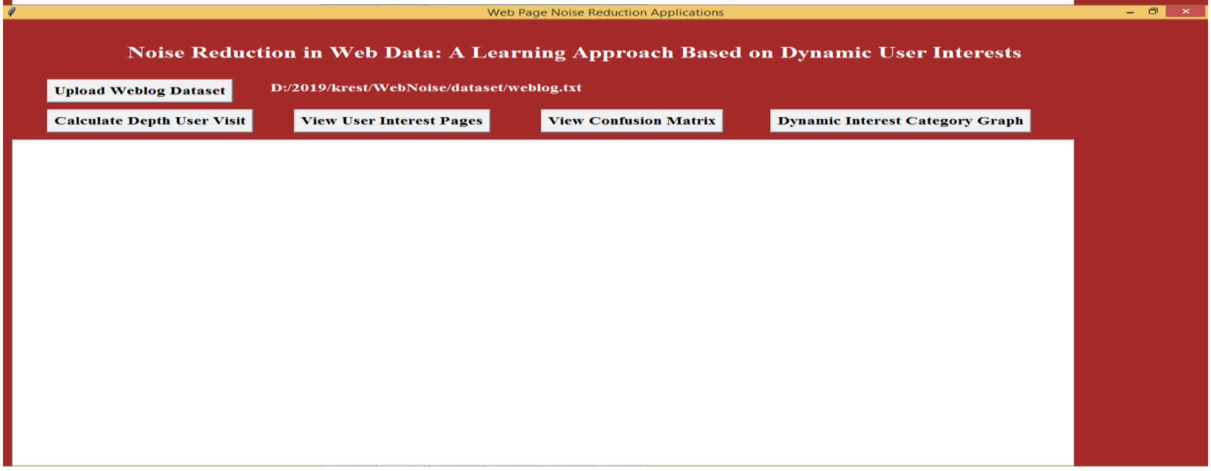
In the second scenario, the user's text will be rewritten to have an academic tone in this context, there are three hyperplanes denoted as A, B, and C.


Web Page Noise Reduction Applications
Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests
Upload Weblog Dataset
Calculate Depth User Visit
View User Interest Pages
View Confusion Matrix
Dynamic Interest Category Graph

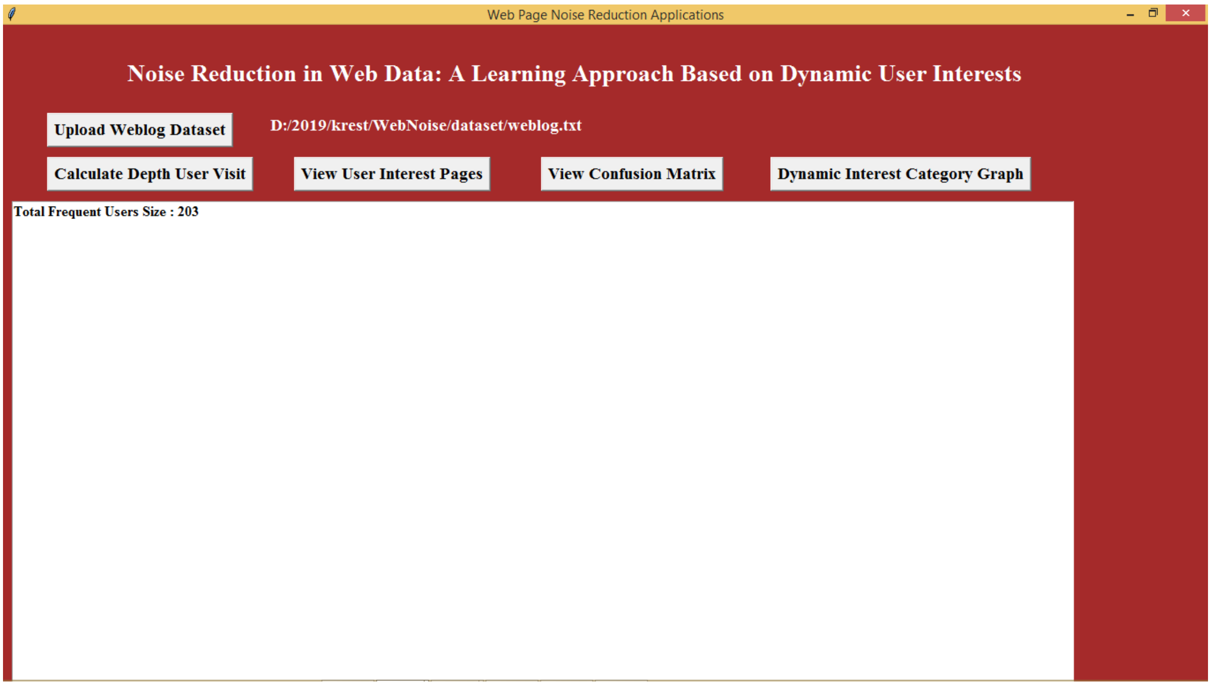


Name	Date modified	Type
New folder	15-08-2019 15:21	File folder
weblog	15-08-2019 19:47	Text Document

File name: weblog

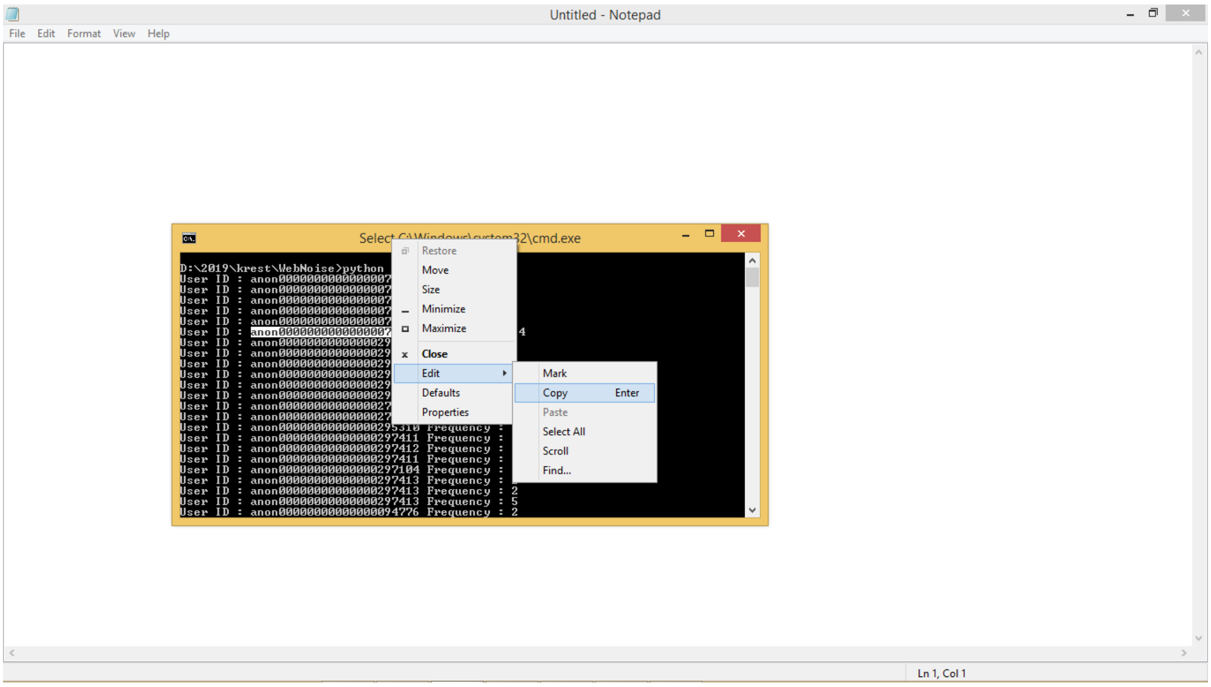

Web Page Noise Reduction Applications
Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests
Upload Weblog Dataset
Calculate Depth User Visit
View User Interest Pages
View Confusion Matrix
Dynamic Interest Category Graph

D:/2019/krest/WebNoise/dataset/weblog.txt

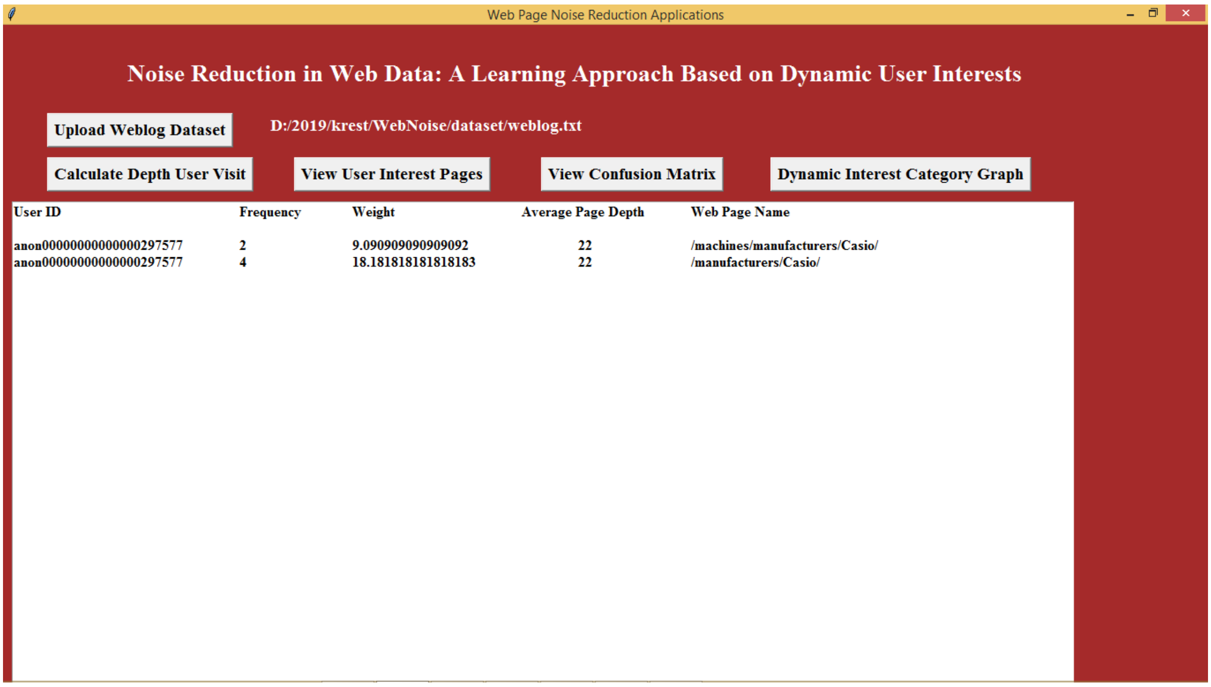

Web Page Noise Reduction Applications
Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests
Upload Weblog Dataset
Calculate Depth User Visit
View User Interest Pages
View Confusion Matrix
Dynamic Interest Category Graph

D:/2019/krest/WebNoise/dataset/weblog.txt

```
Total Frequent Users Size : 203
```



```
D:\2019\krest\WebNoise>python
User ID : anon0000000000000007
User ID : anon0000000000000007
User ID : anon0000000000000007
User ID : anon0000000000000007
User ID : anon0000000000000007
User ID : anon0000000000000007                    4
User ID : anon0000000000000029
User ID : anon0000000000000029
User ID : anon0000000000000029
User ID : anon0000000000000029
User ID : anon0000000000000029
User ID : anon0000000000000029
User ID : anon0000000000000027
User ID : anon0000000000000027
User ID : anon00000000000000295310 Frequency :
User ID : anon00000000000000297411 Frequency :
User ID : anon00000000000000297412 Frequency :
User ID : anon00000000000000297411 Frequency :
User ID : anon00000000000000297104 Frequency :
User ID : anon00000000000000297413 Frequency :
User ID : anon00000000000000297413 Frequency : 2
User ID : anon00000000000000297413 Frequency : 5
User ID : anon00000000000000094776 Frequency : 2
```


Web Page Noise Reduction Applications
Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests
Upload Weblog Dataset
Calculate Depth User Visit
View User Interest Pages
View Confusion Matrix
Dynamic Interest Category Graph

D:/2019/krest/WebNoise/dataset/weblog.txt

User ID	Frequency	Weight	Average Page Depth	Web Page Name
anon00000000000000297577	2	9.090909090909092	22	/machines/manufacturers/Casio/
anon00000000000000297577	4	18.181818181818183	22	/manufacturers/Casio/

STEP BY STEP PROCESS:
1. Web User Profile
2. Learning of User Interest
3. Page Visit Duration
4. Frequency of a User Visit
5. Determine the Weight of the kth Web Page
6. Depth of User Visit
7. Web Page Category Weight
8. Learn Noise in Web data

Step 1. Web User Profile

The first step is to check the user profile or to create a user profile A user profile has a set of URLs that represent a user interest. Creating a user profile is based on a set of web pages accessed by a user taking into account the relevance of his/her interest. The user profile denoted by Uj contains several sessions i.e. Uj = (S1, S2,…Si…..SI) where SI are the number of user sessions. The i th user session is defined as a sequence of accessed pages for the j th user, i.e. Si j= (url1, url2,…urlk,...urlK) where urlK is the number of web pages for the j th user. After creating a user profile, this work learns user interest levels on visited web pages to determine useful information from noise data. Various measures are considered, i.e. time, frequency and depth of visit of user visit to a web page.

Step 2. Learning of User Interest

Once we create a user profile we look for user interests on specific websites.
The current research work recognises that it is important to learn user interest levels to find useful information. This can be done by collecting user log data, analyzing it and storing the results in a user profile. User interest relies on the basis that the visiting time of a web page is an indicator of a user's interest level. The amount of time spent on a set of web pages requested by the user within a single session reflects the interest of that user. In addition, states that web pages with higher frequency are of stronger interest to a user. Even though this paper considers page visit duration and frequency of visit to learn user interest in visited web pages, it is difficult to measure user interest levels based on page visit duration and frequency of visit alone. For example, a high frequency of visits to a web page may either reflect a user struggling to find useful information or based on website layout, he/she is forced to visit some pages before accessing interested ones. Therefore, the proposed work considers additional measures such as depth of visit and frequency of visit to a web page category to learn user interest before the elimination of noise data.

Step 3. Page Visit Duration

After learning user interest on page next we should check for page visit duration. Page visit duration is one of the metrics widely used by current research work to measure user interest level on a web page. Page visit duration is the amount of time a user spends viewing a web page, it reflects the relative importance of each page to the j user. Generally, a user spends more time on a more useful page, if a user is not interested in a page, he/she will exit or move to another page. Therefore, page visit duration defines the length of user interest on a web page. argue that the calculation of page visit duration is a bit skewed because it is not possible to determine the time a user exits a web page as it is always 0. Therefore, the last page visited is not counted as a part of the page visit duration, but as the number of pages visited/viewed. The number of page views (NPV) represents the number of times a user views a specific web page in th session

Step 4. Frequency of a User Visit

Once after complete of page visit duration, we must check the frequency. The frequency of a user's visit to a web page is determined by the number of times k th web page appears in ith session for the first user. The frequency of the jth user on a kth web page is presented.

Step 5. Determine the Weight of the kth Web page

Once after completing of frequency of user visits, we must check the weight of the required page The weights signify the importance of the k web page in j th user profile. The weight of k th web page is the interest degree of the j th user on k th web page, it is denoted as Wk j which is determined by the length and frequency of the j th user visit.

Step 6. Depth of User Visit

Once after complete of weight of the required page, we must check the depth of the user's required page. Page visit depth is defined as the average number of pages viewed by visitors during a single browser session. The depth of the user's visit on k th web page is an indicator of a user's interest level. The proposed tool considers the depth of the user visit not only in terms of the number of page views but also the route a user takes to navigate through a website. The jth user creates a path of page views when searching for information on a specific website. For example, a user may enter a website from the home page but only be interested in finding delivery charges for a specific item under accessories. Even though the j th user is likely to visit other web pages to get to the information of interest, it is difficult to assume that every page visit is of user interest unless measures such as time duration and frequency to visit over several sessions are considered. Therefore, the path taken by the j th user from the entry to the exit page and the weights associated with each k th web page are considered in this paper to learn interest levels for the j th user.

Step 7. Web Page Category Weight

In this work, a web page category is defined as a set of related web pages. The weight of a web page category is determined based on the frequency of user visits to a particular web page category. The more frequently a user visits the same category the higher the level of interest. Unlike the frequency of visits to k th web page discussed in (2), the frequency of visits to a web page category determines if a user is interested in information from a given category of web data. For example, a high number of visits to footwear web pages under the men's category depict interest in information regarding men's shoes. Based on this concept, the weight of a web page category is presented to learn the user interest level of a particular web page category. The weight of a web page category is defined by taking into account the number of times a particular category of a web page denoted as Cm appears in ith session for the jth user, where mth is an indicator of a web page category. The proposed tool determines the number of times k th web page of Cm appears in j th user profile Average length of time spent on Cm by the j th user profile

Step 8. Learn Noise in Web data

At last, by calculating all the frequency, and depth we determine the noise in web data.

Conclusion

In conclusion, it can be inferred that the aforementioned points support the notion that this paper presents a machine-learning system that can learn and identify noise in online data before it is eliminated. This paper commences by delineating and characterising the obstacles associated with the existing research endeavours in the process of reducing noise in web data. For instance, the process of removing noise from online data relies on preceding patterns of noise. However, when a user's interests change, the previously recorded noise data patterns become unreliable and therefore irrelevant. Furthermore, contemporary research endeavours regard noise as any information that does not constitute an integral component of the primary webpage.

Hence, the task of identifying and eliminating noise in web data is challenging due to the need to consider the dynamic interests of a web user.

This study employs a series of measures to tackle the highlighted issues. The present study introduces a machine learning technique that takes into account the dynamic fluctuations in user interests by acquiring knowledge about the extent of a user's engagement with a particular webpage. Additionally, this study proposes an algorithm that effectively incorporates evolving web data and changes in user preferences when learning from noisy web data. The algorithm under consideration possesses the capability to ascertain the specific interests of users at any given moment, as well as their search behaviour and whether they exhibit interest in their past search queries before discarding them. Ultimately, the suggested tool serves to enhance the quality of a web user's 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 behaviour 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. Secure. 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.
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 a 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.