Noise Web Data Learning from a Web User Profile: Position Paper

Julius Onyancha, Valentina Plekhanova and David Nelson

Abstract- This position paper 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.

Keywords: Web Usage Mining, Web Log Data, Noise Data, User Profile, User Interest.

I INTRODUCTION

The information available on the web is increasing rapidly with the explosive growth of the World Wide Web[1], [2] 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 [3]. 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 [4], [5]. 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 [4]. WUM attempts to discover useful information from web logs which are interactions of users on the web [5]. 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 [6]–[8], 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 [9], [10].

Manuscript received March 28, 2017; revised April 6, 2017

J. Onyancha is with the Faculty of Computer Science, University of Sunderland, Edinburgh Building, Chester Road, Sunderland SR1 3SD (julius.onyancha@research.sunderland.ac.uk).

V. Plekhanova is with the Faculty of Computer Science, University of Sunderland, Edinburgh Building, Chester Road, Sunderland SR1 3SD (valentina.plekhanova@sunderland.ac.uk).

D. Nelson is with the Faculty of Computer Science, University of Sunderland, Edinburgh Building, Chester Road, Sunderland SR1 3SD (david.nelson@sunderland.ac.uk).

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. [1], [11] 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.[12] 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 [13], [14], there is a need to improve quality 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 e.g. [1], [2], [15], [16] 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 [17]–[19]. 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, [20] 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 [16] defines it as the core information of a web page that a user needs to view. For

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

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. [9], [16] have used noise and irrelevant data interchangeably. In addition, [9] 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 [1], [11], [21], [22], 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 p rofile. Some web data eliminated at the pre-processing 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

The rest of this paper is organised as follows; section II is a critical analysis and evaluation of current research work on noise web data reduction. Section III is a discussion of our research position based on contribution made by current research work. Finally, future work this research will undertake to address issues identified from current research work.

II CRITICAL ANALYSIS AND EVALUATION OF RELATED RESEARCH WORK IN NOISE WEB DATA REDUCTION

In this section, we explore various machine learning tools currently applied in noise elimination from web data. We also evaluate contribution of current research works in relation to noise web data reduction and the ability to discover useful information based on a specific user interest. In order to justify the position of this paper, we aim to address the following key aspects:

- To establish from current research work different types of noise web data eliminated by existing tools
- To identify measures used by current available tools to eliminate noise from web log data
- To find out how current available tools take into consideration interest of a user on noise web data prior to elimination

Current tools developed to identify and eliminate noise from web pages are mainly based on two different approaches i.e. the underlying structure of document object model (DOM) tree [1], [23] while the other approach solely depends on the visual layout of web pages [24]. DOM tree is a data structure used to represent the layout of a web page, it is build using web page's html parser from which a web content structure is created represent areas of a website with relevant and noise data [11], [25]. For example, [1] proposed a tool known as Site Style Tree (SST) to detect and eliminate noise data from web pages 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 map the page to the main web page to determine if the page is useful or noise based on its presentation style. [26] developed a tool based on case base reasoning (CBR) and neural network to eliminate noise data from web pages. CBR is a machine learning approach which makes use of past experiences to solve future problems i.e. detects noise from web pages using existing stored noise web data. Different noise patterns in websites are stored in form of DOM tree, the case base is then searched for similar existing noise pattern. 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 content 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 quality of data is determined using case based approach then the output will be misleading. [27] proposed pattern tree algorithm to eliminate noise data from web pages, pattern tree algorithm is based on 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. [25] applied Least Recently Used paging algorithm (LRU) 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 frequently visited pages and those that are not been visited over a certain period.

In this paper, we do not focus on structure and layout of web data to identify and eliminate noise but instead we focus on interest of a user to available web data. However, the above tools play a significant role in defining user interest level on web log data. Reduction of noise from web pages based on the structure of data on a given website will subsequently affect the quality of extracted web log data. In this paper, we argue that noise in web data should be determine based on web user's level of interest on available data. In order to understand user interest from extracted web log data, pre-processing is an essential process for purpose of improving quality of output from web usage mining process [28]. Current research works have applied existing tools to determine interest of a user from extracted web data logs. For example, [29] applied Naïve Bayesian classification algorithm to identify interested users based on web log data extracted from a website. This authors' main objective was to classify extracted web data logs and study how useful is the extracted information based on a user interest. Their 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 number of pages viewed and time taken on a specific page. Neural network (NN) is another tool widely applied in web usage mining to reduce noise from web log. [30] state that neural networks use frequency of a web page in web log data to determine its weight. The authors define weight as a statistical measure used to evaluate how important is the information to a given user. [15] applied kNN on web data logs to find information

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

of a user interest by removing noise data, their main focus was on local noise for example advertisements, banners, navigational links etc. Web log data was extracted and surveyed to which web server they belong. If the address belongs to a list of already defined advertisement server, then the link is removed. [31] proposed Weighted Association Rule Mining 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 user interest based page weight is used to eliminate noise web pages from useful information. In their research work, weight of a web page to a user interest is estimated with the frequency of page visit and number of page visited. Where pages visited only once by only one user, they will be assigned low weights and subsequently considered noise.

This work considers noise removal of noise and extraction of useful information from web log data as critical aspects to consider in improving quality of a web user profile. Our justification to this claim is that if noise in web data is not clearly defined and analysed through learning, the purpose and use of data extracted will be compromised hence its overall quality. This paper outlines some of the major contributions made by currently available noise web data reduction tools. For example

- Existing tools applied in noise web data reduction process focus mainly on improving quality of data on web pages by removing noise at pre-processing stage.
- Automatically detecting and removing noise data by matching noise data in current web data with the previously stored noise web data patterns. The contribution is not only to remove multiple noise patterns in a web page but also to enable classification of noise based on defined patterns.
- Protecting useful data regions by identifying boundaries between noise and useful data [21], [32]–[34]. This is due to their assumption that the main web data contains only useful information.

Despite efforts from current research work to address problems with noise in extracted web data logs, this work identify a research gap that has not been fully addressed in relation to noise web data reduction with a view of creating a dynamic and interest specific web user profile. This work argues that eliminating noise from extracted web data logs should be more user interest specific rather than determined by the relationship of data to the web page it resides. In our next section, we establish critical aspects that justify our position based on current research works contribution.

III OUR POSITION

In this section, we conclude that the following points justify our position for the proposed research work:

- It is widely acknowledged by current research works discussed in section II that eliminating noise from web page mainly focus on how relevant is web data to a website as opposed to how relevant is the data to targeted web users.
- In current research works for example, [21], [26], [34] noise in web pages is eliminated based on pre-existing noise data patterns. The problem with this concept is that interest of a user to web data is dynamic and so is the data on the web.
- Even though there are current available tools and techniques which address noise data reduction from web pages, to the best of our knowledge there are no discussions on how loss of useful information can be reduced by learning noise in web log data prior to elimination.

IV CONCLUSION AND FUTURE WORK

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. 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.

ACKNOWLEGMENT

Special thanks to University of Sunderland, Faculty of Computer Science for the continued support in my research progress and development as a researcher.

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.

[8] S. Kanoje, S. Girase, and D. Mukhopadhyay, “User profiling trends, techniques and applications,” *ArXiv Prepr. ArXiv150307474*, 2015.

[9] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, “Enhancing data analysis with noise removal,” *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 3, pp. 304–319, Mar. 2006.

[10] 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.

[11] 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.
[12] X. Wu and X. Zhu, "Mining with Noise Knowledge: Error-Aware Data Mining," *IEEE Trans. Syst. Man Cybern. - Part Syst. Hum.*, vol. 38, no. 4, pp. 917–932, Jul. 2008.
[13] B. H. Kang and Y. S. Kim, "Noise elimination from the web documents by using URL paths and information redundancy," 2006.
[14] R. K. Lomotey and R. Deters, "Analytics-as-a-service framework for terms association mining in unstructured data," *Int. J. Bus. Process Integr. Manag.*, vol. 7, no. 1, pp. 49–61, Jan. 2014.
[15] H. K. Azad, R. Raj, R. Kumar, H. Ranjan, K. Abhishek, and M. P. Singh, "Removal of Noisy Information in Web Pages," 2014, pp. 1–5.
[16] S. Lingwal, "Noise Reduction and Content Retrieval from Web Pages," *Int. J. Comput. Appl.*, vol. 73, no. 4, 2013.
[17] T. T. Aye, "Web log cleaning for mining of web usage patterns," in *2011 3rd International Conference on Computer Research and Development*, 2011, vol. 2, pp. 490–494.
[18] S. Kankane and V. Garg, "A Survey Paper on: Frequent Pattern Analysis Algorithm from the Web Log Data," *Int. J. Comput. Appl.*, vol. 119, no. 13, 2015.
[19] Thakur, A. S and Richhariya, V., *Line Up: A Technique for Semantic-Synaptic Web Entropy Visualization*, 2 vols. International Journal of Modern Engineering & Management Research Technology, 2015.
[20] E. S. Laber *et al.*, "A fast and simple method for extracting relevant content from news webpages," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 1685–1688.
[21] 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.
[22] H. F. Eldirdiery and A. H. Ahmed, "Detecting and Removing Noisy Data on Web Document using Text Density Approach," *Int. J. Comput. Appl.*, vol. 112, no. 5, 2015.
[23] M. John and J. S. Jayasudha, "Methods for Removing Noise from Web Pages: A Review," 2016.
[24] M. E. Akpınar and Y. Yesilada, "Vision Based Page Segmentation Algorithm: Extended and Perceived Success," in *Revised Selected Papers of the ICWE 2013 International Workshops on Current Trends in Web Engineering - Volume 8295*, New York, NY, USA, 2013, pp. 238–252.
[25] A. Garg and B. Kaur, "Enhancing Performance of Web Page by Removing Noises using LRU," *Int. J. Comput. Appl.*, vol. 103, no. 6, 2014.
[26] 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.
[27] 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.
[28] P. Nithya and P. Sumathi, "Novel pre-processing technique for web log mining by removing global noise and web robots," in *2012 NATIONAL CONFERENCE ON COMPUTING AND COMMUNICATION SYSTEMS*, 2012, pp. 1–5.
[29] A. K. Santra and S. Jayasudha, "Classification of web log data to identify interested users using Naïve Bayesian classification," *Int. J. Comput. Sci. Issues*, vol. 9, no. 1, pp. 381–387, 2012.
[30] J. Sripriya and E. S. Samundeeswari, "Comparison of Neural Networks and Support Vector Machines using PCA and ICA for Feature Reduction," *Int. J. Comput. Appl.*, vol. 40, no. 16, pp. 31–36, Feb. 2012.
[31] S. P. Malarvizhi and B. Sathiyabhama, "Enhanced reconfigurable weighted association rule mining for frequent patterns of web logs," *Int. J. Comput.*, vol. 13, no. 2, pp. 97–105, 2014.
[32] X. Wang, B. Chen, and F. Chang, "A Classification Algorithm for Noisy Data Streams with Concept-Drifting," *J. Comput. Inf. Syst.*, vol. 7, no. 12, pp. 4392–4399, 2011.
[33] H. Wang, Q. Xu, and L. Zhou, "Deep Web Search Interface Identification: A Semi-Supervised Ensemble Approach," *Information*, vol. 5, no. 4, pp. 634–651, Dec. 2014.
[34] F. Hu, M. Li, Y. N. Zhang, T. Peng, and Y. Lei, "A Non-Template Approach to Purify Web Pages Based on Word Density," in *Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012*, Springer, London, 2013, pp. 221–228.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)