IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

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

Dr. Sumagna Patnaik, A. Niharika Reddy, N. Varun Goud, CH. Sai Venkat

Professor, Student, Student, Student
Information Technology,
JB Institute of Engineering and Technology, Hyderabad, India

Abstract: This explores how current available tools address problems with noise in web user profile. We establish that current research works eliminate noise from web data mainly based on the structure and layout of web pages i.e. they consider noise as any data that does not form part of the main web page. However, not all data that form part of main web page is of a user interest and not every data considered noise is actually noise to a given user. The ability to determine what is noise and useful to a dynamic web user profile has not been fully addressed by current research works. We aim to justify a claim that it is important to learn noise prior to elimination, to not only decrease levels of noise but also reduce loss of useful information. This is because if noise in web data is not clearly defined and analysed through learning, the purpose and its use will be compromised hence its overall quality

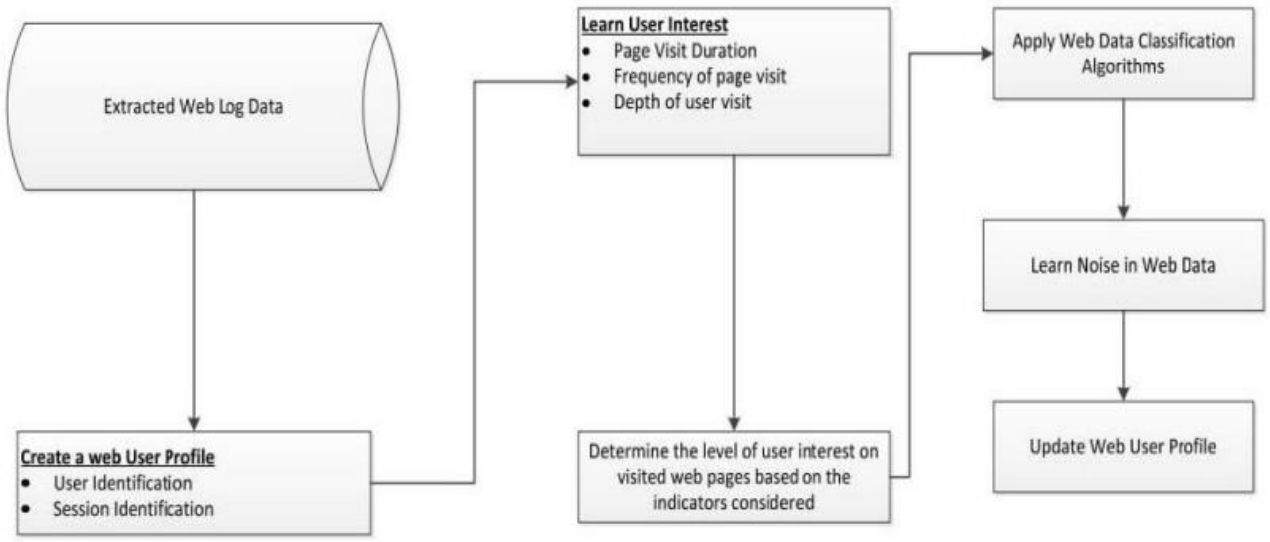
Index Terms **- Web Usage Mining, Web Log Data, Noise Data, User Profile, User Interest.**

1.INTRODUCTION

The information available on the web is increasing rapidly with the explosive growth of the World Wide Web While users are provided with more information, it has become more difficult to extract useful information from the web due to its size and diversity. Moreover, a lot of web data is buried further deep on the web and only a small percentage of useful data is available to users. Extraction of interesting information from web data has become popular in the recent past with more research focus on web usage mining. Web usage mining (WUM) is defined as application of data mining techniques to discover usage patterns from web data in order to understand the needs of web users. WUM attempts to discover useful information from web logs which are interactions of users on the web. In real world, it is practically impossible to extract web log data and create a user profile free from noise. A user profile is defined as a description of user interests, characteristics, and preferences on a given website, user interest is measured by looking at user web log data to determine time spent on web pages and frequency of visit to a web page Useful information on the web is often accompanied by high level of noise e.g. advertisements, navigation panels, copyrights notices, web page links from external web sites, etc., which hinder the process of finding information that meet the interest of a user. Define noise web data as any data that is not part of the main content of a web page, and such data can harm web usage mining process. However, our view on this definition is that noise is not necessarily advertisements from external web pages, duplicate links and dead URL or any data that does not form a part of the main content of a web page but also useful web data that is incorrectly assigned to different data class hence affecting usefulness of web data to a specific user interest. It is recognised that noise web data is an unavoidable problem that affects web usage mining process e.g. mainly because the source of web data is uncontrolled hence difficult find useful data from extracted web log data with high noise levels. Presence of noise in extracted web log data can adversely affect the output from web usage mining process. For example, seasonal data can be useful in a given time period but not useful to a specific user in different occasion, this dynamic change of event to web data may cause current machine learning tools to extract data that does not meet interest of a user. According to there is a need to improve equality of the web data by eliminating noise so as to ensure web data available to a specific user is of their interest. It is widely discussed in current research work that web data need to be pre-processed before applying web data mining tools. The main objective of data pre-processing is to remove noise data, and to reduce the size of data. However, our opinion is that useful information can easily be eliminated at pre-processing stage. It is therefore important to understand the nature of noise data on the web. For example, the home page of a website is likely to contain advertisement banners relevant to geo-location of a user determined through an IP address. The user might be interested to click through the links and spend some time on suggested pages or choose not to. Therefore, to determine if such type of data is relevant to a user or not does not only depends on the relationship of different types of web data to the main web page content but the user's interest determined by frequency and time spent on a given web page. Relevant data in web usage mining field has been defined by various researcher for example, defines relevant web data as sections of a web page that more objectively describe the main content of a web page. This includes the title and the main body section, and excludes comments about the story and presentation elements while defines it as the core information of a web page that a user needs to view. For example, the main content in the web page of a news article is the core information, while anything that does not form part of the main web page is irrelevant. Some current research work e.g. have used noise and irrelevant

data interchangeably. In addition, defines noise as irrelevant data. In this work, noise will be used predominantly which also refer to as irrelevant. Even though current works has played a significant role in reducing noise levels present on the web there is still limited discussion on how loss of useful information otherwise considered noise at pre-processing stage can be decreased as well as reducing levels of noise data in a web user profile. Some web data eliminated at the preprocessing stage can be noise to a specific user but useful to another user. Moreover, the main web page content is likely to contain data relevant to the website but noise to a user. Therefore, it is important to understand the nature of noise data identified against interest of a user prior to elimination. In this work, we aim to establish and justify our position as to why there is a need to learn different type of noise from extracted web log data prior to elimination 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.

2.IMPLEMENTATION



An 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, that the interest of a user on a web page is measured by how often they visit the page , how long they spend on the page, how recently they visit the page and the and the numbers of links on 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 elimination.

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 Weight of 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 user profile or to create 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 pages accessed by a user taking into account relevance of his/her interest. User profile denoted by Uj contains a number of sessions i.e. Uj = (S1,S2,…Si…..SI) where SI are a 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 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.

Step 2. Learning of User Interest

Once we created user profile we look for user interests on specific website.

The current research work recognises 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 . 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, 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.

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 spent viewing a web page, it reflects the relative importance of each page to the j th 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 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 page visit duration, but in 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 i th session.

Step 4. Frequency of a User Visit

Once after completion of page visit duration we must check frequency.

Frequency of a user visit to a web page is determined by the number of times k th web page appears in ith session for the jth user. Frequency of the jth user on a kth web page is presented.

Step 5. Determine Weight of kth Web page

Once after completion of frequency of user visit we must check weight of the required page

The weights signify the importance of the k th 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 completion of weight of required page we must check depth of user 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 j th user visit on k th web page is an indicator of a user interest level. The proposed tool considers the depth of the j th user visit not only in terms of a number of page views but the route a user takes to navigate through a website. The j th 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 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 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 j th user from entry to exit page and the weights associated with each k th web page is considered in this paper to learn interest levels for the j th user.

Step 7. 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 k th 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. 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: 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 , depth we determine the noise in web data.

3.CONCLUSION

Our aim in this work is to justify the need to learn noise in web log data prior to elimination. It is important to take into account both available web data and interests of web users to determine which data is noise or useful given varying interest levels of a user. Our position is based on the fact that interests of a web user are dynamic and so is web data. For this reason, it is difficult with current available tools to eliminate noise without affecting its usefulness.

4.FUTURE ENHANCEMENT

In our ongoing research work, we propose a noise web data learning tool capable of learning noise from a web user profile prior to elimination. In addition to frequency of visit and time duration which are widely applied in current research work, we introduce depth of visit to measure interest level of a user on visited web pages. An example to justify our proposed measure is if a web page appears only once in a user web log it does not mean the user is not interested. There is a possibility that a user will only be interested in a given time period hence a need to introduce this measure to learn dynamic interests exhibited a user before eliminating any noise data from a user profile.

5.REFERENCES

[1] 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.
[2] C. Ramya, G. Kavitha, and D. K. Shreedhara, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process," ArXiv Prepr. ArXiv11050350, 2011.
[3] S. Dias and J. Gadge, "Identifying Informative Web Content Blocks using Web Page Segmentation," entropy, vol. 1, p. 2, 2014.
[4] 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.
[5] 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.
[6] S.Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," in The adaptive web, Springer, 2007, pp. 54–89.
[7] 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.