

WEB DATA LEARNING (NRWDL) ALGORITHM TO DETECT DYNAMIC CHANGES IN USER INTERESTS

Dr.D.V.Manjula, Ms.G.Naga Rani, Mrs.K.Shamsri, Mrs. Aditi Nautiyal
Department of CSE, PRAGATI Engineering College (Autonomous), Surampalem, A.P

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.

I 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 recognized 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. 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. User interests can be implicit or explicit. Explicit interests are where a user tells 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 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. To address dynamic issues in noise web

data reduction.

II. 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. While the authors discussed in this section aimed at finding useful information from web log data, they concentrated on eliminating noise data based on existing noise data patterns and page visit duration to determine interest level of a user. Despite efforts from current research work to address problems with noise in extracted web log data, this work observes some critical issues not fully addressed by current research work. For example, 1) The Web is dynamic where a high volume of data is posted and updated every minute. The majority of web data only remains useful for a very short period of time. 2) User interests on available web data tend to change as web data evolves. In essence, web users express 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 analyzed 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.

III. PROPOSED NOISE WEB DATA LEARNING (NWDL)

In this section, a machine learning algorithm capable of learning noise in a web user profile prior to elimination is proposed. A key focus is to learn, identify and eliminate noise, taking into account the dynamic interest of a user and the evolving web data. Eliminating noise in extracted web log data is determined based on what a user is interested and not interested in. It is widely discussed in current research work, [5], [8], [27], [28] that the interest of a user on a web page is measured by how often they visit that page, how long they spend on the page, how recently they visited the page and the number of links on the page that they visit. To some extent, current research works measure user interest in extracted web data logs but there is inadequate evidence to demonstrate how noise in a web user profile is determined

prior to elimination. A summary of the proposed work is shown in Fig. 1. A Web User Profile A user profile has a set of URLs that represent a user interest. Creating a user profile is based on a set web page accessed by a user taking into account relevance of his/her interest. User profile denoted by contains a number of sessions where are a number of user sessions. The user session is defined as a sequence of accessed pages for the user, whereis the number of web pages for the user. After creating a user profile, this work learns user interest levels on visited web pages so as 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.

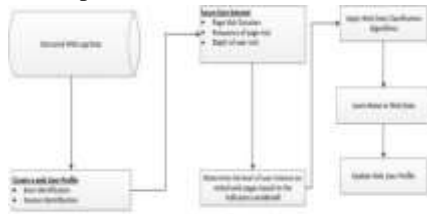


Fig. 1 Proposed process flow

B. Learning of User Interest

The current research work [29] recognizes that it is important to learn user interest level 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 [30]. The amount of time spent in a set of web pages requested by the user within a single session reflects the interest of that user. In addition, [31] 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 on visited web pages, it is difficult to measure user interest levels based on page visit duration and frequency of visit alone. For example, high frequency of visit 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 prior to elimination of noise data.

1. 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 [32]. Page visit duration is the amount of time a user spent viewing a web page, it reflects the relative importance of each page to the 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

$$T_k^j = \sum_{i=1}^{I_j} \frac{T_{k_i}^j}{NPV_i^j - 1} \quad (1)$$

2. Frequency of a User Visit

Frequency of a user visit to a web page is determined by the number of times web page appears in session for the user. Frequency of the user on a web page is presented in the following equation:

$$Freq_k^j = \frac{\sum_{j=1}^{K_j} url_k^j}{I_j} \quad (2)$$

3. Determine Weight of Web Page

The weights signify the importance of the web page in user profile. The weight of web page is the interest degree of the user on web page, it is denoted as which is determined by the length and frequency of the user visit using the following equation:

4. Depth of User Visit

Page visit depth is defined as the average number of pages viewed by visitors during a single browser session [34]. The depth of the user visit on web page is an indicator of a user interest level. The proposed tool considers the depth of the user visit not only in terms of a number of page views but the route a user takes to navigate through a website. The user creates a path of page views when searching for information on a specific website. For example, a user may enter a website from home page but only interested in finding delivery charges for a specific item under accessories. Even though the 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 a user interest unless measures such as time duration and frequency to visit over a number of sessions are considered. Therefore, the path taken by the user from entry to exit page and the weights associated with each web page is considered in this paper to learn interest levels for the user.

Algorithm: Depth of user visit

Input: Extracted web user logs

Output: A set of links associated with the user profile
1. Define the user profile; // see definition 1

2. for each web page in the j user profile
3. find the web page category

4. if two web page from the same category are both included in session

5. flag_Link = 1; // a link between two web page from the same category is found

6. else

7. flag_Link = 0; // no link between the two web pages
8. if (flag_Link = 1)

9. out_List.put (url:url); // web pages visited by the user profile are connected

10. end if

11. end if
12. end

5. Web Page Category Weight

In this work, 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 frequent a user visits the same category the higher the level of interest. Unlike frequency of visit to web page discussed in (2), the frequency of visit 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 men category depict interest on information regarding men shoes. Based on this concept, the weight of a web page category is presented for the purposes of learning user interest level to a particular web page category. Definition 4. 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 appears in session for the user, where is an indicator of a web page category. The proposed tool determines: technique called back propagation to adjust the outcome with the expected result or class.

Algorithm: Learn noise web data

Input: Weighted url for the user profile
Output: A set of web pages assigned to a class
1. Define the user profile

2. for each web page in j user profile do
3. Determine the weight of web page using (3)
4. if url weight > threshold set then

5. assign to
6. else
7. assign to
8. for do

9. Create a simple page link of the user profile
10. Determine frequency to web page category using
11. if Freq < threshold set then

12. assign to
13. else
14. update
15. end if
16. end if
17. end

IV. EXPERIMENT AND ANALYSIS OF RESULTS

This paper demonstrates the performance of the proposed tool over well-known algorithms applied in noise web data reduction

process. The experimental design aims to demonstrate how noise in web data is identified and eliminated by existing algorithms and the proposed algorithm taking into account changes in user interests. Firstly, the choice of the dataset and various tasks carried

out in this experiment are presented and discussed. Secondly, the performance of the proposed algorithm is demonstrated by the results obtained from experiments carried out. Finally, the results from existing algorithms and the proposed algorithms are compared. This is to validate the performance of the proposed algorithm over the existing algorithms.

A. Experimental Design Scheme

The dataset used to carry out the experiments are user web log data extracted from an e-commerce site for a period of one month. The proposed algorithm identifies user activities on the website to determine dynamic changes over the specified period. A number of current machine learning algorithms are considered in this work to validate the performance of the proposed algorithm. The objective of this experimental direction is to demonstrate if existing tools identify and classify extracted web log data in relation to varying user interest on web pages visited. We validate



Fig. 2 Experimental design setup scheme

the performance of the proposed tool by comparing its output against the output from existing algorithms considered in this paper. Summary of the experimental design setup scheme is presented in Fig. 2. To evaluate the

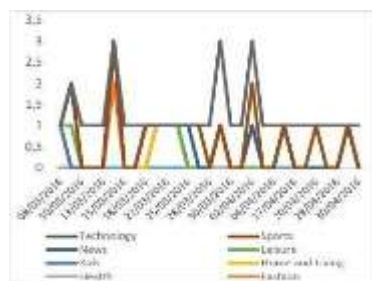


Fig. 3 Dynamic interests of a user in a web page category

proposed algorithms in terms of classification performance, we divide the dataset into training and test data. We perform ten-fold cross validation and report the classification performance in terms of Precision and Recall. The choice of cross-validation is to iterate separations into a training set and test set in order to get more reliable results in terms of classification accuracy.

B. Experimental Results

The proposed tool aims to demonstrate how dynamic change in user interest affects the elimination of noise in a interests to a specific web page category are obtained using (3). Unlike current algorithms, eliminating noise from web data in a user profile is determined based on changes in user interest over time and not existing noise data patterns. For example in week one, the user visit to sports category was none, which means it is considered as not useful. But over time a user has visited the category if existing noise data patterns will be used to eliminate noise in web data, then useful information is likely to be eliminated. Frequency and duration of a user visit to a web page alone cannot determine the level of a user interest. There is a possibility that a user will only be interested in a given time period. Therefore, frequency and duration of visit will not provide a clear picture of a user interest when eliminating noise web data. 1. Baseline Selection The objective of applying a number of well-known machine learning algorithms, as shown in Fig. 4, was to establish a baseline among them for the given data set. A baseline is a method that uses machine learning to create predictions for a dataset. The results obtained are then used to measure the baseline's performance in terms of Precision and Recall. The best performing algorithm is used to validate the performance of the proposed algorithms using confusion matrix as shown in Fig. 5. C. Evaluation of Results In this section, we analyze the experimental results of the proposed algorithm in terms of classification accuracy. Accuracy is the fraction of correct classification out of total possible data classification in a web user profile. Although it is common in experimental studies to iterate separations into a training set and test set and to use cross-validation in order to get more reliable results, this paper also considered a random split of the original data set because of the size of the dataset used. The training data set used is randomly split from the original dataset i.e. we have considered 70% training data and 30% test data. However, the results presented in Fig. 4 do not show much difference despite the size of the dataset used. The proposed algorithm/tool is compared against a baseline machine learning algorithm as shown in Fig.5

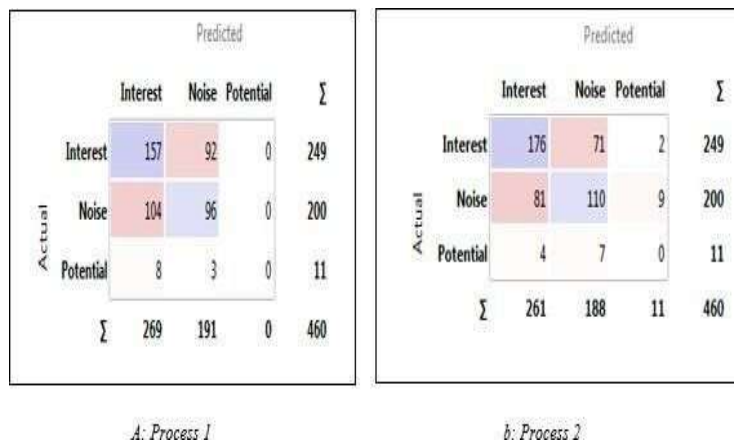


Fig. 5 Comparing performance of the proposed algorithm over current algorithms using confusionmatrix

V. 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.

VI. 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, 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.
- [5]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.
- [6]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. [7]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.
- [8]S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, 'User profiles for personalized information access', in The adaptive web, Springer, 2007, pp. 54–89. [9]P. Peñas, R. del Hoyo, J. Veja-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. [10]S. Kanoje, S. Girase, and D. Mukhopadhyay, 'User profiling trends, techniques and applications', ArXiv Prepr. ArXiv150307474, 2015.
- [11]H. Kim and P. K. Chan, 'Implicit indicators for interesting web pages', 2005.
- [12]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.
- [13]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.
- [14]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.

- [15]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.
- [16]R. P. Velloso and C. F. Dorneles, 'Automatic WebPage Segmentation and Noise Removal for Structured Extraction using Tag Path Sequences', J. Inf. Data Manag., vol. 4, no. 3, p. 173, Sep. 2013.
- [17]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.
- [18]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.