



**University of
Sunderland**

Onyancha, Julius (2019) Learning Noise Web Data Prior to Elimination: Classification of Dynamic Web User Interests. Doctoral thesis, University of Sunderland.

Downloaded from: <http://sure.sunderland.ac.uk/id/eprint/13533/>

Usage guidelines

Please refer to the usage guidelines at <http://sure.sunderland.ac.uk/policies.html> or alternatively contact sure@sunderland.ac.uk.

Learning Noise Web Data Prior to Elimination: Classification of Dynamic Web User Interests

By

Julius Onyancha

A Thesis Submitted in Partial Fulfilment of the
Requirements of the University of Sunderland
for the Degree of Doctor of Philosophy

Jan 2019

Acknowledgement

First and foremost, I thank Almighty God for his guidance all through. My gratitude goes out to my family who sacrificed all they had to support me financially and morally. My parents Mr Zablon and Mrs Eunice Onyancha, my brothers and sisters, I'm indebted to you forever and may God bless you, because of your continuous prayers and patience, I found peace and confidence that enabled me to keep pushing.

Secondly, I wish to acknowledge the support extended to me by my Director of Studies Dr Valentina Plekhanova. Her dedication and motivation kept me going throughout the PhD journey. Through her mentorship, I have developed myself as a young researcher, I will ever be grateful. I also wish to thank my co-supervisor Dr David Nelson for his tremendous support throughout, your feedback was invaluable.

Special thanks to the Faculty of Computer Science and the University of Sunderland Library for financially supporting my research publications; the university has been my home since my undergraduate. Finally, I would like to thank my friends who have kept up with me when they felt I was no longer of value to them, I appreciate you were kind enough to accomodate my awkward silence.

Abstract

The amount of noise in web data is rapidly increasing, as is the number of users searching for information to suit their interest. The increase of web data has led to some critical issues, such as a high level of noise and irrelevant data. Given that the web is noisy, inconsistent and irrelevant by nature, finding useful information that defines interest of user has become a challenge. Existing research acknowledges that there is a need to propose machine learning tools capable of addressing problems with noise web data. Identifying and eliminating noise web data is critical to the web usage mining process. As the web evolves and more web data sources emerge, the level of noisiness also increases.

Despite efforts by existing research to address noise in web data, a number of critical issues remain unresolved. For example, existing research work considers noise web data as irrelevant data that does not form part of the main content of a web page. Therefore, current machine learning tools focus on protecting the main content of a web page by eliminating noise/irrelevant data, such as advertisements, banners and external links etc. However, the main content of a web page can potentially be noise when user interests are considered. The position taken by the proposed research is based on the fact that noise web data can itself be useful when the interests of a user are considered prior to elimination.

To justify this position, a Noise Web Data Learning (NWDL) approach which aims to learn noise web data prior to elimination is proposed. To the best of our knowledge, learning noise web data prior to elimination has not been addressed by current and relevant research works. The objective is to ensure that the interestingness of data on the web is defined by user interests over time. The proposed NWDL considers the following key aspects, 1) the significance of exit page in defining user interest level on web pages visited by a user. 2) The effect of the dynamic change of user interests towards the classification of web pages.

Experiments conducted in this research shows that noise web data reduction process is user-centric, i.e., the dynamic changes of user interests influence

the interestingness of web data. As a result, what is currently identified and eliminated as noise can be useful when user interests and their changes over time are considered. The findings from validation and evaluation of the proposed NWDL against existing tools shows that user interest over time significantly impacts the importance of data on the web. Given that current research work mainly identifies and eliminates noise web data without defining user interests, the process is not user-centric. A key contribution of the proposed research is to identify and learn noise web data taking into account user interests as they change over time prior to elimination. Ultimately the proposed NWDL contributes towards minimising loss of useful information otherwise considered as noise by the existing tool as well as reduce the level of noise data suggested to a user.

Disseminations based on this thesis

Book Chapter

Onyancha J., Plekhanova V., Nelson D. (2019), "Learning Noise in Web Data Prior to Elimination," In: Ao Sl., Gelman L., Kim H. (eds) Transactions on Engineering Technologies, WCE 2017, Springer, pp.177-187

Journal Article

Onyancha J. V. Plekhanova V., (2018) "Noise Reduction in Web Data: A Learning Approach Based on Dynamic User Interests," International Journal of Computer and Information Engineering, Vol: 12, No: 1, pp 7-14

Published Conferences Papers

Onyancha J., Plekhanova V., Nelson D. (2017) "Noise Web Data Learning from a Web User Profile: Position Paper," in Proceedings of the World Congress on Engineering, 2017, vol. 2, pp 608-611

Onyancha J., V. Plekhanova V., (2017), "A User-Centric Approach towards Learning Noise in Web Data," 12th International Conference on Intelligent Systems and Knowledge Engineering, IEEE 2017, Nanjing, Jiangsu, 2017, pp. 1-6.

Accepted Conference Papers

Onyancha J., Plekhanova V., Nelson D. (2018), "Effect of Dynamic Thresholds Values on Interestingness of Noise Web Data", International Conference on Information Science and System (ICISS 2018), Jeju Island, South Korea, April 27-29, 2018, Proceedings Series by ACM

J Onyancha J., Plekhanova V., Nelson D. (2018), "Interestingness of Web Data Based on Dynamic Change of User Interests," International Conference on Machine Learning and Data Analysis 2018, San Francisco, USA, 23-25 October 2018

Presentations

"Noisy Web Data Learning (NWDL)," MSc. Information Technology Management Seminar, University of Sunderland, 24 January 2017.

"Problems with Noise in Web Data." MSc. Data Science Seminar, University of Sunderland, 12 April 2018.

Posters

"Noise Web Data Learning (NWDL) - A User-Centric Approach," University of Sunderland Research Conference, 5th June 2017

"Learning Noise in Web Data – A User Centric Approach," The 12th International Conference on Intelligent Systems and Knowledge Engineering," November 24–26, 2017, Nanjing, China.

Table of Contents

Acknowledgment	i
Abstract	ii
Disseminations based on this thesis	iv
Chapter 1: The Research Background	1
1.1. Introduction	1
1.2. Research Motivation	2
1.3. Problem Statement	2
1.4. Proposed Research Position	4
1.5. Research Aims and Objectives	4
1.6. Research Questions	5
1.7. Proposed Research Contributions to Knowledge	6
1.8. Thesis Structure	7
Chapter 2: Noise Web Data – A Literature Review	10
2.1. Introduction	10
2.2. Noise Data – Definition and Relevant Research Issues/Theory	11
2.2.1. The context of web page content in the noise elimination process	14
2.2.2. Common Sources of Noise in Web Data	16
2.2.3. Different Types of Noise in Web Data	17
2.2.4. The Impact of Noise Web Data in the Web Usage Mining Process ...	17
2.3. Critical Analysis and Evaluation of Relevant Research Works	19
2.3.1. Noise Web Data Reduction: Layout and Structure of Web Data	19
2.3.2. Noise Web Data Reduction: Extracted Web Data Logs	25
2.4. The Influence of User Interest on the Noise Web Data Reduction Process	26
2.5. Discussion of Critical Aspects	27
2.6. Chapter Summary	29
Chapter 3: Web User Profiling Based on Web Data	31
3.1. Introduction	31
3.2. Collection of User Interest Information	33
3.2.1. Explicit User Interest Information	34
3.2.2. Implicit User Interest Information	35
3.3. Extraction and Pre-processing of Web Log Files	36
3.3.1. User Identification	39
3.3.2. User Session Identification.....	41
3.3.3. Page View Identification	47

3.4. Web User Profile Construction	48
3.5. Web User Profiling: its Significance in Noise Web Data Reduction ...	49
3.6. Chapter Summary.....	50
Chapter 4: Learning Noise in Web Data	52
4.1. Introduction.....	52
4.2. User Interest Learning.....	53
4.2.1. Identifying Key Indicators for Learning User Interest.....	54
4.2.2. Interestingness of a web page based on time and frequency of user visits	59
4.2.3. Influence of Time and Frequency of Web Page Visits on Noise Elimination	60
4.2.4. Learning User Interest Based on Depth of User Visit.....	62
4.2.5. Interestingness of a Web Page Based on User Interest Category	66
4.3. Addressing Dynamic Change in User Interests	73
4.3.1. Interestingness of a Web Page Based on Recency of Visit.....	74
4.3.2. Dynamic Threshold Values	75
4.4. Learning Noise Web Data by Classification	78
4.5. Noise Web Data Learning: its Significance to Web Usage Mining.....	81
4.6. Chapter Summary.....	82
Chapter 5: Experimental Design Setup	85
5.1. Introduction.....	86
5.2. Experimental Data Preparation.....	87
5.3. Experimental Setup.....	88
5.3.1. Interestingness of web page based on exit page user visit duration..	89
5.3.2. Interestingness of Web Data Based On Dynamic Change of User Interest	92
5.3.3. The overall performance of the proposed noise web data learning approach	96
5.4. Chapter Summary.....	99
Chapter 6: Evaluating Performance of Proposed NWDL	100
6.1. Introduction.....	100
6.2. Evaluation Metrics.....	102
6.2.1. Confusion Matrix	102
6.2.2. Accuracy/Error rate.....	103
6.2.3. Precision, Recall and F-Measure	104
6.3. First Direction – Black-box Validation Process.....	106
6.3.1. Evaluating Performance of NWDL using a ‘Black Box’ Approach	106

6.3.2. Discussion of the results.....	107
6.4. Second Direction- Baseline Model.....	110
6.4.1. Evaluating the performance of NWDL against the Baseline Model ..	113
6.4.2. Discussion of the results.....	114
6.5. Evaluating performance of the proposed NWDL approach using a noise dataset	116
6.5.1. Validation Process.....	116
6.5.2. Discussion of the results.....	117
6.6. Evaluating performance of the proposed NWDL using Open Source Dataset	118
6.7. Discussion of critical aspects based on performance of NWDL	121
6.8. Chapter Summary.....	122
Chapter 7: Conclusion and Future Work.....	123
7.1. Introduction.....	123
7.2. Critical Discussions Based on Research Objectives.....	123
7.3. Key Findings Based on the Proposed Research Questions.....	124
7.4. Research Contribution	127
7.5. Future Work.....	128
References	131
Appendices	145
Appendix I: Samples of noise web data	145
Appendix II: Comparative analysis of data mining techniques applied in noise web data reduction	146
Appendix III: Data Modelling: Use Case Diagram	148
Appendix IV: Sample raw datasets	149
Appendix V: Experimental Design Procedures and Validation.....	150

List of Algorithms

Algorithm 1: User Identification.....	40
Algorithm 2. Session Identification based on dynamic time-out.....	47
Algorithm 3. Depth of User Visit	64
Algorithm 4. Learning noise web data.....	81

List of figures

Figure 1.1:Thesis Structure.....	8
Figure 2.1: Literature review structure.....	11
Figure 2.2 Different types of Noisy Data.....	12
Figure 2.3: A web page containing noise data.....	15
Figure 2.4: Relationship between a user and web data	24
Figure 3.1: Extraction and pre-processing of web log files	33
Figure 3.2: Explicit user feedback.....	34
Figure 3.3: Session identification based on 30 min threshold value.....	42
Figure 3.4: User profile construction.....	49
Figure 4.1: Noise web data learning process flow diagram	54
Figure 4.2: Depth of a user visit to a website	63
Figure 4.3: Categories of web pages.....	67
Figure 4.4: Frequency of visits to a web page category	69
Figure 4.5: Frequency versus visit to a web page category	72
Figure 5.1: Database Model.....	88
Figure 5.2: Page visit duration based on fixed vs dynamic time-out session.....	90
Figure 5.3: Comparing page weight based on fixed vs dynamic time-out session ..	91
Figure 5.4: Dynamic change of user interest over 90 days' period	93
Figure 5.5a: User interest level after one week	95
Figure 5.5b: User interest level after 7 weeks.....	95
Figure 5.6: Noise Web Data Learning Process Flow Diagram.....	97
Figure 5.7: Overall Performance of NWDL.....	98
Figure 6.1: Black-Box Validation Process.....	106
Figure 6.2: The output from black-box approach.....	107
Figure 6.3: The output from the black-box approach	108
Figure 6.4: Noise output from black-box approach	108
Figure 6.5: Interest output from black-box approach	109
Figure 6.6: Defining a Baseline Model	111
Figure 6.8: Baseline vs NWDL – A validation process.	114
Figure 6.9: Performance evaluation using Confusion Matrix.....	114
Figure 6.10: Performance evaluation in noise data set	118
Figure 6.11: Classification Performance of the Baseline Model.....	119
Figure 6.12: Classification Performance of the Proposed NWDL.....	120

List of Tables

Table 1: Raw web log file	38
Table 2: A set of records in the j th user profile	40
Table 3: Session identification <i>for the jth</i> user	44
Table 4: Visit duration for j th user on k th web page in i th session	56
Table 5: Time Duration versus Frequency of User Visit.....	61
Table 6: Length of visit to a web page category	71

Abbreviations

WWW:	World Wide Web
NWDL:	Noise Web Data Learning
WUM:	Web Usage Mining
ANN:	Artificial Neural Networks
k-NN:	k Nearest Neighbours
CBS:	Case Based Reasoning
NB:	Naïve Bayes
ML:	Machine Learning
DOM:	Document Object Model
LRU:	Least Recently Used
AUC:	Area Under Curve
RF:	Random Forest
CA:	Classification Accuracy

Chapter 1: The Research Background

1.1. Introduction

The World Wide Web (WWW) has in the recent past emerged as the main source of information, but its rapid growth has also made it more difficult for web users to actually find useful information (Gao et al., 2016; Varnagar et al., 2013; Dohare et al., 2012). The process of extracting information from the web that meets the needs of a web user, referred to as web usage mining (WUM), has thus become a popular research area (Jafari et al., 2013; Srivastava et al., 2000). However, the level of noise data on the web is rapidly increasing, making it difficult to find useful information for a given user at any time. In order to improve the web usage mining process, a number of machine learning (ML) algorithms are proposed by the current research (Htwe et al., 2011; Kabir et al., 2012; Dutta et al., 2014; Lopes and Roy, 2015; Sirsat and Chavan, 2016).

Web data is noisy, inconsistent and often irrelevant by nature (Singla and White, 2010; Dwivedi and Rawat, 2015). The presence of noise web data hinders the discovery of relevant and useful information in relation to the given user interests (Jafari et al., 2013; Xiong et al., 2006). Lingwal (2013) and Yi et al (2003) define relevant data as the main content of the web page that a user needs to view. Content pages are web pages where a user can find useful information, while anything that does not form part of the main web page is considered noise (Kapusta et al., 2012). Noise is defined as irrelevant data that is not part of the main content of the web page (Yi et al., 2003; Laber et al., 2009; Lingwal et al., 2013; Bhamare and Pawar, 2013). Problems with noise web data have been explored by Goel (2014), Dutta et al. (2014) and Aldekhail (2016); such problems include the fact that data available on the web comes from different sources. Given that there are few measures in place to address the noise levels in web data, web users can write and post anything, thereby subjecting web data to low quality, erroneous or even misleading content (Srivastava et al., 2000; Jafari et al., 2013). Moreover, identifying information considered useful, as opposed to noise, for a specific

user varies from one person to another as not all web data are of equal interest to every user, and this also varies over time because user interests change.

1.2. Research Motivation

Existing research studies have made an effort to address problems associated with noise web data; for example, Yi et al (2003), Lingwal (2013) and Sivakumar (2015) propose machine learning tools that can identify and eliminate noise in web data based on the structure and layout of web pages. Determining noise based on the structure and layout of web data, however, ignores the fact that the main content within a web page can be noise if it does not meet the interests of a user. Htwe and Kham (2011) and Velloso and Dorneles (2013) propose a mechanism whereby the noise associated with web pages is matched to stored noise data for classification and subsequent elimination. As a result, the elimination of noise in web data is based on pre-existing noise data patterns. It is thus evident that current research concentrates on the structure and content of the web pages as a means of determining the usefulness of web data, under the assumption that the main web page content contains useful information (Yi et al., 2003; Ansari et al., 2011; Htwe and Kham, 2011). However, the proposed research argues that noise is not necessarily data that does not form part of the main content of a web page; instead, the main web page content can also be noise if it does not reflect the user's interest at the given time.

1.3. Problem Statement

Discovery of useful information from the noisy web is an area that has received extensive consideration in the existing research, as mentioned in previous sections of this chapter. Despite the contributions made by the existing research to address problems with noise web data, there are still critical issues that have not thus far been fully addressed. The existing research works that address problems with web data typically focus on identification and classification of web data, but it is unclear if they consider user interests prior to elimination of noise in web data. For example, Kakol et al. (2017) argue that the relationship between main content of a web page can have a substantial effect on identifying useful information from the noisy web. However, as user interests change over time, the interestingness of web data is also affected,

thus making it difficult to identify and eliminate noise in web data. The interestingness of web data is defined either subjectively or objectively (Huang et al., 2002; Dimitrijevic et al., 2014). For instance, objective interestingness depends on the structure of data and the patterns extracted from it, while subjective interestingness relies on the specific needs and interests of a user (Sahar, 2010; Jiang et al., 2013). Pazzani and Billsus (1997), Cooley et al. (1999) and Dimitrijevic et al. (2014) further define interestingness of a web page as the relevance of the page with respect to the interests of a user. The interestingness of web data may vary in line with different levels of user interests. This is due to the fact that information on the web can be useful only in some respects or for a certain period, which is defined by a sequence of user events that happen over time. For example, a user interest in shopping for a wedding is regarded as an event. Therefore, learning the interestingness of web data is critical to identification of information from the web that reflects the interests of a user. This is to ensure that the web data identified and eliminated as noise is first determined to be against the interests of the user.

Noise web data elimination is a research area that cannot be investigated independent of user interests. This is because identification and elimination of noise web data is dependent on user needs and interests. The proposed research focus is to learn about noise in web data based on user need and interests. For example, what is noise to one user can be interesting to another, and current user interest data can become noise in the future. It is therefore important to note that as the web evolves, user interests change as well, and such changes influence the process of eliminating noise in web data. Changing user interests opens up challenges in determining whether the main web content itself is useful or noise for a given user at a given time. The dynamic aspects of user interest are critical to the identification and subsequent elimination of noise web data. It is therefore necessary to determine noise web data by considering a change of user interests. Given the current state of the web, eliminating noise in web data prior to learning its interestingness to a user can lead to the loss of useful information (Onyancha et al., 2017).

1.4. Proposed Research Position

The proposed work is inspired by the existing research's effort to address problems with noise web data. A number of critical issues discussed in sections 1.2 and 1.3 of this chapter define the gap the proposed research attempts to address. For example, the existing research on noise elimination from web data fails to acknowledge that as user interest change over time, the interestingness of web data changes as well. In addition, the main content of a web page is currently perceived to hold useful information, but in reality this does not necessarily mean it addresses the user interests. These are some of the critical issues the proposed research aims to address in order to ensure user interests play a part in eliminating noise in web data.

With the rapid growth of data volume on the web, finding information that is interesting to the user is becoming even more challenging due to the level of noisy data. The proposed research revisits this problem from a user interest perspective. Specifically, it focuses on avoiding the elimination of useful information otherwise identified as noise by existing tools. Rather than just eliminating noise in web data, the proposed approach will learn about the interestingness of web data, taking into account user interests prior to noise elimination. Even though the structural layout of data on a web page creates a boundary between useful and noisy data (Velloso and Dorneles, 2013), the proposed research's viewpoint is that **"noise in web data can be useful if changes in user interests are considered prior to elimination"**. Therefore, this research challenges the assertion of existing research on noise web data that: (1) the main content of the web contains useful information, while the rest is noise; and (2) noise web data is data that does not form the main content of a web page.

1.5. Research Aims and Objectives

The aim of this research is to explore how current research addresses problems with noise web data and to identify critical issues in relation to noise in web data that current research works have not managed to fully address. The limitations with the existing tools, as discussed in section 1.3, highlight problems that need to be addressed given the current state of the web. The proposed research work aims to ensure that the process of finding and

eliminating noise web data is conducted in a way that allows user interests to play a central role. In order to fulfil the proposed research aims, the following objectives are considered:

1. Identify and critically evaluate current research that addresses noise in web data. This includes exploring the types of noise data eliminated, problems addressed in relation to noise data eliminated, contributions and limitations of the current tools used.
2. Identify gaps in current research based on the performance of existing tools to existing problems. The outcome aid in positioning the proposed research and its attempts to address the gap.
3. Determine how changes in user interests impact identification and subsequent elimination of noise in web data.
4. Propose a noise web data learning approach that can identify and learn noise in web data prior to elimination.
5. Validate and evaluate the performance of the proposed algorithms against existing tools applied in the noise web data elimination process.
6. Identify and evaluate key findings from the proposed research work against research objectives.

1.6. Research Questions

This section identifies the research questions addressed by this thesis. The objectives outlined in the previous section are expressed in the following research questions:

Question 1: In what ways do current research works define and address noise in web data?

This question is answered through a critical review and evaluation of current research work in chapter 2 of this thesis. The existing research work is evaluated in terms of how noise web data is defined, machine learning tools proposed to address noise data, measures applied to evaluate performance of tools proposed by existing research and their contributions to address problems with noise in web data.

Question 2: What are the key indicators for learning user interests and how could the interests of a user influence identification of noise web data?

This question is answered through a critical analysis and evaluation of measures used to learn user interests on the web. A number of key process and measures used to determine user interest levels on the web are explored in chapters 3 and 4.

Question 3: How can learning noise web data better address problems with the noisy web in comparison to contributions made by the existing research?

Chapters 5 and 6 introduce a number of experiments using the proposed machine learning algorithms. In order to verify whether the proposed noise web data learning approach performs better than existing tools, the proposed algorithms are validated by comparing the results using key measurement metrics such as confusion matrix, precision, recall and F-measure. In order to ensure the proposed research objective is achieved, the overall performance of the proposed algorithms is evaluated against the research objectives.

1.7. Proposed Research Contributions to Knowledge

Based on the existing literature, the proposed research explores how the relevant current research studies address problems associated with noise web data. Noise web data is currently defined based on structure and layout of web data. Therefore, interestingness of web data depends on the structure of data and the patterns extracted from it. To the best of our knowledge, existing research do not clearly define noise web data taking into account change user interest. Moreover, they do not learn noise web data based on change of user interest prior to elimination. Without learning the interestingness of web data based on a user's interest, the process of eliminating noise web data is limited to simply recognising how web data is presented and not what users are interested in at any given time.

To justify this position, a noise web data learning (NWDL) approach is proposed, this approach aims to learn noise web data prior to elimination. NWDL approach introduces a number of measures capable of learning interestingness of web data prior noise identification. The proposed measures consider the following key aspects, 1) the influence of exit page in defining interestingness of web page visited by a user. 2) The effect of dynamic change of user interests towards the classification of web pages.

A key contribution of the proposed research is to identify and learn noise web data taking into account user interests as they change over a time prior to elimination. Ultimately, the proposed NWDL contributes towards minimising loss of useful information otherwise considered as noise by existing tool as well as reduce level of noise data suggested to a user.

1.8. Thesis Structure

The thesis is divided into seven chapters, as presented in Figure 1. These include the definition of the research problem; rationale and contribution to existing research, as discussed in the previous chapter; a critical review and evaluation of current research work; the positioning of the proposed research's methodological approach used to address the defined research problems, experimental design and validation process; and finally evaluation of the research outcomes against the defined objectives and conclusion.

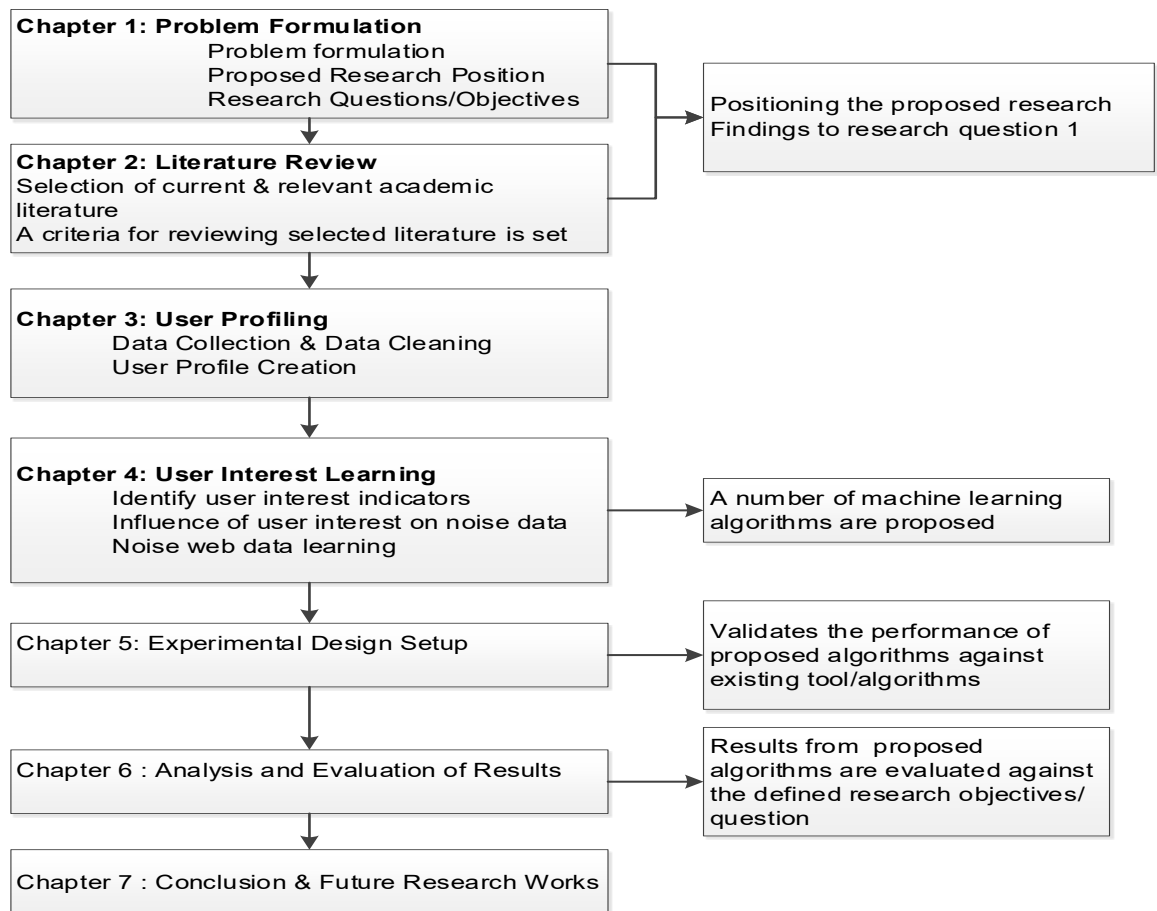


Figure 1.1: Thesis Structure

Chapter 2 is a critical review and evaluation of the existing researches that address problems with noise web data. This chapter further justifies the proposed research’s position and the gap it attempts to fill.

Chapter 3 presents a methodological approach to address the defined research problem. The chapter first defines a user profile that captures the interests of a user on a web page. The key aspects of user profiling are: understanding varying user interests on the web and learning how changes in user interests influence interestingness of web data. The key stages explored in this chapter include: data extraction from web servers, pre-processing of extracted web log data, and finally building a user profile based on the extracted web log data.

Chapter 4 proposes a noise web data learning approach that takes into account a change in user interests prior to noise elimination. In this chapter, a number of algorithms are defined to learn the interestingness of web data

when user interests are considered. The defined algorithms are based on a number of user interest indicators: duration, frequency and depth of a user visit to a web page. This chapter aims to demonstrate how user interests and the changes in this influence identification and subsequent elimination of noise web data.

In **Chapter 5**, a number of experiments are conducted based on proposed algorithms in order to validate the performance of the proposed algorithms. The objective of conducting experiments is to determine how the proposed machine learning algorithms perform over existing tools applied to address noise web data. The experiments are designed in such a way as to verify whether the proposed noise web data learning approach minimises loss of useful information, reduces noise in web data and considers change of user interests prior to elimination of noise in web data.

Chapter 6 is based on the results from the experiments conducted; the performance of the proposed algorithms is validated against existing tools that address noise web data. The goal of the validation process is to demonstrate that the proposed noise web data learning approach produces better results than existing tools. The outcome of the validation process aims to respond to the third research question: *How can learning noise web data better address problems with the noisy web in comparison to contributions made by the existing research?*

Chapter 7 provides a summary of the critical issues addressed in relation to the defined research problem and objectives. A discussion of the contribution made and further research work in the noisy web is also presented.

Chapter 2: Noise Web Data – A Literature Review

2.1. Introduction

Chapter 1 introduces the proposed research by examining problems with noise web data. It explores the contribution made by existing research to address the defined problems and, as a result, a number of critical issues in relation to the problem are identified. A gap in existing research has thus been defined that precipitates the need to propose an approach to address challenges with noise web data. In order to justify the rationale of the proposed research, it is necessary to understand how noise in web data is addressed by current research, the machine learning tools used by current research to address existing problems, and the measures used to evaluate the performance of the existing tools.

This chapter provides a critical analysis and evaluation of current research on eliminating noise web data. The main objective is to find out how existing research identifies and addresses noise web data. The outcome of this chapter aids in identifying the gap in the literature that the proposed research attempts to address. Overall, this chapter answers the following question:

Question 1: In what ways do current research works define and address noise in web data?

The chapter comprises the following sections: **Section 2.2** describes relevant research issues and theories associated with noise web data. For example, an understanding of how noise web data is currently defined, its source and how it affects the process of finding useful information from the web. **Section 2.3** provides a critical analysis and evaluation of the relevant and most current research work in the noise web data reduction process. A review of existing research works aids in understanding the problems with noise web data, the methodology used to address such problems, the contributions and limitations. **Section 2.4** explores how user interests impact the noise web data reduction process. This section ascertains whether existing research considers the influence of user interest when identifying and eliminating noise

web data. **Section 2.5** presents a discussion of critical aspects in relation to the current approach to noise web data reduction and dynamic change in user interest with regard to web data. This will identify the gap in the literature and position the proposed research. Finally, **Section 2.6** summarises the entire chapter. The literature review structure is outlined in **Figure 2.1**.

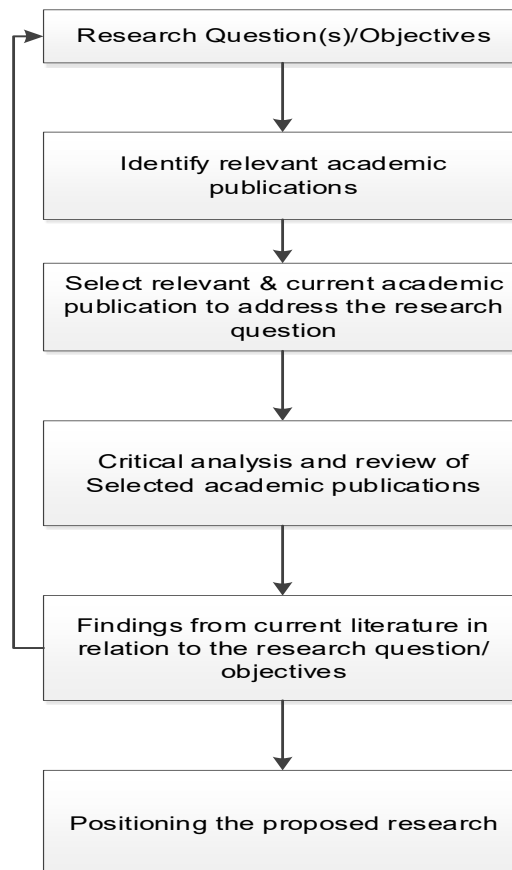


Figure 2.1: Literature review structure

2.2. Noise Data – Definition and Relevant Research Issues/Theory

Noise data is a broad term that has been interpreted and applied in various different ways by current research. In data mining, several definitions have been proposed; for example, Yang and Fong (2011) define noise data as irrelevant or meaningless data that does not typically reflect the main trends, but instead makes the identification of these trends more difficult. Sunitha et al. (2013) add that noise data is meaningless or corrupted data, i.e. any data that cannot be understood and interpreted correctly by a machine, or data that is incorrectly typed or dissimilar to other entries.

Presence of noise in data affects the intrinsic characteristics of classification problem (López et al., 2013). For this reason, classification problems as a result of noise data are often difficult to address. Noise data is differentiated into two categories, i.e., attribute and class noise (Zhu and Wu, 2004; Frenay, B. and Verleysen, M., 2014). Attribute noise refers to corruption in the values of one or more attributes (Wu and Zhu, 2008). Examples of attributes noise are erroneous attribute values, missing or unknown attributes values, and incomplete attributes. This type of noise can be tackled from the input end by eliminating data objects that are suspect of noise according to certain evaluation mechanisms. When removing noisy objects from data, there is a trade-off between the amount of information available for building the classifier and the amount of noise retained in the data set. Class noise occurs when a data object is incorrectly labelled as shown in Figure 2.2. This type of noise usually occurs on the boundaries of the classes, where the examples may have similar characteristics. (Sáez, et al 2013) highlight some of the reasons why class noise occur; firstly subjectivity, i.e., the information used to label an object is different from the information which learning algorithm will have access to, secondly inadequacy of the information used to label each data object, for example in medical domain it is difficult to perform diagnosis of 100% accuracy with test data available because it is highly likely that the information available is incomplete or sometimes misleading. Lastly, data entry errors which will affect the meaning of data more especially when devices are used to capture and automatically create data classes (Pechenizkiy, et al., 2010).

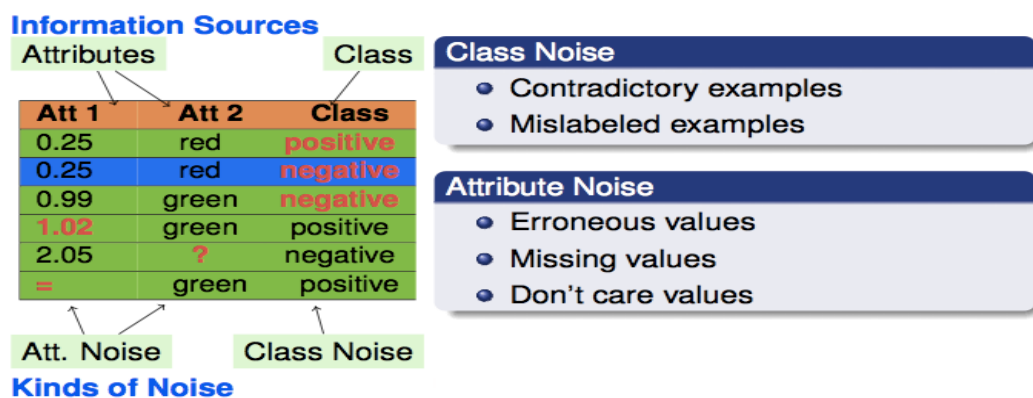


Figure 2.2 Different types of Noisy Data

Source: <https://sci2s.ugr.es/noisydata>

In web data mining, noise is defined as irrelevant data that is not part of the main content of a web page (Yi et al., 2003; Laber et al., 2009; Lingwal et al., 2013; Bhamare and Pawar, 2013). For example, advertisements banners, graphics, web page links from external websites etc., which surround the main content of the web page, as shown in Figure 2.2. Dutta et al. (2014) argue that noise removal from web pages is an important task that helps to extract the actual content for the web usage mining process. Based on this viewpoint, the actual content where useful information resides is protected from noisy data prior to web usage mining. However, the proposed research's viewpoint is that identifying useful information from the noisy web should not only be limited to the main web page content perceived to contain useful information, but instead considered from a user interest perspective. In order to ensure an efficient web usage mining process, it is important to identify and eliminate noise data that may misguide users without loss of any useful information. However, various challenges are encountered in this regard; for example, the dynamic nature of web data makes it difficult to detect noisy data, and various sources of web data are seen to generate voluminous data characterised by noise, thus making it difficult to identify the useful information (Castellano et al., 2009; Cai and Zhu, 2015).

It is widely acknowledged that the web has become the main source of information in the modern world. Indeed, 90% of the data in the world today has been created in the last two years (IBM, 2013). However, as the amount of data on the web increase, the levels of noise data also increase. For this reason, machine learning tools have emerged as critical in the web usage mining process. The main sources of data used in the web usage mining process are web server logs, proxy servers and client or browser logs (Srivastava et al., 2014). Web servers are considered the richest and most common source of data for the web usage mining process (Adeniyi et al., 2016; Kaddu and Kulkarni, 2016; Mobasher et al., 2000; Ramya et al., 2011). The information related to a user request is recorded on the server in a web log file. Log file data represent the fine-grained details of user activities on the web; they also give an idea of what a user is interested in.

One of the key issues in web usage mining is the ability to extract information in relation to the interests expressed by a user. In order to discover useful information, extracted web log data need to be pre-processed. Pre-processing is necessary to identify interesting data patterns in relation to user interest (Srivastava et al., 2000). However, there are a number of challenges that hinder the web usage mining process (Goel, 2014; Mobasher and Nasraoui, 2011). Firstly, the dynamic tendencies of the web, as well as evolving user interests, affect the process of separating which data is useful or not for any given user. Therefore, a critical issue the proposed research aim to address is how change of user interests influence interestingness of web data. Secondly, the web is noisy; data available on the web comes from various different sources. Due to the fact that the web does not have control over the quality of data available, web users can write and post anything, hence subjecting web data to low quality, erroneous or even misleading data (Cai and Zhu, 2015).

Due to the richness and diversity of data available on the web, machine learning algorithms have been proposed to address the above challenges. For example, Dutta et al. (2014), Qi and Sun (2011) and Zhang and Deng (2010) propose noise elimination tools based on the fact that information that does not form part of the main web page contents is noisy and should be eliminated. This thesis recognises that since the data available on the web is heterogeneous and rapidly increasing, integration of user interests and available web data is becoming a challenge. Therefore, prior to determining user interest in extracted web log data, there is a need to identify and learn noisiness in web data in relation to the interests of a user. This is done to ensure that the process of eliminating noise in web data does not lead to a loss of useful information. For example, what is noisy to one user can be useful to another, hence the need to learn and define noise in web data in relation to user interests (Onyancha et al., 2017).

2.2.1. The context of web page content in the noise elimination process

The web page is considered the main source of useful information (Dutta et al., 2014; Kaddu and Kulkarni, 2016). Alongside the main content, however, a web page also comprises noisy parts, such as advertisements that surround

the main content, as shown in **Figure 2.3**. Further samples of noise web data are shown in Appendix I.

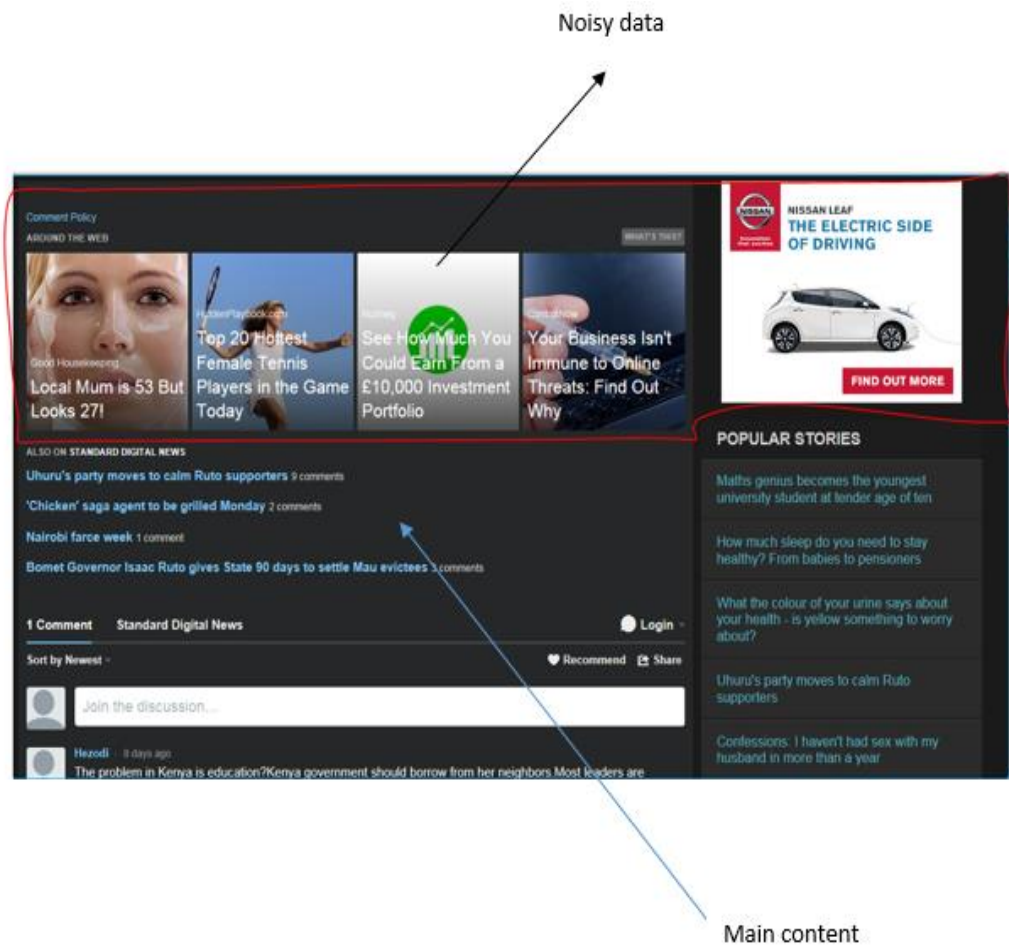


Figure 2.3: A web page containing noise data

Source: <https://www.standardmedia.co.ke/> accessed on 23/02/2015

A number of machine learning tools applied in the noise web data reduction process are based on the followings concepts: (1) web page segmentation, i.e. finding the boundary between the main web page content and noisy data (Hu et al., 2013; Velloso and Dorneles, 2013); and (2) detecting and removing noisy data based on the visual layout of the web page as per the assumption of Htwe and Kham (2011) that relevant information mainly exists in the middle of a web page, while the rest of the page contains noise data. Based on these two critical aspects, web data is considered relevant if it is part of the main content of a web page. However, this thesis argues that the main content can be noise if a user finds the information irrelevant. Therefore, this thesis considers web page content for a very basic reason: web logs contain data

extracted from the main content of a web page. Given that a user is looking for interesting information from the main content, it is important to learn its interestingness in relation to dynamic change in user interests. The existing tools applied to eliminate noise from web pages mainly focus on the web content, but there are no clear discussions how noise is defined in web data in relation to the web user.

2.2.2. Common Sources of Noise in Web Data

Understanding the source of noise in web data is the first step to the noise web data reduction process (Stoica, 2012). Noise in web data comes from two main sources (Liu, 2011; Sáez et al., 2013, 2016). Firstly, a typical web page contains many pieces of information, mainly in form of navigation links. For any particular application, only part of such information is useful, while the rest is considered noise (Liu et al., 2011; Patil, 2012). Secondly, there are no clear measures in place to control the information available on the web; a person can write and post anything on the web, hence subjecting web data to low quality, erroneous or misleading information. Noise data is mainly generated by inaccuracies in data collection, transmission and storage (Garcia et al., 2012; García-Gil et al., 2017). Frenay and Verleysen (2014) add that some noise in web data can simply come from data communication problems, such as spam and accidental clicks. Fan et al. (2014) note that noise occurrence in data is related to the way the data is accessed and pre-processed.

Current research works acknowledge that data is noise-free if it is accurately processed and transmitted (Sáez et al., 2015). However, accuracy can be defined based on a number of aspects and its application domain. For example, Garcia-Gil (2017) suggests that data must be accurate: it must be what it says it is with enough precision to drive value. On the other hand, information that is only relevant to a specific user over a specified period of time can be noise if suggested to a user outside the time of need or use. The proposed research therefore argues that the usefulness of web data should be determined based on a user-centric approach that considers user change of interests, as well as the evolving web. This is due to the fact that the web is not only about data or information, but also about interaction among users (Yazidi and Granmo, 2011). Understanding the sources of noise in web data

opens up a discussion as to why learning noise data prior to elimination is critical. This is precisely the proposed research's focus: it makes a significant contribution to reducing noise levels with minimal loss of information relevant to the given user's interest at the time.

2.2.3. Different Types of Noise in Web Data

The concept of noise in web data has been defined and discussed by various research works from different application domains. As defined in section 2.2.2, noise web data is mainly external data that does not make up the main web page content. Noise web data is categorised as global and local noise (Liu et al., 2004; Nithya and Sumathi, 2012; John and Jayasudha, 2016). Global noise is redundant web pages over the internet, such as mirror sites and illegal or legal duplicated web pages. Local noise is any irrelevant, incoherent data in the main content of a web page. Global noise is caused by duplicated web pages, whereas local noise is caused by irrelevant content on a web page, such as advertisements, navigational panels, announcements, etc.

The existing machine learning tools applied in local noise elimination have been developed on the basis of the web content – the main web page content being considered useful, and any other external data irrelevant (Dutta et al., 2014; Wang et al., 2007; Zhang and Deng, 2010). Gunduz-Oguducu (2010) adds that the presence of local noise in web pages makes it difficult to extract useful information, which also decreases the quality of information available on the web. Therefore, prior to finding the interestingness of web data given a user interest, pre-processing becomes one of the critical stages in the web usage mining process. Understanding the source and type of noise in web data simplifies the process of exploring its impact on the process of web usage mining (Mobasher and Nasraoui, 2011)

2.2.4. The Impact of Noise Web Data in the Web Usage Mining Process

Noise in web data is an unavoidable problem that can affect the process of extracting useful information from the web (Xiong et al., 2006; Wu and Zhu, 2008; Yu et al., 2016). The performance of machine learning tools applied in the web data mining process can also be affected by the presence of noise data (Shanab et al., 2012; Sáez et al., 2015-16; García-Gil et al., 2017). This

includes pre-processing of data to remove irrelevant data, extraction of useful information, and classification of similar data. The research proposed in this thesis identifies and discusses the impact of noise in web data in relation to the end users, as well as website developers/owners. The rationale for this consideration is based on the fact that information that is noisy to end users could actually be of commercial benefit to website owners, hence their determination to push this type of content on users.

Web Users: Web users get frustrated when a website promotes content that isn't tailored to their interests. According to a study by Janrain (2013), 74% of web users get frustrated with websites that suggest content that has nothing to do with their interests; 57% say they will leave the site if they are married and shown ads for a dating site. Various machine learning algorithms/tools have been proposed to address such issues experienced by end users (Adeniyi et al., 2016; Santra and Jayasudha, 2012; Yi et al., 2003). However, since web log data are being continuously generated, in some cases amounting to a dynamic change of user interests, existing tools have not fully addressed such changes while eliminating noise data.

Website developers (owners): Website owners attempt to suggest more content to end users, especially promotional information for marketing purposes. Their approach is based on the fact that if some content is popular with the majority of users, or within a certain geo-location, then it should be widely disseminated. One of the problems with this approach is that users have no choice other than to view or click on the suggested web pages. The feedback received can be misleading where user interest learning is not considered. Khasawneh and Chan (2006) state that analysing web data logs can provide useful information that helps web developers suggest information that meets the needs of a user given the time of interest.

In summary, this section examines critical issues in relation to problems with noise in web data. One of the research objectives is to provide a clear understanding of how existing research defines noise and addresses problems with noise in web data. The criterion is to identify and critically analyse current research work in order to position the proposed research. In

the following section, this thesis will explore various machine learning tools applied by existing research in order to address the identified problems. At the end of this chapter, the thesis will be able to position its proposed research and respond to the defined research question.

2.3. Critical Analysis and Evaluation of Relevant Research Works

This section explores how existing research addresses the problem of noise in web data. In order to justify the position taken by the proposed research, this section undertakes the following critical investigations:

- 1) Establish from current and relevant research work the different types of machine learning tools applied to identify and eliminate noise in web data.
- 2) Identify and critically analyse machine learning tools proposed by existing research to address problems with noise in web data; this includes considering the contribution, limitations and how the performance of existing tools are evaluated.
- 3) Ascertain whether the existing tools take into consideration the interests of a user prior to eliminating noise in web data.

2.3.1. Noise Web Data Reduction: Layout and Structure of Web Data

The current tools developed to identify and eliminate noise from web pages are generally based on (1) the underlying structure of the document as appraised using a document object model (DOM) tree (John and Jayasudha, 2016; Yi et al., 2003); and (2) entire dependence on the visual layout of web pages (Akpınar and Yesilada, 2013). The DOM tree is a data structure used to represent the structure of a web page; it is built using a web page's HTML parse, from which a web content structure is created to distinguish areas of a website based on relevance and noise data (Dutta et al., 2014; Garg and Kaur, 2014).

Existing research works have proposed a number of tools using the DOM approach that eliminate noise in web data. For example, the Site Style Tree (SST) proposed by Yi et al. (2003) detects and eliminates noise data from web pages based on the observation that the main web page contents usually

share the same presentation style and that any other pages with different presentation styles may be considered noise. To eliminate noise from web pages, SST simply maps the page against the main web page to determine whether the page is useful or noise based on its presentation style. With regard to removing noise from web data, SST only considers the structure and layout of a web page, thus neglecting the needs of web users when it comes to identifying information that fits their interest or not (Dutta et al., 2014).

A new tree structure was proposed by Yi et al (2003) to capture the general presentation style and usefulness of a page on a specified website. A Site Pattern Tree algorithm (SPT) is used as a measure that determines the parts of the web representing noise and those that represent the core information of the web page. By mapping any web page to the SPT, noise data is identified and eliminated. Narwal (2013) also proposed an algorithm to eliminate noise in web pages based on the DOM approach. The Pattern Tree algorithm proposed by Narwal captures the layout pattern and actual content of the web page. The objective is to improve the web usage mining process through classification of web pages prior to removing noise data. The algorithm analyses different web pages to formulate two key measures based on the style and similarity of web pages. Using a defined threshold value, the two measures are then engaged to identify noise data from the main content of a web page. The general observation made by Narwal is that web pages in a given website often follow a similar layout. Therefore, any content with a dissimilar pattern will be considered as noise. Based on such view point, it is clear that the importance of web data is subjected to the layout of the website and not interests and needs of a user.

Swe Swe Nyein (2011) proposed the Content Structure Tree algorithm (CST), which also uses the DOM tree to identify and extract irrelevant data from the main web page content. The proposed tool has the ability to rank the content using similarity value and it subsequently extracts relevant data based on the given search criteria. CST uses the cosine similarity measure to evaluate which parts of the web page contain relevant and irrelevant data. The cosine similarity measure is widely used in web data mining to determine how different data values are likely to be identified based on their interestingness

(Bhattacharjee et al., 2015). However, classification of similar web pages does not necessarily mean absence of noise in web data due to the fact that content such as navigations panels, copyright and privacy notices, as well as advertisements, can always have a serious impact on the quality of information available on the web when user interest is taken into account.

Dutta et al. (2014) proposed a machine learning tool to remove noise from web pages based on structural analysis and the regular expression of web pages. The two main steps applied in their proposed work are tag-based filtering, which means that information with positive tags form the useful part of the web page, while negative tags contain information that is noise. The assumption made in their proposed work is that noise present on every page of a website has the same presentation style. Therefore, the process of identifying and eliminating noise in web data is based on the principle that the consistency of web pages on a given website separates noise from useful web data. Similarly, Sivakumar (2015) proposed a tool to remove a large amount of information that is not part of the main web page content. The proposed tool aims at identifying and removing banner advertisements, navigation bars, copyright notices, etc. This type of web data is considered noise and thus likely to not fit with the user interest. The author uses keyword redundancy, link-word percentage and title-word relevancy to identify noise in web data. These parameters are used to compute the importance of each web page based on a defined threshold value so as to determine whether the page is relevant or noise prior to elimination. Even though the author acknowledges that it is important to eliminate noise in web data that may affect user interest, there is no evidence in the work to suggest how the interests of a user influence the elimination of noise.

Jiang and Yang (2015) proposed an algorithm that uses the DOM tree to identify and eliminate noise from web pages. The objective is to preserve the original structure and layout of the web page so as to ensure an efficient web usage mining process. The observation made by Jiang and Yang indicates that extracting the main content from web pages has recently become more difficult due to the fact that all web pages contain information that are irrelevant to the main content. Jiang and Yang argue that there is no algorithm that can

completely solve problems with noise web data independently. Based on their literature findings, measures such as style and layout are considered independently in noise reduction, rather than a combination of both. However, this thesis proposes acknowledging that problems with noise in web data cannot be solved independently, suggesting that the analysis of both web data and user interest is key to an efficient web usage mining process. The rationale for this viewpoint is that a web page contains information that is aimed at addressing the needs of the 'target audience', which is the user. Therefore, regardless of the structure and layout of web data, the interests of a user should be considered critical and therefore overrule how data is structured on the web.

Htwe and Kham (2011) developed a tool using case-based reasoning (CBR) and neural networks to eliminate noise data from web pages. CBR is a machine learning approach that makes use of past experience to solve future problems; in this case it detects noise from web pages using existing stored noise data for reference. The elimination of noise from web pages not only depends on the DOM tree, but also the classification of results of neural networks. Artificial Neural Network is used to match noise patterns with those stored in the case base. This approach is based on the idea of using case-based reasoning to identify noise data by matching existing noise patterns stored in the case base. However, it is difficult to determine if such content is relevant or noise to a particular user interest because: (1) it matches existing patterns of noise data and the output can be misleading because web data is dynamic, as is the user's interest; and (2) Htwe and Kham (2011) argue that relevant information mainly exists in the middle of a web page, while the rest of the page contains noise data, but again information in the middle of the web page can be noise if user interest is considered.

To address these challenges, Pappas et al. (2012) proposed an algorithm that takes into account the non-visual characteristic of a web page to identify and eliminate noise data. The Least Recently Used (LRU) paging algorithm is used to detect and remove noise from web pages. LRU considers both the visual and non-visual characteristics of a web page and is able to remove web data noise, such as news, blogs and discussions. The LRU algorithm determines

frequently visited pages and those that have not been visited over a certain period. The algorithm then classes least recently used pages as noisy or irrelevant content that needs to be eliminated. However, various aspects can contribute to infrequent use of a web page. For example, in the case of seasonal data, which a user will only be interested in for a specific time or occasion. LRU was also applied by Garg and Kaur (2014), but there is no clear discussion of whether least visited web pages might be considered as useful in the future.

The proposed research outlines some of the major contributions made by the existing research. For example:

- i. Protecting useful data regions by identifying boundaries between noise and useful data (Hu et al., 2013; Velloso and Dorneles, 2013; Wang et al., 2014, 2011). This is tied to the assumption that only the main web data contains useful information.
- ii. Automatically detecting and removing noise data by matching noise data in extracted web log data with previously stored web data noise patterns. The contribution of the proposed research is not only to remove multiple noise patterns from logs relating to a web page, but also to enable classification of the noise encountered based on defined patterns.
- iii. The existing tools applied in the noise web data reduction process mainly focus on improving the quality of web data by removing noise at the pre-processing stage. Borzemski (2007), Narwal (2013) and Azad et al. (2014) add that the discovery of useful information that characterises the interests of the user is dependent on the performance of the machine learning tool applied.

From the above literature, current research acknowledges that the noise web data reduction process is aimed at improving the process of mining useful information from the web (Jiang and Yang, 2015; Xiong et al., 2006; Yang and Fong, 2011). They also acknowledge that there is a need to extract relevant information from the noisy web to ensure that the main content of the web is protected from noisy data. However, relevant data and the main content of the

web are defined and interpreted in the current literature from the following viewpoint: (1) relevant data is the main content of a web page that a user needs to view (Lingwal, 2013); and (2) the content pages are web pages where a user can find useful information, while anything that does not form part of the main web page is noise (Kapusta et al., 2012).

The proposed research's viewpoint is that the usefulness of web data is determined by user interests, which also change over time. In order to justify this perspective, **Figure 2.3** presents the relationship between web users and web data. The interdependence between the two actors – user and website – demonstrates the need to learn noise in web data, taking into account the interests of a user prior to elimination. It is widely recognised that web servers are the richest source of web data (Mehak et al., 2013; Ramya et al., 2011). Data stored in web log files defines the relationship between a user and the web. The interestingness of information provided to users by web servers can therefore be influenced by what an individual user is interested in. The following section explores how current research addresses noise in web log data. The findings will aid in understanding whether noise in web data is influenced by user interests discovered from web log files.

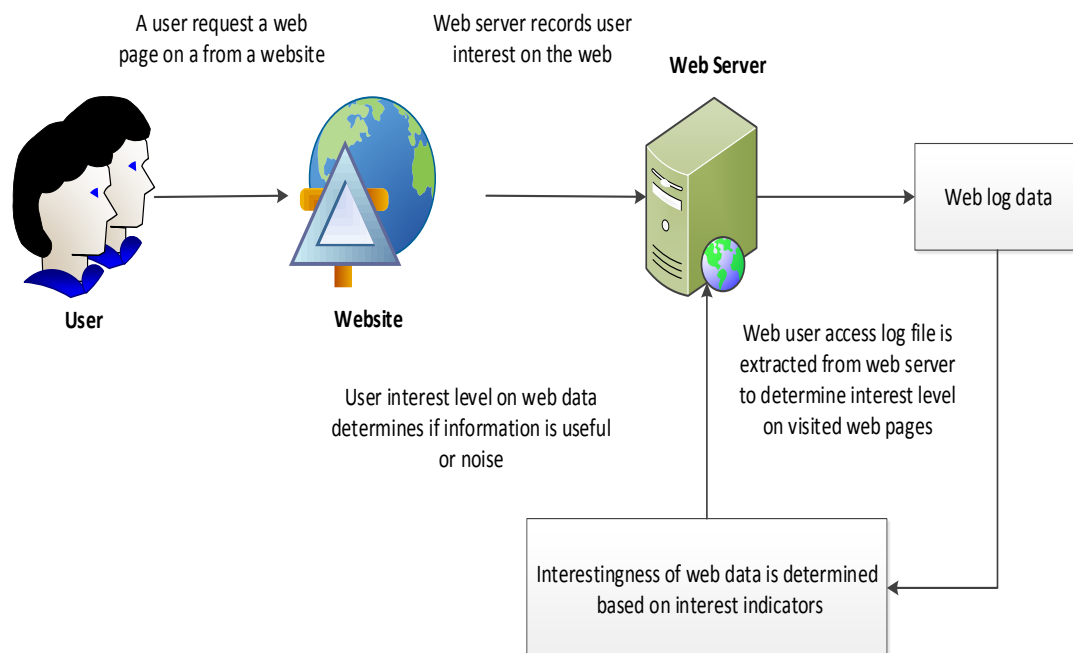


Figure 2.4: Relationship between a user and web data

2.3.2. Noise Web Data Reduction: Extracted Web Data Logs

The proposed research does not simply rely on the structure and layout of web data to identify and eliminate noise; instead, it focuses on the interests of a user in relation to the available web data. However, the existing research does play a significant role in defining user interest level based on web log data. The reduction of noise from web pages based on the structure of data on a given website will thus affect the quality of the extracted web log data. The proposed research work argues that noise in web data should be determined based on the web user's level of interest in the available data. The rationale for this argument is that while the web provides valuable information to its users, the interestingness of data can only be determined if there is any evidence that a user is interested; otherwise such information is noise regardless of its structure on a website.

Valuable information about user interests on the web is hidden in user logs extracted from a web server. In order to ensure useful information is extracted from raw web log files, understanding user needs and how they vary is critical to this process. The outcome will provide a more dynamic approach to finding useful information that reflects changes in user interests. In order to understand user interest from the extracted web log data, pre-processing is essential for the purpose of improving the quality of the output from the web usage mining processes (Nithya and Sumathi, 2012). Pre-processing eliminates noise data from web user logs (Vidyavathi and Begum, 2016). The process involves building a web user profile through user, session and page view identification in order to determine user interest level in relation to visited web pages. Han and Xia (2014) acknowledge that the web usage mining process can identify useful information from noisy data if the pre-processing of the web log takes into account the level of user interest. A more detailed discussion and critical evaluation of pre-processing web log files is presented in the next chapter.

A number of machine learning tools have been proposed for extracting useful information from the noisy web. For example, Santra and Jayasudha (2012) applied a Naïve Bayesian Classification (NB) algorithm to identify the interests of users based on web log data extracted from a website. Their main objective

was to classify extracted web data logs and study how useful the extracted information was based on a user's interest. Their initial processing phase involved removing noise data, such as advertisement banners, images and screensavers, from the extracted web data logs. They used a Naïve Bayesian Classification model to classify useful, as opposed to noise data, based on the number of pages viewed and the time spent on a specific page.

Sripriya and Samundeeswari (2012) proposed an algorithm based on Neural Networks to determine the frequency of a web page in extracted web log data. The frequency of occurrence of a web page in a user log file shows the level of user interest. The authors define weight as a statistical measure used to evaluate interestingness of a web page to a given user. Azad et al. (2014) applied kNN to web data logs to find useful information from noisy web log files; their main focus was on local noise, for example, advertisements, banners, navigational links etc. Web log data was extracted and surveyed with regard to which web server they belonged. If the address belonged to a list of already defined advertisement servers, the link was removed. Similarly, Malarvizhi and Sathiyabhama (2014) proposed a Weighted Association Rule Mining method for extracting useful information from web log data. Their objective was to find web pages visited by a user and assign weight based on interest level. The user interest-based page weight is used to eliminate noise web pages from useful information. In their research work, the weight of a web page in relation to a user interest is estimated from the frequency of page visits and the number of different pages visit. Where pages are visited only once, for instance, they will be assigned low weight and subsequently considered irrelevant (Gupta et al., 2016; Tyagi and Sharma, 2012).

2.4. The Influence of User Interest on the Noise Web Data Reduction Process

Finding useful information based on user interest is challenging due to the increasing amount of data available on the web (Nanda et al., 2014). The interestingness of web data is dependent on a user and the interests of a user may change over time. Wu and Liu (2014) acknowledge that user interest relies on the principle that the visiting time of a page is an indicator of the level of user interest. The amount of time spent on a set of pages requested by the

user within a single session forms the aggregate interest of that user in that session.

User interest can be determined in two ways: explicitly or implicitly (Nanda et al., 2014; Rao et al., 2017; Wei et al., 2015). Explicit interest is where users provide feedback concerning what they think about the information that they have received. However, many users are unwilling to state what their true intentions are in terms of the visited websites. For this reason, implicit learning of user interest is also considered by this proposed research. The implicit method uses logs created by a user's visit to a website to learn their interests, instead of requesting user feedback. There are two ways of capturing implicit user interest: from browsing behaviour and from browser history (Grčar et al., 2005; Kim and Chan, 2005a). Browsing histories capture the relationship between a user's interests and their click history; this is necessary for the identification of useful information from the extracted web log data. Learning user interests plays a fundamental role in understanding how useful web data is to a user at a given time, thereby improving the process of noise web data reduction. It is therefore important to understand how the dynamic nature of the web and varying user interests influence the identification of noise in web data. In the proposed research work, the focus is on learning the interests of a user in relation to available web data with the aim of reducing the amount of useful information eliminated as noise.

2.5. Discussion of Critical Aspects

This thesis highlights a number of critical issues that are widely discussed in recent research, but with a limited or different approach to addressing problems with noise in web data. For example, in terms of the recent works' emphasis on the need to identify and eliminate data that does not form part of the main web page content. This is done to ensure that the core information that forms the main content of the web is isolated from noise data (Lingwal, 2013). The machine learning algorithms that have been proposed by the existing research are based on the observation that web pages usually share common layouts and presentation styles (Das et al., 2012; Jiang and Yang, 2015; Nithya and Sumathi, 2012). However, the dynamic nature of the web makes it difficult to rely on presentation of web pages for the identification and

elimination of noise data. Secondly, the common objective identified from existing research on noise web data elimination is to improve the performance of the web usage mining process (Ramya et al., 2011). Performance in this respect focuses on easy access to information from the web and discovery of useful information. However, there are no clear discussions of how the performance of existing tools is evaluated and whether user interests are considered. Azad et al. (2014), Narwal (2013) and Ting and Wu (2009) argue that the performance of a web usage mining process is evaluated based on the discovery of useful information that characterises the interests of end users. Therefore, eliminating noise in web data should consider the interests of the web user in order to determine the interestingness of data on the web. Finally, there is a great deal of focus on identifying and eliminating noise, such as advertisement banners, failed https links, mirror sites, duplicated web pages, copyright, external links etc. Since the process of identifying and eliminating this type of noise is mainly based on its relationship with the main web page content, the process is not user-driven, hence the outcome will not reflect user interests.

Appendix II is a summary of some of the recent research works that have applied data mining techniques to extract useful information from web pages, analyse web log data by removing any irrelevant data and subsequently identifying useful information based on a specific user interest. Data attributes considered as input mainly include user IP address, page URL, time of access etc. This indicates that creating a data class, cluster or associating a user to data available on the web takes into account these attributes. On the other hand, IP address used by the user to access a web page can be used to determine interest web pages but user's interest cannot only be directed towards the IP address. Sometimes it may require a combination of other data attributes such as source of web page visited, type of request, time of request, frequency of visits etc. to determine user level of interest.

In summary, this work's main focus is to learn to recognise noise in web data in order to reduce the loss of useful information otherwise considered as noise, as well as to decrease noise levels. Rather than isolating the main web page content and relying on its layout and content, the proposed research

aims to focus on how user interest can influence the type of noise present in web logs.

2.6. Chapter Summary

This chapter undertook a critical review and analysis of the existing research work addressing problems with noise in web data. The aim was to find out how current research defines and addresses noise in web data; the findings respond to the first research question identified in the first chapter of this thesis. At the onset, the criteria was to review and evaluate existing literatures covering the following aspects: definition of noise in web data, tools and techniques proposed to identify and eliminate noise data, measures employed by existing research to evaluate the performance of existing tools, contribution and limitations taking into account the defined problems, and the current situation.

The critical review and evaluation of the existing research conducted in this chapter acknowledges the contribution made by using existing machine learning tools to address problems with noise in web data. This thesis found that although there are a number of tools and techniques that identify and eliminate noise in web data, there are still critical issues that have not been fully addressed in relation to noise in web data. For example, there are no tools currently applied to learn noise web data prior to elimination. There are no discussions on how the existing tools used to eliminate noise in web data take into account evolving user interests. The existing research work has therefore not explicitly defined measures that will aid in understanding the influence of user interests and how change in user interests is modelled to minimise loss of useful information. Therefore, the discussions presented in this chapter begin to respond to research question 1 below:

Question 1: In what ways do current research works define and address noise in web data?

Noise in web data is defined by the existing research mainly based on the structure and layout of web pages. It is also widely acknowledged that noise in web pages often follow a similar layout pattern, which is used to distinguish between useful and noise web data.

The rationale to propose a new approach is based on the fact that if noise in web data is not defined in relation to web users and their evolving interests, the interestingness of web data will be misinterpreted during the web usage mining process. In the following chapter, a research methodology framework is presented. The methodological approach considered in this thesis defines the process of collecting user interest information, as well as different phases of pre-processing data. A user profile, which is defined in the next chapter, plays a key role in learning the interestingness of web data and subsequent elimination of noise based on user change of interests.

Chapter 3: Web User Profiling Based on Web Data

In chapter 2, a critical review and evaluation of the existing research that addresses problems with noise web data was presented. A number of critical issues were examined, for example, ways in which the existing research address problems with noise web data, their contribution and limitations. The objective was to understand to what extent recent research addresses defined problems vis-à-vis the current situation. The position taken by the proposed research was then defined with a justification as to why there is a need to propose a new approach to address the defined problems. This chapter examines the methods recent research considered for learning about web users and their interests.

3.1. Introduction

This chapter makes reference to the proposed research focus defined in chapter 1: learning noise web data taking into account changes in user interests. Learning the interestingness of web data based on user interest involves finding information on the web that defines user interests, and building a user profile based on user interest information. User profiling is defined as the process of identifying data from the web in relation to user interest (Gauch et al., 2012; Kanoje et al., 2014; Dias et al., 2017). The goal of user profiling is to find and extract information from the web on what a user is interested in while on the web. Key aspects of user profiling involve understanding varying user interests and learning how such changes influence the interestingness of web data. Generally, a user profile evolves over time, which means that information that defines user interest is time variant. Therefore, there is a need to examine how changes in user interest impact the identification and subsequent elimination of noise in web data.

Building a user profile can be considered a process in which machine learning tools are applied to understand user interests on the web (Kanoje et al., 2015). The process mainly relies on a user's browsing behaviour while on a given web page. Machine learning algorithms are then applied to analyse user visits in order to discover user interest level, taking into account a number of measures. For example, the time spent on a given web page and the number

of visits made within a given period of time can signify interestingness of a web page.

This chapter introduces the methodological approach the proposed research uses in an attempt to address the defined objective: finding how changes in user interests impact identification and subsequent elimination of noise in web data. The framework of the proposed research methodology is outlined in **Figure 3.1**. A typical user profiling process consists of a number of phases (Gauch et al., 2012; Kanoje et al., 2014; Rathipriya and Thangavel, 2014). The initial phase is collecting raw information about user interests, involving, for example, extraction of web user access logs from the web server. Access logs contain user records, such as IP address, request time, requested URL and agent. This type of information is used by search engines to better understand user interests. It aids in discovering trends, patterns of user interests and data that does not fit the interests of a user, such as noise data. The second phase focuses on pre-processing raw information; the pre-processing phase includes the user, session and page view identification process. In the final phase, a user profile is constructed.

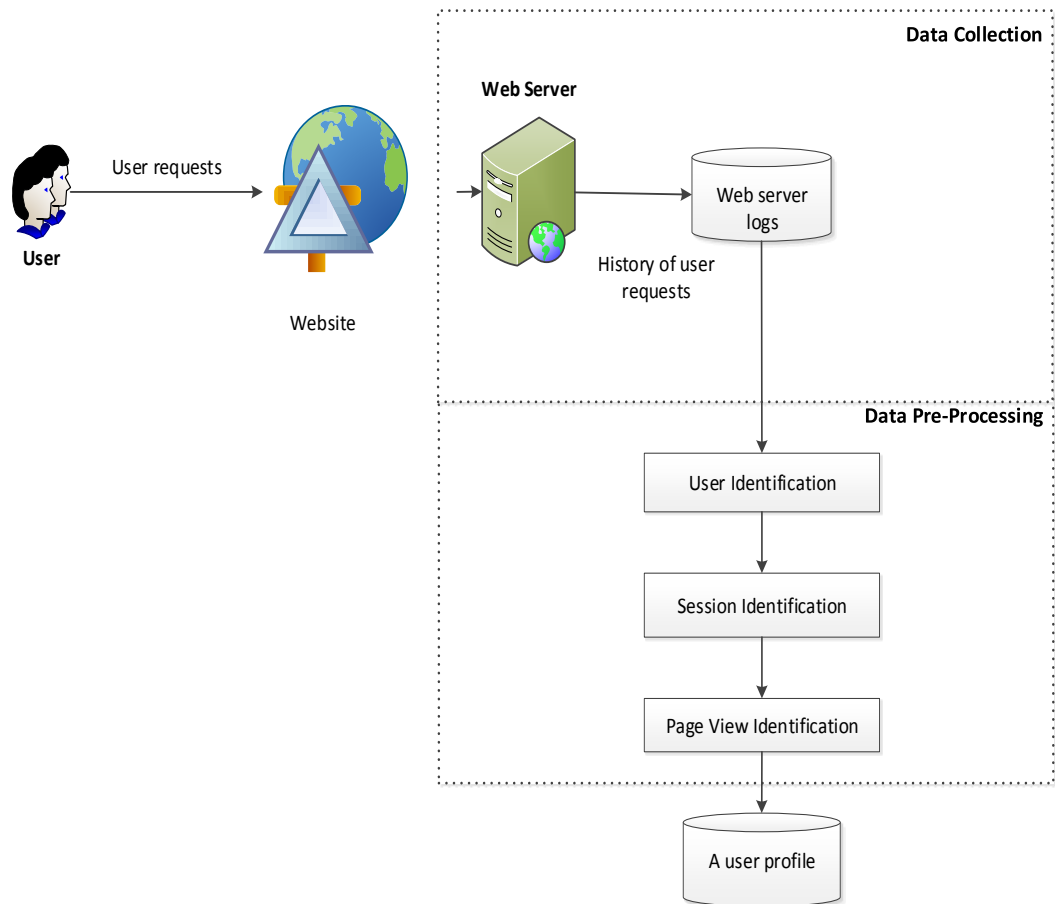


Figure 3.1: Extraction and pre-processing of web log files

This chapter explores in detail the key stages of processing raw web log files, as outlined in **Figure 3.1**, and subsequent creation of a user profile prior to learning user interests.

3.2. Collection of User Interest Information

The initial phase of user profiling is to collect information about a user that defines his/her interests (Isinkaye et al., 2015). In order to identify and define user interests on the web and manage changes in this over time, the proposed approach should be able to tell a story about the user. Therefore, collecting information about user interests is critical to learning the interestingness of web data prior to noise elimination. User interests are determined from the user’s journey on the web, which is defined by information about the user collected either explicitly or implicitly. The explicit method of collecting user interest information involves asking web users directly about their interests, usually in the form of rating web pages they have visited. The implicit method

is based on analysing user visits to a given website and this is done without the user's knowledge.

3.2.1. Explicit User Interest Information

The explicit method of collecting user interest information, often referred to as explicit user feedback, mainly relies on web users providing direct feedback on information of interest (Reusens et al., 2017). Explicit user feedback is usually collected in the form of user ratings. For example, in **Figure 3.2**, users are required to rate information as either relevant or not relevant. Website owners will thereafter analyse user feedback in order to improve the web usage mining process, thus minimising the noise data suggested to a user.

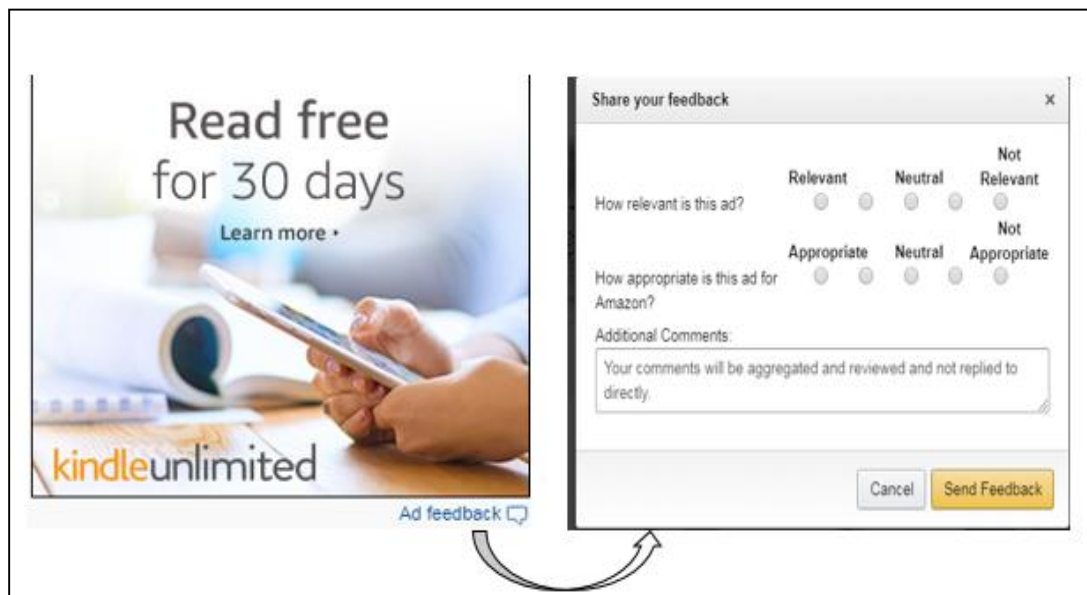


Figure 3.2: Explicit user feedback.

Accessed via <https://www.amazon.co.uk/>

Even though explicitly collecting user interest information is easy to implement, the method has some limitations. For example, it requires users to take time and explicitly rate the web page before proceeding to another page. Furthermore, it is considered difficult to motivate users to continuously provide an explicit rating (Kim et al., 2002). The user's unwillingness to provide accurate information about their interests on the web is another challenge; some may consider it to be time consuming and hence they may opt not to participate (Ajabshir, 2014).

This method therefore falls short of adaptability in line with a change in user interest over time. It is difficult for this to reflect the user's change of interest due to its static nature (Pasi, 2014). As a result, a user profile built using the explicit method does not usually reflect changing user interests because they degrade over time. Gauch et al. (2012) and Jawaheer et al. (2014) add that web users are explicitly requested to rate a web page based on a given level of interest, i.e. 'not relevant...to relevant'. However, whenever the interest of the user changes, the previous rating of the web page will not change unless the user updates the profile to reflect their current interest. Therefore, using the explicit method requires a lot of effort to update a user profile in order to ensure the noise data eliminated reflects the user interests at any given time.

3.2.2. Implicit User Interest Information

Defining a user profile based on information collected explicitly is only relevant for a given period of time (Akuma et al., 2016). However, as user interests change, the current profile information may become noisy unless the user advises the website owner about their change of interests. Instead of relying on a user to advise on changing interest, an implicit method of collecting user interest information is considered by this research. The implicit method is the process of extracting and analysing user visits in order to determine their interest level with regard to web pages visited (Kim and Chan, 2005; Nanda et al., 2014). There are a number of ways to implicitly collect user information; these include: web user logs, clickstream, browsing histories and content information from visited web pages (Fan et al., 2014). Existing research studies argue that web log data are a rich and common source of implicit user information (Gu et al., 2016; Kanoje et al., 2015; Gauch et al., 2012). Web log data contains links to visited web pages, the date and time of every user visit. These records are captured by the web server, thereby allowing the website to define interestingness of a web page in relation to user interest levels. Kim and Chan (2005) support the idea that interestingness of a web page should be captured from user's visits to a given website in order to assess their interest level. The main advantage of implicitly collecting user interest information is that it does not directly require user effort when constructing a

user profile (Zahoor et al., 2014). It also allows for easy and continuous access to data, hence the ability to learn user interests as they change over time.

Based on the discussion of the recent research, this work suggests that the implicit method is easily updatable to a dynamic change of user interest. Therefore, the proposed research's choice to use web log data to learn noise web data prior to noise elimination is based on the following key aspects.

1. *Users rarely give explicit feedback regarding their interests:* It is widely acknowledged that users are reluctant to perform actions such as rating a web page they have visited. Asking a user to rate web pages is not only time consuming, but can also rely on a user's willingness to disclose their interests.
2. *User interests change over time:* User interests are bound to change over time (Kellar et al., 2004; Pasi, 2014), which means that using explicitly collected user interest information from the past is less reliable. This is because a user will be required to manually advise whenever their interests change by rating or updating preference forms. Consequently, if the past interests of a user are used to determine interestingness of a web page, the information available to a user is likely to be noise.

3.3. Extraction and Pre-processing of Web Log Files

The web server logs used by the proposed research are extracted from a web server of an ecommerce website for a period of 90 days. The objective of using this range is to collect a wide range of data capable of understanding user interests, as well as with the ability to identify any changes within the specified time period. Each record is associated with a unique IP address that has been anonymised with a User_ID. The assumption taken is that the IP address is unique to the user where operating system and browser type/version is considered. 50 users were randomly selected with the criteria that they had to have an average of 10 unique clicks a day for a period of 90 days.

The objective of data pre-processing is to transform log files extracted from a web server into a user profile (Ansari et al., 2015). Every visit a user makes to

a website is recorded in a web server and stored as a web log file. As discussed in chapter 1, web log files store useful data patterns that define the interests of a user on a web page, but it is difficult to extract such data without the pre-processing phase. Pre-processing involves cleaning raw data with the aim of ensuring the extracted web log file provides a clear picture of the type of user interest data, the level of interestingness and whether the interests of a user change over time (Aye, 2011; Dhandi and Chakrawarti, 2016; Hussain et al., 2010). Since it is difficult to identify the interestingness of a raw web log file, pre-processing of web log files is considered a critical phase in the web usage mining process (Lokeshkumar and Sengottuvelan, 2015). This is to ensure useful information is identified from the web log file, which is believed to contain noise data. A log file is a plain text file that records information about each user visiting a website (Nithya and Sumathi, 2012); for example, the IP address that identifies a user visiting a specific website, the timestamp that reports the time of the visit, the web page requested, browser and operating system used, etc. A web server writes information into a log file each time a user requests a web page from a specific site and every request is recorded in a web log file. A record of a user visit to a web page comprises:

- IP address – in this research, IP address has been anonymised with the User_ID
- A link to the page visited, i.e. Uniform Resource Locator (URL)
- Time of visit, which is presented by the Time_Stamp
- Agent that stores the browser used; operating system, etc.

Table 1 shows a sample log file containing user records extracted from a web server

Table 1: Raw web log file

User_ID	URL_ID	Time_Stamp	Agent
150	58	22/01/2016 17:40:02	Mozilla/5.0 (X11; Linux i686; rv:17.0) Gecko/20100101 Firefox/17.0
150	56	22/01/2016 17:29:36	Mozilla/5.0 (X11; Linux i686; rv:17.0) Gecko/20100101 Firefox/17.0
150	53	22/01/2016 17:21:35	Mozilla/5.0 (X11; Linux i686; rv:17.0) Gecko/20100101 Firefox/17.0
150	1	22/01/2016 17:21:08	Mozilla/5.0 (X11; Linux i686; rv:17.0) Gecko/20100101 Firefox/17.0
126	55	21/01/2016 10:37:10	Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75 Safari/601.1
126	56	21/01/2016 10:24:10	Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75 Safari/601.1
126	53	21/01/2016 10:13:16	Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75 Safari/601.1
126	1	21/01/2016 10:13:10	Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75 Safari/601.1
173	56	19/01/2016 06:19:30	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	53	19/01/2016 06:13:29	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	55	19/01/2016 06:02:59	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	1	19/01/2016 06:02:52	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	53	15/01/2016 11:33:30	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	56	15/01/2016 11:25:29	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	54	15/01/2016 11:20:12	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30
173	1	15/01/2016 11:19:52	Mozilla/5.0 (Linux; U; Android 4.1.2; de-at; GT-I8190 Build/JZO54K) AppleWebKit/534.30 (KHTML, like Gecko) Version/4.0 Mobile Safari/534.30

Munk et al. (2015; Zhang and Chen, 2012) argue that the process of finding useful information from a noisy web log file is dependent on the pre-processing stage, which involves a number of phases. Firstly, users are identified based on information requests logged on a web server. Secondly, sessions from different users are identified. A session is a sequence of

records accessed by a user visit to a website within a defined time duration. Following user session identification, the sequence of user records accessed by the user is generated. A sequence in this case is a set of items with a specific relationship between them. Finally, a user profile is created; the profile aids in learning user interest levels with regard to the visited web pages.

3.3.1. User Identification

A user is an individual accessing a web page through a web browser (Grace et al., 2011). User activities are recorded as web logs on the server based on the time-stamp that notes when they occurred. The relationship between a user and web log record is considered to be one-to-many, i.e. each user is identified by one or more records. User identification is the process of identifying each user who has visited a given web site (Grace et al., 2011; Neelima and Rodda, 2016). The user identification process is based on the following rules: different IP address reflects different users, the same IP with a different operating system or different browser should also be considered as a different user (Patel and Parmar, 2014). Prior to learning the interests of a user from visited web pages, it is important to identify and study every user's visit characteristics, such as, IP address, location, new or returning user, operating system, browser used to visit the website, etc. These characteristics aid in creating a user profile that is adaptable enough to reflect changes in user interest.

The following is an extracted web log file presented as a set of records $N = (rec_1, \dots, rec_n, \dots, rec_N)$, where N is the total number of records in a web log file, as shown in **Table 2**. The n^{th} record of the j^{th} user is defined by the following attributes: $rec_n^j = (user_id_1^j, url_id_n^j, time_stamp_n^j, \dots, agent_n^j)$, where

$user_id_1^j$ = the IP address that identifies a user visiting a specific website

$url_id_n^j$ = a link to the web page requested by the j^{th} user

$time_stamp_n^j$ = date and time the j^{th} user visited a web page

$agent_n^j$ = the browser version and operating system used by the j^{th} user

From the above representation of records in web log file N , **Algorithm 1** is used to identify the n^{th} record of the j^{th} user, as shown in **Table 2**.

Algorithm 1: User Identification

Input: N // web log file

Output: rec_n^j //A set of records for the j^{th}

Begin

 Read the logs in N

 For every entry in N

 If (IP address of first log entry = IP address of second log entry)

 Compare the user browser and operating system of both entries

 Else

 If both user browser and operating system are the same,

 Assign both entries to the same user_id.

 Else

 a different user_id

 End if

 End for

Table 2: A set of records in the j^{th} user profile

$user_id^j$	url_id^j	$time_stamp^j$
173	1	02/01/2016 08:02:08
173	54	02/01/2016 08:07:52
173	53	02/01/2016 08:29:29
173	56	02/01/2016 08:33:30
173	55	02/01/2016 08:39:54
173	1	06/01/2016 19:32:27
173	53	06/01/2016 19:33:10
173	56	06/01/2016 19:34:19
173	20	06/01/2016 19:40:10
173	54	06/01/2016 19:51:02
173	1	09/01/2016 08:02:52
173	54	09/01/2016 08:02:56
173	55	09/01/2016 08:17:29
173	56	09/01/2016 08:24:30
173	1	15/01/2016 11:19:52
173	54	15/01/2016 11:20:12
173	56	15/01/2016 11:25:29
173	53	15/01/2016 11:33:30

Table 2 shows access records for the j^{th} user. Each record includes the User_ID, the timestamp of each user access record and the web page visited,

which is determined by the URL_ID. The timestamp of each page visit plays a critical role in identifying user sessions because the difference between two timestamps determines whether access records are within a session given a specified threshold value.

3.3.2. User Session Identification

A session is defined as a sequence of records accessed by the same user within a single visit to a website (Dwivedi and Rawat, 2015; Kapusta et al., 2012). The user session identification process is considered one of the key phases of pre-processing web log data; it segments records of each user visit to a website into sessions (Patel and Pamar, 2014). The process of identifying user sessions from a web log file aids in finding a sequence of user records on a website from the time of entry until he/she exits the website. The session identification process can either be navigation-oriented or time-oriented (Castellano et al., 2013; Srivastava et al., 2000).

Navigation-oriented user session identification: The time a user spends on a web page can illustrate the interestingness of the web page. The navigation-oriented approach takes into account the sequence of web page access by a user based on the structure and layout of a web page (Kapusta et al., 2012). Each page visit can be categorised as navigational/auxiliary and content page (Varnagar et al., 2013; Kapusta et al., 2012). Content pages are perceived as the ultimate destination of a web user and a user is likely to spend more time on such pages, whereas navigational features are hyperlinks that simply connect to content pages (Mayil, 2012; Aldekhail, 2016). However, in the proposed research, the usefulness of web data is determined by the interests of a user rather than the structure or layout of a website. Importantly, the navigation-oriented approach to session identification overlooks the interests of a user with regard to available web data. To address the limitations of the navigation-oriented approach, Yuankang and Huang (2010) and Kapusta et al. (2014) argue that time-oriented session identification approach is better at identifying sessions that correspond to user interests on a web page.

Time-oriented user session identification: The time-oriented session identification process is considered the most common technique used to identify user sessions from raw web access log files (Srivastava et al., 2000; Jafari et al., 2013; Kapusta et al., 2014; Rao et al., 2017). A number of algorithms based on fixed time-out values have been proposed by existing research. For example, Guerbas et al. (2013), Halfaker et al. (2014) and Verma and Kesswani (2014) propose a time-oriented algorithm based on a fixed time threshold value for user session identification. Time-oriented session identification defines a user session as a sequence of requests made to a web server by the same user within a specified time. The assumption applied in the time-oriented approach is that if there is a break between user requests that is reasonably long, it is likely that the user is no longer active, and therefore the session is considered to have ended. A new session will then start when the next user request is reported on the server. For example, consider ts_k as the initial timestamp of the first page request by the j^{th} user in i^{th} session, and ts_{k+1} is the timestamp for the page put into the current session. If $ts_{k+1} - ts_k > time_{threshold}$, then a new session is created. **Figure 3.3** shows that the first session was created when the j^{th} user visited the home page and shoes categories, 30 minutes later he/she visited the computers and phones categories.

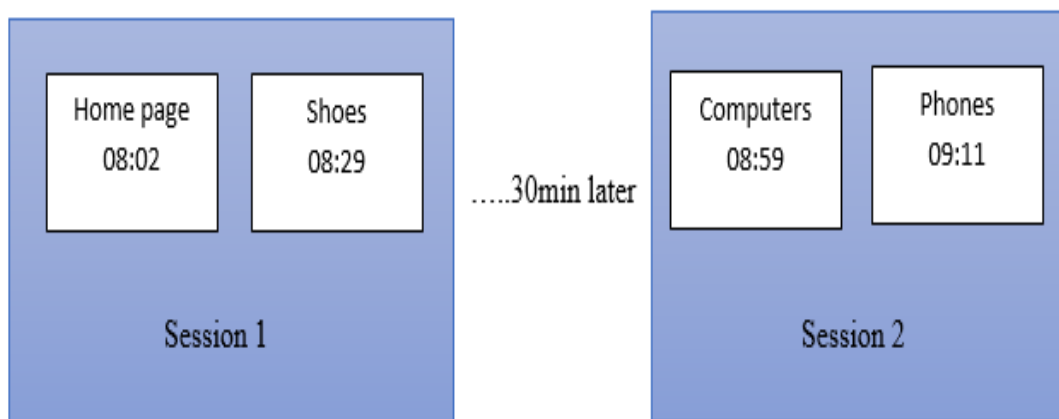


Figure 3.3: Session identification based on 30 min threshold value

It is thus observed that a session is created if the timestamp between two pages requested exceeds a given threshold value, such as 30 minutes, as

shown in **Figure 3.3**. Where the proceeding timestamp is not recorded, the session will end 30 minutes after the current timestamp. Zhang and Ghorbani (2004) argue that each user request and time spent on a page can be affected by website structure and layout, as well as varying user interests. Hence, using a fixed threshold value to determine a session does not reflect the actual time on a page. It is therefore important to note that determining the interestingness of a web page based on fixed time-out threshold value can be misleading. For example, the time spent on the current page may be less than 30 minutes. If the interestingness of a web page is measured based on fixed time-out value, it can lead to misclassification of the web page as either noise or useful.

Illustrative example: A session on a web server is defined by the n^{th} record requested by the j^{th} user within a specified time. Every j^{th} user has i^{th} session, each of which has a sequence of records, such that $S_i^j = (rec_1^j, \dots, rec_n^j, \dots, rec_N^j)$. A record of a user visit to a web page comprises a user-id, time of visit, web page visited, etc., where $rec_n^j = (user_id_{in}^j, url_id_{in}^j, time_stamp_{in}^j)$.

The concept of user session is important because it corresponds to what is often considered to be a visit to a website (Yuankang and Huang, 2010). The following stages are used to create a user session based on a fixed time-out threshold value, which is 30 minutes.

1. The time spent on a page must not exceed a specified threshold, for example, 30 minutes in this case.
2. Let ts_k be the timestamp of the initial k^{th} web page request.
3. Let the next ts_{k+1} be the timestamp for the next k^{th} web page request.
4. The next k^{th} web page request belongs to the same session if $ts_{k+1} - ts_k < 30$ minutes, otherwise it becomes the first of the next user session.

Table 3 shows a session created from a web log file with a default time-out threshold value of 30 minutes. However, the proposed research argues that it is difficult to set a fixed time-out threshold value because each user's time intervals between page visits may vary for a number of reasons. For example,

interruptions occur when browsing the internet, when a user is likely to attend to other activities like making a cup of tea or answering a phone call. Based on **Figure 7**, if we were to determine the time taken by the j^{th} user on the shoes category with the timestamp 08:29, it would be 30 minutes, which is reflected in the next timestamp at 08:59. Fatima et al. (2016) argue that session identification based on fixed time-out threshold value fails to consider key aspects in relation to the interests of a web user. For instance, different users have different reading speeds and the content of a web page may vary in structure and layout; thus the time taken will vary. Moreover, where a fixed time threshold value is used to determine session duration, a long session can easily be divided into two sessions. For these reasons, there is a need to determine user session based on a dynamic threshold value rather than using a fixed time-out value.

Table 3: Session identification *for the j^{th} user*

$user_id_i^j$	s_i^j	$url_id_i^j$	$time_stamp_i^j$
173	1	1	02/01/2016:08:02
173	1	54	02/01/2016:08:07
173	1	53	02/01/2016:08:29
173	1	56	02/01/2016:08:33
173	1	55	02/01/2016:08:39
173	2	55	02/01/2016:17:24
173	2	53	02/01/2016:17:36
173	2	54	02/01/2016:17:48
173	3	1	06/01/2016:19:32
173	3	53	06/01/2016:19:33
173	3	56	06/01/2016:19:34
173	3	20	06/01/2016:19:40
173	3	54	06/01/2016:19:51
173	4	1	09/01/2016:08:02
173	4	54	09/01/2016:08:02
173	4	55	09/01/2016:08:17
173	4	56	09/01/2016:08:24

The effectiveness of finding useful information from the noisy web depends on how accurate the process of user session identification is (Patel and Parmar, 2014). The proposed research argues that it is difficult to measure the accuracy of user sessions because any fixed time-out threshold value is subject to incorrect identification of user sessions. It is therefore difficult to determine the interestingness of a web page to a specific user based on a fixed time-out threshold value. This is due to the fact that sessions based on a fixed time-out fail to consider change of user interests. For example, any fixed threshold will be too short for some sessions with relatively long breaks, but too long for other sessions where the access time is too short (Xinhua and Qiong, 2011). This will lead to incorrect classification of web pages, thereby affecting the process of identifying useful information in relation to a user interest.

With regard to problems with the navigation-oriented and fixed time-out session identification approach, the proposed research considers user session identification based on dynamic time-out adjustment values (Sharma and Makhija, 2015). In the dynamic time-out adjustment approach, more emphasis is given to page requests where the time intervals are large or the last timestamp cannot be determined. This is due to the assumption that the large time interval between page requests signifies the end of a session and start of a new one (Zhuang et al., 2004).

User session identification based on dynamic time-out adjustment: User interests on the web vary, and so does the time taken to access useful information. For example, the time spent on a web page by a specific user on different occasions of page requests will vary (Xinhua and Qiong, 2011). Where a long user session is present in the web access log, the page will be divided with the next session where a fixed time-out threshold is applied. In order to ensure page requests with long time intervals reflect the time a user spends on the requested pages, there is a need to adjust the time threshold used to define a session (Sengottuvelan et al., 2017).

Dynamic time-out adjustment has been used to address the challenges faced by the fixed time-out process (Halfaker et al., 2014; Sharma and Makhija,

2015). One of the key objectives is to ensure time intervals between page requests reflect user interests on the page based on the duration of the user visit. This approach identifies all page visits within a user session prior to making any time-out adjustments, thus avoiding misclassification of information requested by a user. For example, if a fixed time threshold is used to determine the interestingness of a web page, it can either be identified as useful or noise. In the dynamic time-out adjustment process, visit time is calculated for each page visit by the user by using a consecutive timestamp value. For example:

t_1 = the primal time-out of a web page

t_{new} = the time-out of a web page that is put into the current session

The average time-out, denoted as t' , is defined as: $t' = \frac{(t_1 + t_{new})}{n}$, where n is the total number of web pages considered. In order to apply the adjustment to other pages, the adjustment ration η is defined using the following equations:

$$\eta = \frac{(t' - t_1)}{t_1} = \frac{(t_{new} - t_1)}{n(t_1)} \quad (1)$$

To apply the adjustment to all pages, the adjusted time δ denotes

$$\delta = \delta_0(1 + \eta) = \delta_0(t_{new} + t_1) / n(t_0) \quad (2)$$

where, δ_0 denotes the time-out by the last adjustment time-out.

The dynamic threshold value considered in the proposed research uses the average time of visiting web pages to indicate the end of a session. Time on a web page, also referred to as page visit duration, is calculated for each page visited by the user by using a consecutive timestamp value. The average duration of a particular page is the average of all the times spent on that page. At the beginning of a new session, the initial time-out t_1 is set for each page, while the requested page is put into the current session t_{new} . The time-out will be computed dynamically in order to make it reflect the actual time a user spends on a web page. The dynamic adjustment means that only requests

with a long interval will be considered. Algorithm 2 describes the stages of determining user session based on dynamic time-out value.

Algorithm 2: Session Identification Based on Dynamic Time-out

```

Input:  $rec_n^j$  //A set of records for the  $j^{th}$  user
Output:  $S_i^j$  //A set of sessions for the  $j^{th}$  user
Begin
Read user log_file // extracted access logs from a web server
Sort all web logs by user_ID and Time_stamp
For every unique user_ID do
    Create a new user session  $i^j$ 
    If time interval  $T_{k+1} - T_k$  is  $< \delta$  //adjusted time-out
        Assign  $k^{th}$  into  $S_i^j$ 
    Else
        Create a new user session  $S_i^j + 1$ 
    End if
End for

```

The success of the web user mining process depends on the effective identification of user sessions implicitly recorded in a web log file (Pater and Parmar, 2014). A number of algorithms proposed by current research to define user sessions mainly rely on fixed time-out threshold value to determine the end of a session, as well as the beginning of a new one. However, the proposed research points out a number of limitations associated with fixed-time-out value. Fatima et al. (2016) argue that the user session identification approach based on dynamic time-out threshold value precisely captures user interests on the web as compared to a fixed time-out approach. This is due to the fact that it eliminates the assumptions made when using a fixed threshold value. The adjacent time-out value is also dynamic enough to reflect the interestingness of a web page in relation to the interests of a user.

3.3.3. Page View Identification

Page view is a collection of information on the web linked together in a particular page representing a user event (Srivastava et al., 2000). This is what the user actually sees while on a website. The page view identification process determines which pages accessed by a user lead to the display of web content the user is interested in. Identifying page views based on a user visit is heavily dependent on the layout and structure of a website (Zubi and

Raiani, 2014). Each page view can be considered a collection of web pages that represent information the user is interested in. However, where measures such as duration and frequency of visit are considered, the interests of a user will vary between the relevant page views.

3.4. Web User Profile Construction

A user profile is a collection of information that describes the interests of a user on the web (Gasparetti et al., 2014; Hasan et al., 2013). The main characteristic of a web user profile is the ability to determine the information from the web that defines the interests of a user. The initial step in constructing a web user profile is analysing web user access logs extracted from a web server. The pre-processing of web log data discussed earlier in this chapter is critical to user profiling, as well as extraction of useful information that reflect user interests.

The main actors in building a user profile are the user and the web (Amato and Straccia, 1999; Grčar et al., 2005). The user specifies what he/she is looking for while on the web. However, user interests may change over time and this is one of the aspects that need to be taken into account during the user profiling process. On the other hand, the web acts as the information source, where all different types of information the user might be interested in resides. The web aims to satisfy users by providing all information, but the key aspect is ensuring that the right information is available to the user at the right time. For example, one user may be searching for “python” due to an interest in computer programming language, while a different user will be interested in studying different species of reptile. Such requests should consider a user profile in order to provide results that meet the interests of a specific user. In essence, user profiling aims to address classification problems that contribute to noise in web data. Between these two actors, there are a number of key stages, which include: a user request to the web server, extracting and analysing user interest information, determining user interest level on visited web pages, etc.

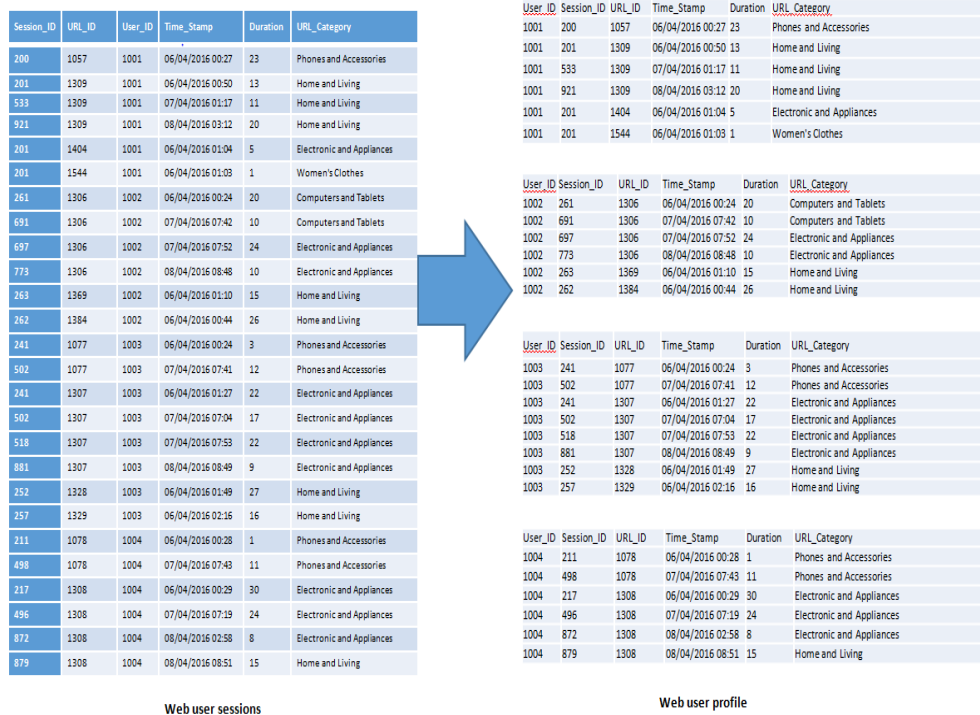


Figure 3.4: User profile construction

The user profiling process shown in **Figure 3.4** implicitly learns user interests from extracted web user logs. Huidrom and Bagoria (2013) consider user profiling based on web user access log files as one of the most efficient ways of determining the interestingness of web data in relation to a specific user. However, challenges such as the presence of noise levels in web log files hinder the process of finding useful information in relation to user interest. The following section explores the significance of web user profiling in the identification and subsequent elimination of noise web data.

3.5. Web User Profiling: its Significance in Noise Web Data Reduction

Every piece of information on the web that a user requests is recorded on a web server (Wiedmann et al., 2002). When user requests are combined with the information available on the web, a clear picture of user needs and interests is defined. Determining the interestingness of web data requires finding the level of user interest to ensure that problems with noise web data are addressed in relation to change in user interests. One way to ensure useful

information is available to a user at the right time is by building a user profile based on web log data. The main goal of user profiling is to learn user interests on the web and provide them with what they need without explicitly requesting input. Each web page in a user profile is presented in the form of a URL, from which user interest level can be measured; for example, based on duration and frequency of user visits.

Web user profiles are considered to be dynamic in relation to changes in user interests. A number of recent research works have emerged that address changes in user interest with regard to web data (Alphy and Prabakaran, 2015). When user interests change, the user profile needs to be dynamic enough to learn such changes and thus minimise the noise data suggested to a user. User interests can be quite wide and the user can at any time focus on a small subset of his/her broad interests. In the case of internet browsing, the entire set of user interests can include interests that are relevant to his/her job and hobbies. For instance, after the birth of a baby users will naturally be interested in parenting issues and their preferences in automobiles may change (Ahmed et al., 2011). Some events, such as planning a holiday, purchasing a car, obtaining a mortgage, etc., will lead to a marked change in user interests. Therefore, it is clear that user interests are subject to change over time. Due to such dynamic tendencies in the web usage mining process, it is important to consider various measures that are key in learning user interests. These include: duration, frequency and depth of a user visit to a web page, as well as how recently a user has visited a specific page. These measures are extensively analysed and discussed in the following chapter.

3.6. Chapter Summary

One of the proposed research aims is to learn the interests of a user and how noise in web data is affected by a dynamic change of user interests. In this chapter, a user profile that captures the interests of a user on a web page was defined. The rationale was based on the key aspects identified and discussed in previous chapters. For example, this thesis acknowledged that the interestingness of web data is influenced by user interests and dynamic changes in this. User interests can be subjective; in the context of the

proposed research, a web page is considered useful if it is available to a user when relevant, and otherwise it is noisy.

The key stages involved in defining a user profile were explored. These include: user and session identification stages of pre-processing web log data, which plays an important role in finding data on the web and defining a user and his/her interests. A session identification algorithm based on dynamic threshold value was considered when defining user sessions. The rationale for using the dynamic time-out threshold approach in the user session identification process is to ensure that the importance of a web page reflects the interests of a user. The proposed research's position aims to ensure that noise web data is defined, with clear consideration of user interests and their changes over time. Therefore, using the dynamic user session identification approach ensures all web pages visited by a user are considered when defining the level of user interest in requested web pages. In the next chapter, a number of the measures used to learn user interests on the web are critically evaluated. Even though many users may show an interest in the web pages requested, the level of interest varies. The objective of learning user interests is to examine how different measures affect the interestingness of web data. The outcome of the user interest learning process will lead to a user-driven approach to learning noise web data prior to elimination.

After collecting user interest information and building a user profile, it is important to determine to what extent the web data requested by the user is of interest so as to minimise nosiness in the web user profile. Current research considers various measures to define the interestingness of web data in relation to user interest levels. Measures such as duration and frequency of page visit are defined and critically evaluated in the following chapter.

Chapter 4: Learning Noise in Web Data

Chapter 3 identified and critically evaluated the various stages current research considers when defining a user profile based on user interest information extracted from web servers. One of the objectives of this thesis is to examine how user interests influence the interestingness of web data. This is to ensure that noise data eliminated from the web takes into account the user's change of interests. This chapter first explores how the interests of a user on the web are currently determined and how various measures impact the interestingness of web data. The chapter also attempts to address the following research question:

Question 2: What are the key indicators for learning user interests and how interests of a user could improve noise web data elimination?

4.1. Introduction

Nowadays, the most common challenge a web user faces is finding information of interest from the web without encountering a high volume of noise data. Website owners and developers are also facing challenges in catching up with the dynamic change in user interests in understanding the kind of information a user is interested in. In most cases, user interests on the web are assumed to be fixed over a given period of time (Qiu and Cho et al., 2006). For example, a student is likely to have a fixed interest in a given research domain, hence they will be interested in specific information from the web. However, it is also realistic to suggest that user interests change over time (Wang et al., 2013, Ko and Jiamthapthaksin, 2014). For example, at Christmas time, a person may be interested in shopping, but his/her interests will change during the summer period. Ahmed et al. (2011) and Cheng et al. (2015) suggest the use of time-variant data based on an implicit learning aid in determining interestingness of web data to a user, either dynamic or fixed. As defined and discussed in chapter 3 of this thesis, the implicit approach to learning user interest take into account measures such as duration and frequency of a user visit.

Based on the extracted web log data and subsequent creation of a user profile, the proposed research defines the level of user interests by taking into account key aspects. (1) The visiting time for a web page is an indicator of a user's interest level (Ahmed et al., 2011; Wu et al., 2014). The amount of time a user spends on a web page reflects its interestingness, which is defined as the degree of user interest in a web page (Zhang et al., 2007). (2) The frequency of a user visit to a web page is positively related to his/her interests (Rebon et al., 2015). Furthermore, duration and frequency of visit are interlinked because the longer the time spent on a web page, the higher the user preference for the web page visited. The proposed research aims to identify and examine various measures used by existing research to determine the interestingness of web page in relation to user interest. The outcome is to ascertain the impact of these measures in identifying noise web data when change of user interests are considered.

The rest of this chapter is organised as follows: section 4.2 critically evaluates how current research defines, learns and measures user interests while visiting a website. Section 4.3 examines how a change in user interests influences the interestingness of web data. Section 4.4 proposes a noise web data learning approach that considers change in user interests prior to noise elimination. Finally, section 4.5 discusses the critical aspects in relation to learning noise web data vis-à-vis the current situation, and then the chapter is summarised.

4.2. User Interest Learning

User interest information is regarded as the key indicator when learning how useful data is on the web. Zeng et al. (2012) argue that user interests are not only about finding web pages where a user spends more time, but also when a user appears to be interested or not in a given piece of information. The ability to build a user profile based on web log data is at the heart of learning the interests of a user in a given website (Dong et al., 2008; Bhargava et al., 2015). The learning approach mainly involves determining the interestingness of a web page taking into account change in user interests over time. This thesis proposes machine learning algorithms capable of learning user

interests on the web, as well as how change in user interests impacts the identification of noise web data. The proposed algorithm mainly relies on key user interest indicators to determine interestingness of web data (Claypool et al., 2001; Kim and Chan, 2005; Zahoor et al., 2015). **Figure 4.1** outlines the main stages of the user interest learning process, which are critically evaluated throughout this chapter, further **Appendix III** presents a use case for the proposed NWDL approach.

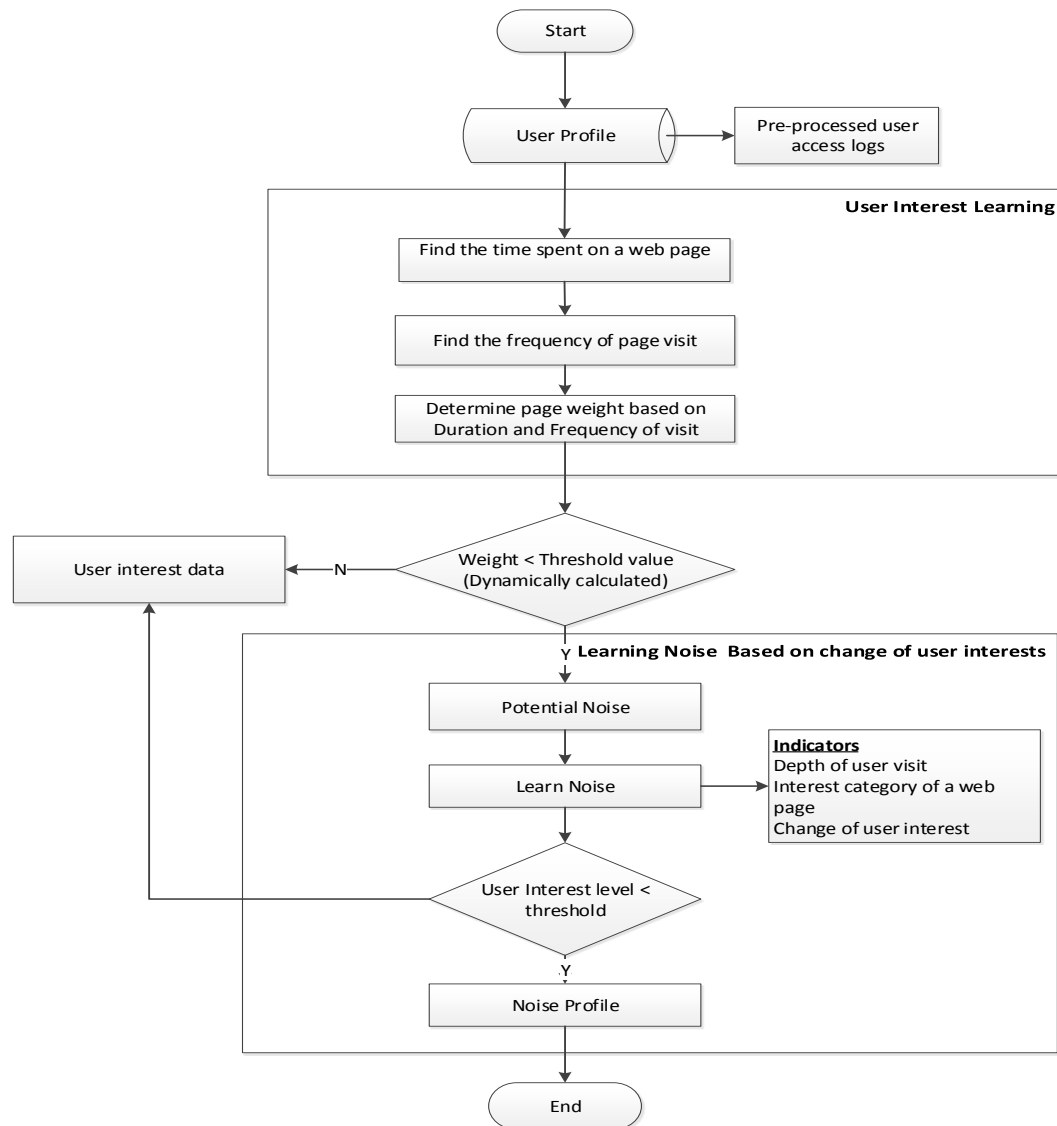


Figure 4.1: Noise web data learning process flow diagram

4.2.1. Identifying Key Indicators for Learning User Interest

User interest is expressed as the weight of a web page visited by a user (Wu and Liu, 2014). The weight of a web page can be determined using binary (0

and 1) or a function of parameters (Forsati and Meybodi, 2010). The binary approach identifies the existence or non-existence of a web page in a user session. This approach fails to measure the level of user interest in the visited web pages, however, because they are all treated equally regardless of the level of user interest. Not all pages present in a user session are interesting; a user can visit a page and find it irrelevant, but the page will be captured in the web log file as part of the user interest information. Page weight based on parameters, such as the frequency and duration of a user visit to a web page, provides an opportunity to measure interest degree in a web page in relation to varying user interests (Kabir et al., 2012). It is believed that the longer the time spent on a web page and the greater the number of visits to the same page, the higher the interest level (Suguna and Sharmila, 2013). Hence, the assumption is that the degree of interest is in proportion to the duration and frequency of a user visit (Wu et al., 2014).

In this thesis, the level of user interest in a web page is determined by a number of key indicators, including: page visit duration, which defines the length of time a user spends on a web page; frequency of page visit, which indicates the number of times a web page has been visited; depth of page visit by a user, which shows the path that leads to user interest information; and the frequency of a web page category, which signifies the preference for a web page category by a given user.

Page visit duration: Page visit duration is one of the indicators widely used to measure user interest level on a web page (Azimpour-Kivi and Azmi, 2011; Kim and Chan, 2005a). Naturally, page visit duration T_k^j is determined by the difference between the timestamp of the current page and the timestamp for the next page view (Yonghong et al., 2016).

$$T_k^j = ts_{k+1}^j - ts_k^j \quad (3)$$

where T_k^j is the time duration of the j^{th} user on the k^{th} web page, ts_{k+1}^j represents the timestamp of the next k^{th} web page and ts_k^j is the timestamp for the current page.

Therefore, the relative importance of each page to the j^{th} user is determined by the duration of the visit. Forsati and Meybodi (2010) acknowledge some of

the key aspects of using page visit duration: this reflects the relative importance of a page in relation to user interest; a user will spend more time on a page of interest. However, the web access logs do not contain enough information to determine the time a user exits a website (Chen and Su, 2013; Hofgesang, 2006). This makes it difficult to measure the time a user spent on the last page of their visit. For example, **Table 4** shows user sessions for a set of web pages whose time duration is determined based on a fixed threshold value.

Table 4: Visit duration for j^{th} user on k^{th} web page in i^{th} session

$user_id_i^j$	$url_id_i^j$	S_i^j	$time_stamp_i^j$	$T_{k_i}^j$
Session 1				
173	1	1	02/01/2016 08:02	05:44
173	54	1	02/01/2016 08:07	21:37
173	53	1	02/01/2016 08:29	04:01
173	56	1	02/01/2016 08:33	06:24
173	55	1	02/01/2016 08:39	00:00
Session 2				
173	55	2	02/01/2016 17:24	12:04
173	53	2	02/01/2016 17:36	11:50
173	54	2	02/01/2016 17:48	00:00

Time spent on url_id in session 1 is defined as $time_stamp$ for url_id 54 - $time_stamp$ for url_id 1

$$= 02/01/2016 08:07 - 02/01/2016 08:02 = 05:44min$$

However, the time spent on url_id 55 in session 1 is unknown (?) due to lack of next page timestamp

$$= ? - 02/01/2016 08:39 = 00:00min$$

The existing research argues that the exit page should be excluded from other user requests due to the lack of an exit timestamp (Prasad and Rao, 2016). However, the importance of the exit page will be affected by its misclassification, hence leading to a misinterpretation of user interests. The last page a user visits on a website is often referred to as the exit page (Sreedhar, 2016). The exit rate is defined as the number of times a user exits

from a particular page divided by the total page views. The exit rate signifies how likely a user is to end his/her journey on a given website. Ultimately, this draws up suggestions on whether the page is interesting or not, but because the time spent on the exit page cannot be determined, it will be difficult to measure the interestingness of such pages. Aldekhail (2016) and Sharma and Makhija (2015) claim that the last page the user visits is always a content page, which is believed to contain useful information (Kapusta et al., 2014; Munk et al., 2015). Therefore, if there is a lack of sufficient information to determine the page visit duration, then the last page, which is a content page, will be excluded when determining user interest level in web pages visited.

Finding how long a user spends on a website, i.e. session duration and the time spent by a user on a specific web page, is critical to addressing the above problem. Chapter 3 of this thesis addressed challenges faced when using fixed default value in the user session identification process. A dynamic time-out session identification approach was considered instead, where long intervals between user requests are recalculated, as shown in equations (1) and (2). Therefore, finding the time a user spent on the last page not only allows the interestingness of all web pages visited by a user to be determined, but also avoids misclassification, thus affecting the quality of information available to a specific user.

Due to the missing timestamp, the time spent on the last page is not calculated. The proposed research thus considers the missing value imputation technique to find the time a user spent on the exit page. Rahman and Islam (2011) and Aljuaid and Sasi (2016) apply the missing value imputation technique to address missing values in a dataset; their aim is to ensure the quality of data is not affected where missing values in a dataset cannot be determined. The proposed research borrows this concept because the interestingness of a web page will be affected if the time a user spends on the exit page cannot be determined. Even though this technique has been used by existing research, there is no existing work that has applied it in defining user session and calculating user page visit duration.

In this thesis, the missing value imputation technique uses previously known time a user spent on the same page to estimate the time duration on the exit page. This technique is considered effective where previous page visits and time spent are known. However, instead of using the time duration on the immediate web page, the proposed research considers average duration in the relevant web page estimated duration for the last three user requests. This takes into account any assumption that a user was struggling to find information of interest or the page was just a link to the destination page. The exit page visit duration, denoted as $T_{k_e}^j$, is determined using equation (4), where N represents the last three user requests for the k^{th} web page:

$$T_{k_e}^j = \frac{(T_{k_1}^j + T_{k_2}^j \dots + T_{k_N}^j)}{N} \quad (4)$$

Therefore, the average time duration on the k^{th} web page in the i^{th} user session for the i^{th} user is defined by equation (5)

$$AVT_{k_i}^j = \frac{\sum_{j=1}^{K_j} T_k^j}{K_i^j} \quad (5)$$

where,

$AVT_{k_i}^j$ = average visit duration on k^{th} web page by the j^{th} user in i^{th}

$T_{k_i}^j$ = page visit duration on k^{th} web page by the j^{th} user in i^{th} session

K_i^j = total number of page visited by the j^{th} user in the i^{th} user session

Page visit duration is widely considered a good indicator to measure user interest level on a web page (Ahmed et al., 2011; Wu et al., 2014). However, time alone cannot provide a 'clear picture' of how interesting a web page is to a given user because of the various reasons mentioned in the previous chapter. For example, different users have different reading speeds and the content of a web page may vary in structure and layout. Therefore, it is important to consider how frequently a user visits the page alongside the time spent when determining its interestingness.

Frequency of user visit: Frequency is the number of times a web page is accessed by a user within a session (Suguna and Sharmila, 2013). This is

considered one of the key indicators used to learn interestingness of a web page in relation to user interests (Booth and Jansen, 2009; Neelima and Rodda, 2016). Where a web page appears frequently in a user session, it might be considered interesting to the user. In the proposed research, frequency of a user visit to a web page is determined by the number of times k^{th} web page appears in i^{th} session for the j^{th} user. Frequency of the j^{th} user on k^{th} web page is defined in equation (6)

$$Freq_{k_i}^j = \frac{\sum_{j=1}^{K_j} url_{k_i}^j}{K_i^j} \quad (6)$$

where

$Freq_{k_i}^j$ = Frequency for the j^{th} user visit on k^{th} web page

$\sum_{j=1}^{K_j} url_{k_i}^j$ = the number of times k^{th} web page appears in the i^{th} user session

K_i^j = total number of page visited by the j^{th} user in the i^{th} user session

The frequency of a user visit may be higher, but the time spent on the web page lower. Therefore, finding the weight of a web page in relation to user interest level involves two aspects: the time spent by a user visiting a web page and the frequency of visits to a web page.

4.2.2. Interestingness of a web page based on time and frequency of user visits

Duration and frequency of page visits are measures widely used to determine the interestingness of a web page (Grace et al., 2011; Wang et al., 2013; Chitraa and Thanamani, 2013). Each web page in a user profile is assigned a weight to reflect the level of user interest. This weight defines the interestingness of a web page to a given user (Kabir et al., 2012; Wei et al., 2015). For each web page visited by a user, the corresponding weight is determined taking into account the amount of time spent and how often the page is visited. Finding the weight of a web page aids in keeping a user profile relevant by identifying web pages that reflect varying user interests. The weight is determined by the degree of the user's interest in the k^{th} web page in i^{th} session, as defined in equation (7):

$$W_{k_i}^j = \sum_{j=1}^J T_{k_i}^j * Freq_{k_i}^j \quad (7)$$

Where $W_{k_i}^j$ = weight of k^{th} web page in i^{th} session for the j^{th} user
 $T_{k_i}^j$ = page visit duration on k^{th} web page by the j^{th} user in i^{th} session
 $Freq_{k_i}^j$ = frequency of the j^{th} user visits to k^{th} web page in i^{th} session

Therefore, interest is calculated by the ratio of the total amount of time spent on a page to the number of times a page was visited by the j^{th} user in i^{th} session.

4.2.3. Influence of Time and Frequency of Web Page Visits on Noise Elimination

Current research acknowledges that frequency and duration of page visits are two major indicators of user's interest levels on a web page (Nanda et al., 2014; Kabir et al., 2012; Liu and Kešelj, 2007). However, the influence of these two measures in determining the interestingness of web data varies; for example, Hofgesang (2006) believes that the frequency of a visit to a web page is a much more relevant indicator of user interest, while Kim and Chan (2016) and Gauch et al. (2012) argue that the interest of a user in a web page is better reflected by the amount of time said user spends on the page. Kabir et al. (2012) and Holub and Bielikova (2010) suggest that time and frequency measures have equal importance in learning the interestingness of a web page.

For example, in **Table 5** URL_ID 1 was visited five times, but the average time duration is lower than that of URL_ID 20, which was visited once. This is due to the structure of a web page where a user's landing page is the home page prior to visiting interest pages. Therefore, a quick move to another page from the entry/landing page reflects its interestingness to a user.

Table 5: Time Duration versus Frequency of User Visit

URL_ID	Frequency	Visit_Duration
56	5	4.19
1	5	1.24
55	4	7.71
53	4	3.5
54	3	6.5
20	1	10.52

URL_ID 1 in **Table 5** is the homepage or user entry page to a website, which is thus traversed more often than the intermediate pages that are more likely to be of user interest. The frequency of a user visit to a web page will influence the interest level of a web page. The proposed research observes the following: (1) a user has no option other than to use the land page to get to the interest page; (2) a user frequently visits a page with the expectation of finding useful information.

Despite the influence of time and frequency measures in learning the interestingness of web data, the proposed research points out some challenges associated with these measures. Firstly, it is recognised that the more time a user spends on a web page, the more interesting the page is (Tan et al., 2012; Umamaheswar and Srivatsa, 2014). The amount of time a user spends on a web page varies from one user to another, mainly due to familiarity with the website and reading speeds. A user struggling to find information of interest may also take longer on a web page, or they may attend to other activities outside the page. Secondly, web pages visited within a session can either be auxiliary or content pages (Munk et al., 2015). Auxiliary pages help a user to find web pages that are of interest; they act as a visiting path to a user 'destination'. The frequency of this type of page will be high, but the duration of the visit will be low, which therefore suggests that frequency on its own may fail to determine the interestingness of a web page. Content pages, as defined in chapters 1 and 2, are pages where a user can find useful information.

The proposed research argues that relying solely on frequency and time duration is inadequate in determining the user interest level of a web page. As

a result, it is difficult to identify and eliminate noise web data based on duration and frequency of page visit. In order to address the above challenges, the proposed research introduces additional measures to learn user interest levels on web pages prior to noise elimination. These include:

1. Depth of user visit, i.e. the path taken to find information of interest. The sequence of page visits by a user is positively related to his/her interest (Rebon et al., 2015; Sambhanthan and Good, 2013). However, very often, a user's visiting path is influenced by the structure and layout of a website, which means that some web page act as a link to interesting pages. Auxiliary pages can either be noise or useful subject to time and changes in user interests. The goal of analysing a user path is to understand varying user interests on the web, how layout and structure of the web influence the interestingness of its data and, more particularly, the impact of user visiting path on the identification of noise web data.
2. Interest category of a web page – a web page category is defined as a set of related web pages in a website (Mishra et al., 2012). As the web evolves, new web pages emerge with no history of user interest (Hu et al., 2007). It is therefore possible to consider useful information as noise due to a lack of previous interest from users. In order to minimise the loss of useful information otherwise considered noise, the proposed research work learns its interestingness based on user's category of interest.

4.2.4. Learning User Interest Based on Depth of User Visit

The journey a user takes on the web is represented as a path (Singh et al., 2013; Joshila Grace et al., 2011), a route used to navigate through a website in order to find the page of interest. Generally, all the web pages a user visits, whether they are content or auxiliary pages, represent information a user might be interested in, but it is up to the user's judgment to determine to what extent the information is of interest (Gasparetti et al., 2014). Given that this perception is mainly dependent on the layout and structure of a website, the interestingness of a web page can be misinterpreted, especially where

frequency of visits is given weight; hence it is difficult to identify noise from user interest data (Hofgesang, 2006). Besides time and frequency of a user visit to a web page, this thesis thus proposes depth of a user visit to learn the interestingness of web data. Jansen et al. (2009) and Lagun and Lalmas (2016) define depth of a user visit as the average number of web pages requested by a user during a single session.

A user visit to a website is presented as a path, a sequence of web pages requested by a user in a session. $s_i = (url_1, url_2, \dots, url_k)$, for $k > 2$, where any ordered pair of k^{th} web page represents the j^{th} user visiting path. A user's visiting path, which defines the depth of visit, is shown in **Figure 4.2**.

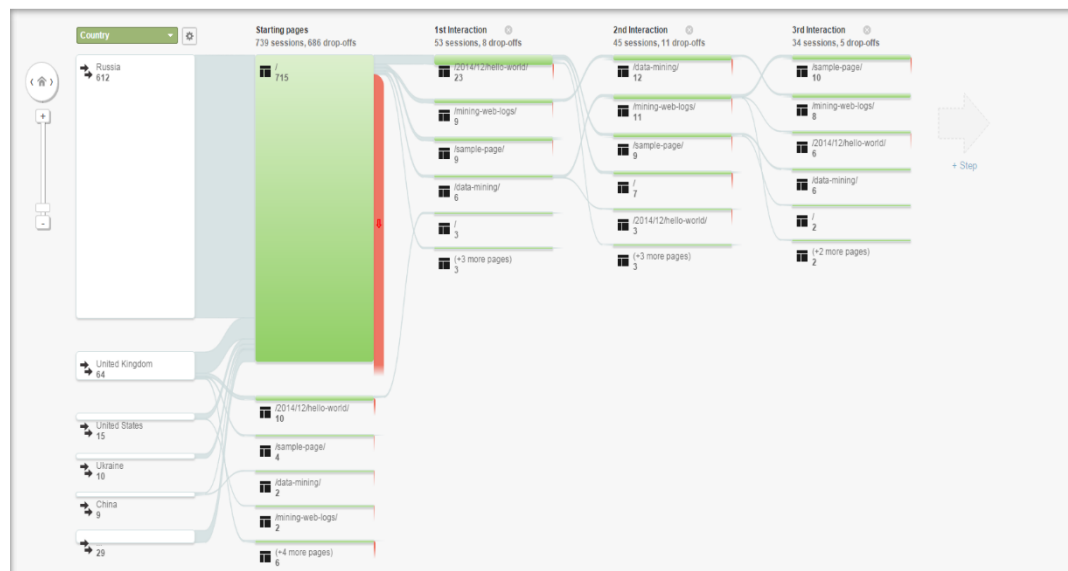


Figure 4.2: Depth of a user visit to a website

In most cases, users tend to visit a web page through the home page. However, this will require a user to visit more pages further down from the home page in order to find interesting web pages. Throughout this process a number of pages have been requested and the duration of every page visit is recorded. Therefore, if the interest level of a web page is determined by the average number of pages viewed in a user session, the interestingness of a web page will be misinterpreted.

Depth of a user visit specifically measures the level of user interest while visiting a specific website. It shows how far a user can go on a website before

finding a page of interest, and points out where a user drops by exiting the website. Unlike time and frequency measures, depth of a user visit reveals something more about a web user. For example, based on depth of a user visit to a given website, it can be ascertained whether: (1) a user is actively interested in specific content occasionally or regularly; (2) a visit to a specific page influences the interestingness of the subsequent page. Thus, depth of a user visit to a web page plays a critical role in understanding varying user interests on a specific website. The proposed **algorithm 3** defines a user's visiting path in order to differentiate noise from useful web pages, not only using auxiliary or content type web pages, but based on the interest levels of a user.

Algorithm 3: Depth of User Visit

Input: Extracted web user logs

Output: A set of web pages visited by the j^{th} user

1. Define the j^{th} user profile
2. For each web page in the i^{th} for the j^{th} user profile
3. Find the web page category
4. If two web pages from the same category are both included in i^{th} session
5. flag_Link = True; //a link between two web page from the same category is found
6. Else
7. flag_Link = False; //no link between the two web pages
8. Else if (flag_Link == True)
9. out_List.put (url₁: url₂) ; // web pages visited by the j^{th} user are connected
10. End if
11. End For
12. End

The proposed research considers depth of the j^{th} user visit not only in terms of number of page views, but also the path a user takes to traverse a website. For example, users can hit the home page of the site and, after few seconds, proceed to a sub-section of the home page, i.e. blogs. After spending some time on the page, a user will either move to another page or exit the website. Even though the j^{th} user is likely to visit other web pages to get to the information of interest, it is difficult to suggest that every page visit is of user interest unless measures such as time duration and frequency of visits over a number of sessions are jointly considered. Therefore, the path taken by the

j^{th} user from entry to exit page and the weight associated with each k^{th} web page are critical to determining the interestingness of a web page.

Depth of user visit: its significance to the user interest learning process:

A user visit to a website is usually random with no logical procedure (Booth and Jansen, 2009). For example, a user can enter a website through the home page or directly visit a specific web page via a URL. In order to learn user interests, the path taken by a user from one page to another immediately following each web page plays a critical role in the process. Some user interest indicators can be considered more important than others. For example, the depth of a user visit may be more significant than the frequency. Even though the user may visit one specific web page more frequently, this does not mean they are interested (Kim and Chan, 2013). Therefore, analysing the path taken by a user to get to the information of interest aids in learning the interestingness of each web page a user visits. The depth of a user visit considered by the proposed research aims to contribute to noise web data reduction in the following ways:

- i. Discovering the user visiting path contributes to developing dynamic websites where only user interest content is made available to a user without necessarily considering how the website is structured.
- ii. Identifying and determining user visiting paths can result in a better understanding of how users visit a website; it identifies users with similar information needs and can also aid in predicting how frequently user interests change.

In summary, if a website has a high page depth in a relatively unimportant part of the site, this may suggest that a user is finding it difficult to locate the information he/she is interested in. To identify and determine user interest levels based on the depth of visit, it is also important to also consider user interest based on web page category. This is due to the fact that the layout and structure of the web is based on relationships between content, in the sense that information on the web is grouped into categories. The following sub-section of this chapter explores web page category and its significance to learning noise in web data.

4.2.5. Interestingness of a Web Page Based on User Interest Category

Web users can spend a considerable amount of time looking for information of interest, but in some cases they may fail to find something that specifically fits their interest. Understanding how long a user spends on a specific category of a web page, rather than a specific web page, aids in learning user interests as well as the interestingness of web data. Learning user interests specifically from URLs accessed by a user can be limited by the fact that only visited web pages are considered (Lavanya and Vardhini, 2014; Poo et al., 2003). As the web evolves, new information in the form of web pages is added to a website. During the web usage mining process, such web pages are likely to either be identified as noise, thus eliminated, or suggested to a user without any learning to determine a user's interestingness.

In order to address this challenge, a learning approach based on a user interest category is proposed in this thesis. The proposed research argues that learning user interest based on the category of a web page aids in understanding the user's potential areas of interest. This is due to the fact that if a user regularly visits web pages within a given category, it is possible that he/she will find other topics within the specified category that meet their interests. For example, a computer student regularly visits pages containing information about 'programming in R', but they have not shown any interest in 'Adobe analytics'. Even though both subjects are within the data science category, 'programming in R' will be more interesting than 'Adobe Analytics'. Therefore, the interestingness of a web page category is important as it implies the significance of the web page to a user's interest (Nanda et al., 2014). The proposed research considers category as a set of web pages, whereas a web page is the internal representation of a category. For example, **Figure 4.3** shows an example of a category of web pages in an ecommerce website.

<p>Home & Living</p> <ul style="list-style-type: none"> Ribbon Kitchen & Dining room appliances Bedding sets & accessories Home Decor Household Cleaning Tools & Accessories Bathroom Products Home Storage & Organization Pet Products Furniture Curtains Home Textile Event & Party Supplies Clocks 	<p>Office Products</p> <ul style="list-style-type: none"> Chair Mats File & Storage Cabinets Desks & Workstations Lighting Shelving Tables Carts & Stands Footrests Bookcases Chairs & Sofas Computer Armoires 	<p>Computers & Tablets</p> <ul style="list-style-type: none"> Notebooks Tablet PC Netbooks Ultrabooks Macbooks <p>Phones & Accessories</p> <ul style="list-style-type: none"> Smart Phones Smart Watches Phablets Featured Phones
---	--	--

Figure 4.3: Categories of web pages

Source: <https://www.kilimall.co.ke/> accessed on 26/4/2017

User visits to a web page category over time signifies the interest level of a user. However, as user interests change over time, user interest categories also change. For example, from the Football World Cup, to summer shopping, to Christmas and winter shopping, etc. Therefore, in order to learn the interestingness of web data in line with a change of user interests, it is important to identify information from the web that has not only been visited by a user, but within a category of web pages. The proposed research considers the following criteria when defining the interest category of a web page:

- i. Visiting frequency: the number of times a user visits a web page category within a specific period of time. For example, the number of web pages within the sports category visited by a user over a specified period of time.
- ii. Length of visit: the duration of a user visit to a web page category. For example, if the amount of time a user spends on a given web page category is above a specified threshold, the category is considered interesting.

A user profile, as defined in the previous chapter, plays a key role in finding user interest information prior to learning the level of interest. For instance, user interest information, such as number of page views, duration of page visits and sequence of page visits, which is defined by the depth of visit, are

key to learning the interestingness of a web page. The previous sections of this chapter examine how duration and frequency of visits to a web page influence the interestingness of a web page. This section further explores the interest category of a web page and ways in which user interest levels are influenced by the category of web pages.

If we define web page category as $M = (ct_1, ct_2, \dots, ct_m)$, where $m \in M$ is the total number of categories of web page. A category visiting path $Ct_{k_i}^j$ = a sequence of m^{th} category visited by the j^{th} user during i^{th} session. For example, the browsing path of j^{th} user is $\{ct_1(url_1, url_2, \dots, url_k), ct_2(url_1, url_2, \dots, url_k), \dots, ct_m(url_1, url_2, \dots, url_k)\}$; this represents that the j^{th} user visits cat_m , then visits url_k which belongs to cat_m .

The proposed research aim is to determine the interestingness of a web page based on the following:

- The frequency of the j^{th} user visit to m^{th} category.
- The time spent by the j^{th} user on m^{th} category.

Interest category based on frequency of user visits: Frequency of visits to a web page category presented as $Freq_{m_i}^j$ is the number of times the j^{th} user visits the m^{th} web page category during i^{th} session. Each visit to k^{th} web page is accumulated to the respective m^{th} category. Interestingness of a web category based on frequency of user visit is therefore defined using equation (8).

$$Freq_{m_i}^j = \frac{\sum_{j=1}^{K_j} K_m^i}{K_i^j} \quad (8)$$

where,

$Freq_{m_i}^j$ = the frequency of m^{th} web page category in i^{th} session for the j^{th} user.

$\sum_{j=1}^{K_j} K_m^i$ = number of k^{th} web page of m^{th} category for the j^{th} user in i^{th} session.

K_i^j = total number of k^{th} web pages visited by j^{th} user in i^{th} session.

Illustrative example: The proposed research examines the relationship between a user and the category of a web page based on frequency of visits. For illustrative purposes, consider M the total number of categories visited by the j^{th} user and $freq_m^j$ the number of times the j^{th} user visits the m^{th} category. When the j^{th} user visits the m^{th} category, its frequency is determined by the number of associated k^{th} web pages, as shown in **Figure 4.4**.

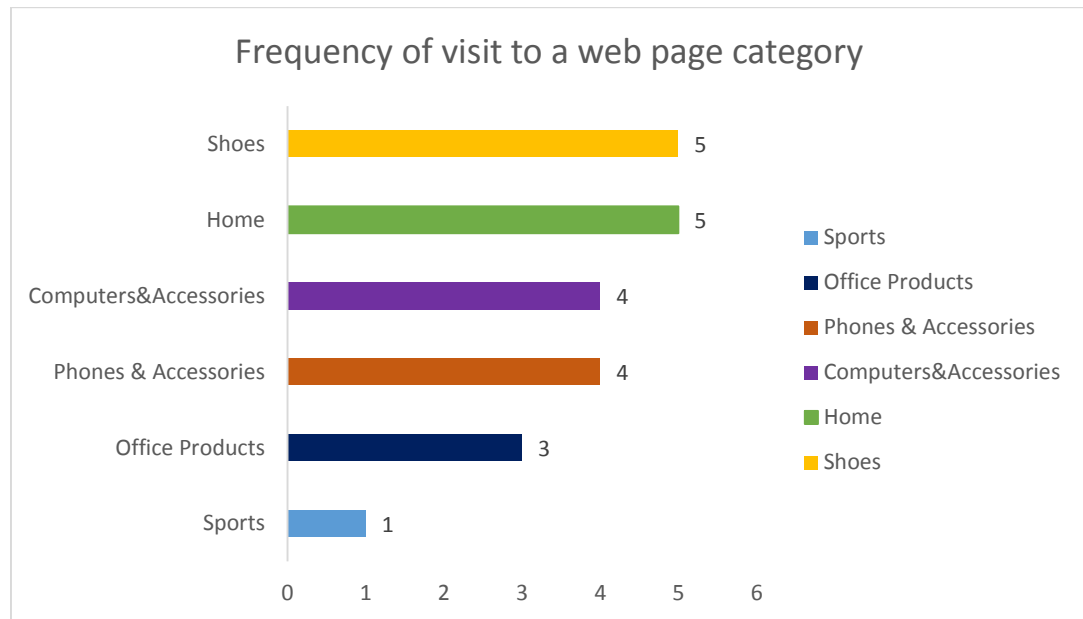


Figure 4.4: Frequency of visits to a web page category

Figure 4.4: Frequency of visits to a web page category. The values shown in this figure are the sum of the number of times a user requested a web page that belongs to the specified web category. Frequency interest is therefore defined by the ratio of number of k^{th} web page of m^{th} category for the j^{th} user in i^{th} session to the total number of web pages requested by j^{th} user in i^{th} session. Frequency of a user visit to a web page category provides an in-depth insight into interest in the category.

Chitraa and Thanamani (2013) argue that frequency of a user visit to a specific web page signifies the interestingness of the requested web page. Unlike frequency of visit to the k^{th} web page, as defined in equation (4), frequency of visit to a web page category measures the interestingness of a web page category to a specific user. The goal is to ensure that website owners/developers understand the type of information a user is likely to be interested in at any given time. This also aids in understanding the interestingness of a web category based on requested web pages that relate to the category of interest.

As shown in **Figure 4.4**, the number of visits to the home page is high, but it is difficult to determine whether a user is really interested in this category without finding the total amount of time spent on this category. Therefore, the length of visit to a web page category is key to determining the interestingness of a web page category and thus eliminating any web pages that seem irrelevant to a user.

Interest category based on length of user visit: The amount of time the j^{th} user spends on m^{th} category of a web page reflects its interestingness to a user. The duration of each k^{th} web page visit by the j^{th} user is accumulated for its respective m^{th} category, which is presented as $T_{m_i}^j$. The interestingness of m^{th} category based on time is defined as the ratio of visit duration to the category by the j^{th} user in i^{th} session, as defined in equation (9):

$$L_{m_i}^j = \frac{\sum_{j=1}^{K_j} T_{m_i}^j}{T_i^j} \quad (9)$$

where

$L_{m_i}^j$ = length of time spent by the j^{th} user on the m^{th} web page category.

$T_{m_i}^j$ = visit duration to m^{th} category by the j^{th} user in i^{th} session.

T_i^j = the total time spent by the j^{th} user in i^{th} session.

Defining the interest category of a web page based on length of user visit aids in describing the relationship between the category of a web page and users who visit such categories within a specified time. Length of access time for each web page is accumulated in its category, as shown in **Table 6**. This is a representation of the visit duration for the j^{th} user in the corresponding category.

Table 6: Length of visit to a web page category

Web Page Categories	URL_ID					
	1	20	53	54	55	56
Computers & Accessories			17.61			
Home	0.98					
Office Products				26.76		
Phones & Accessories					33.81	
Shoes						23.75
Sports		10.52				

Defining user interest category: discussion of critical aspects: The proposed research's viewpoint is that user interest level is better defined by interest category than the requested web pages. The advantages of this approach to learning user interests include: (1) ability to understand changes in user interests and allowing website owners/developers to manage such changes, thus address noisiness in a user profile; (2) justifying the proposed research's viewpoint that the number of times a user visits a web page does not reflect its interestingness. This also addresses the impact of duration of a user visit to a web page, where a user spends more time on a web page, but fails to find information of interest. Therefore, it is important to consider how long a user spends on a given web page category rather than on a specific web page.

Figure 4.5 presents a comparison between frequency and length of visit to m^{th} category of a web page. It can be observed that where only one measure is considered in determining how useful a web page category is to a user, the outcome can be misleading in web page classification. Classification of web pages without learning their interestingness leads to noisiness in a user profile.

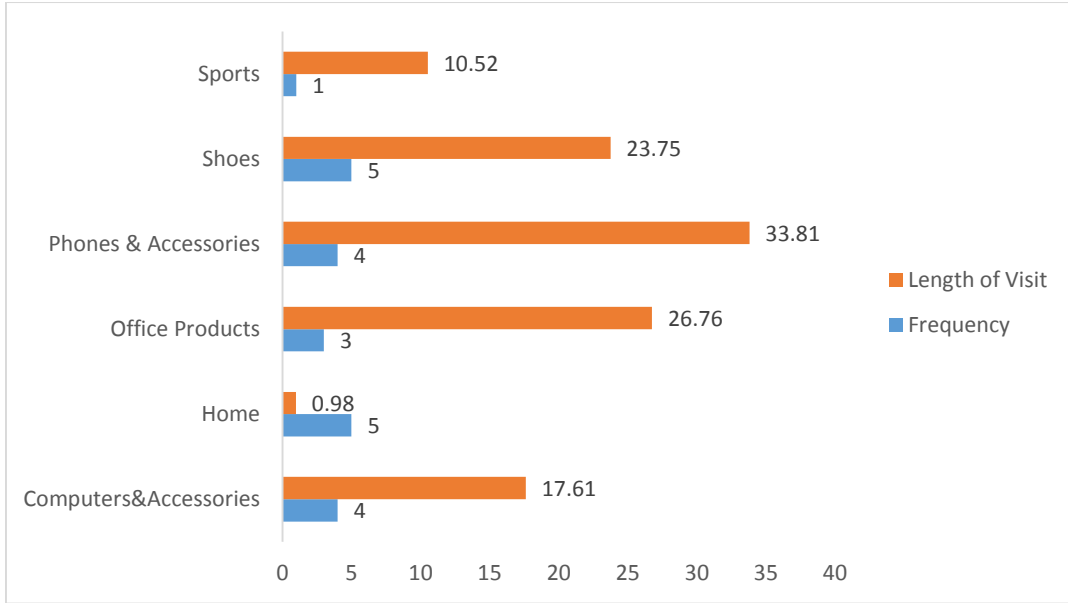


Figure 4.5: Frequency versus visit to a web page category

Figure 4.5: Frequency versus visit to a web page category. The home page shows minimal length of time compared to the sports category. In contrast, the frequency of visit to home page shown in Figure 4.4 are higher than the sports category. Therefore, both frequency and time of visit to a web page category are key in determining the interestingness of a web page category.

The web page category weighting denoted as $W_{m_i}^j$ is defined based on the frequency and length of a user visit to a web page category, as shown in equation (10). The objective of defining the interest category weighting is to understand the interestingness m^{th} category in line with a user's change of interests.

$$W_{m_i}^j = Freq_{m_i}^j * L_{m_i}^j \quad (10)$$

where, $W_{m_i}^j$ is the weight of m^{th} category in the i^{th} session for the j^{th} user,

In summary, the user interest category based on frequency and length of time plays a critical role in finding useful information otherwise considered as noise, where user interest cannot be directly determined. The interest category of a web page reflects its interestingness to a user. It also signifies how identification of noise web data can be influenced by the interest category of a web page. Some information can be seasonal or recur at a given time; hence interest will also vary over time. Since the interest level changes dynamically, a user profile is expected to adjust to changes, as well as to ensure any new web page added is not eliminated as noise but also considered.

4.3. Addressing Dynamic Change in User Interests

The existing research acknowledges that user interests tend to change over a period of time (Ko and Jiamthapthaksin, 2014; Jiang and Sha, 2015). For example, the entire set of user interests can include interests that are relevant to a wedding and thereafter it may change to shopping for a new baby. Therefore, a user profile will contain web pages visited by a user within a given period of time. Thus, if the user's interests were to change over time, the profile would reflect these changes by adding web pages to categories recently viewed and removing web pages from categories no longer found interesting. It is important to note that user interests in web data can change quickly, while others can change gradually over time. Therefore, learning changes in user interest aids in understanding how interesting the requested web pages are to a user. The following are some of the critical aspects examined in this section:

- *Last date of visit to a web page category:* even though a user may frequently request web pages from a specific category, how long is it since the last request?
- *Frequency of change in user interest:* it is important to understand how often user interests change over time. Therefore, defining change frequency for requested web page aids in looking into factors that trigger such changes. For example, if the requested information is seasonal.

The proposed research aims to explore the recency measure, which is considered critical to determining the interestingness of web data. Recency in web data mining is the time within which information on the web is considered relevant to the needs/interests of a user (Dong et al., 2010; Aly et al., 2013; Chakraborty et al., 2017). In other words, recency is measured by the 'freshness' of web data i.e., the time information on the web is interestingness to a specific user which is defined by how recent was a user visit to the web page.

4.3.1. Interestingness of a Web Page Based on Recency of Visit

In the previous sections of this chapter, frequency of user visit to a website has been broadly discussed. Learning how frequently a user visits a web page allows website owners/developers to understand user interest level in order to ensure only relevant information is suggested. It is equally important to examine how long it is since a user requested a specific page. The aim of this is to ascertain whether the interest of a user in such web page or its category has changed. The proposed research considers the recency measure to reveal how recently a web page was visited as compared to how frequently it is requested by a user.

Recency is defined as the time period since the last user visit to a web page (Chakraborty et al., 2017). The best way to capture the interestingness of a web page based on its recency is by determining the number of days since the last occurrence of k^{th} web page in the j^{th} user profile. For example, if x number of days have passed since the last time a user visited the 'home and living' web page category, then it might indicate that though the user was interested before, this may not be the case any longer; the user might have already acquired what he/she was looking for. Instead of setting a fixed time that determines the interestingness of a web page in a user profile, learning interestingness of a web page gradually over time is considered ideal. For instance, a user may have recently visited a specific category of a web page, but the frequency and time spent on the category is gradually decreasing. Even though website owners/developers will continue suggesting relevant information, the time will come when a user will no longer be interested, hence such information becomes noise.

The proposed research considers the interest forgetting function to determine the rate at which user interests change over time. Hawalah and Fasli (2015) and Wu et al. (2015) applied the interest forgetting function to remove data that is outdated from a user profile. A gradual forgetting function determines the weight of the web page category based on the time of its occurrence. The weight is dynamically adjusted each time a user visits the associated web page; thus the interestingness of a web page category is measured based on its appearance in a user profile over time. The concept behind this function is

that interestingness of web data diminishes gradually with time. Therefore, the most recent web page visits in a user profile are considered more interesting than the old ones (Schwab et al., 2001; Huang and Yu, 2014).

To ensure gradual changes in user interests are managed, a half-life of interest, denoted as h_l , is defined. The half-life span of interest is the rate at which user interests change over a period of time (Suksawatchon et al., 2015; Tavakolian et al., 2012). The half-life span is key in learning the interestingness of web pages in a user profile over time, taking into account user interests as they change; it defines the average rate at which a web page within a user profile becomes noise. Existing research argues that user interest reduces to half in a week (Gu et al., 2014; Sugiyama et al., 2004). However, it is important to define a half-life value that reflects the significant change in user interests. In the proposed research, the recency measure based on time of visit is determined using equation (11).

$$Rec_m^j = \frac{\log 2}{h_l} (td_0 - td_n) \quad (11)$$

where Rec_m^j = the recency adjustment weight of m^{th} category for the j^{th} user, td_0 = the current date and td_n = the date of the last occurrence of k^{th} web page in the j^{th} user profile, h_l user is the half-life (in days).

4.3.2. Dynamic Threshold Values

The objective of using the support threshold is to find a single point, which is used to determine the interestingness of a web page. Existing research works in the web usage mining process are based on either the standard/uniform support threshold or the dynamic support threshold (Ou et al., 2008). The standard/uniform support threshold is static and thus does not take into consideration a number of key aspects, such as change of user interests during classification of web page based on user interest levels. User interest information collected explicitly usually relies on a standard threshold value to measure the level of user interest in requested web pages. A number of critical issues arise when using standard threshold support. For instance, Hawalah and Fasli (2015) and Kavitha and Kalpana (2017) argue that the threshold support value remains the same and does not learn change of user interests;

where a low threshold value is assigned, it may lead to a high level of noisy web pages identified as useful and vice-versa.

To overcome challenges associated with uniform threshold, a dynamic threshold support is considered in order to ensure changes in user interests are adequately addressed. Ying et al. (2012) and Wei et al. (2014) observe that the dynamic threshold support value is mainly used to determine a range of user activities in relation to various measures used to learn user interests, for example time and frequency of a user visit to a web page. Based on these measures, it is possible to determine the interestingness of web data in relation to varying user interests when the threshold value is dynamically defined. Dynamic threshold support plays a critical role in ensuring dynamic changes of user interests are considered during the identification and subsequent removal of noise in web data. Instead of setting a standard threshold value based on the weight of a web page, the proposed research makes use of dynamic threshold value. Hawalah and Fasli (2014) proposed a mechanism to calculate a threshold value that reflects changes in users' browsing behaviour; their mechanism is based on frequency of user interest in a given web page category. The proposed research considers a similar approach to Hawalah and Fasli, but instead of using frequency, interest category weight, which incorporates both frequency and length of visit to a web page category, is used. This is to ensure the interestingness of web pages in a user profile reflect the time and number of visits. Recency adjustment measure is considered mainly to learn changes in user interests. Subsequently, the process of identifying noise web data is managed taking into account changes in user interests. The following key issues are considered:

- A web user profile contains all web pages that are perceived as interesting to a given user, but whose degree of interest varies.
- Generally, useful information on the web is assigned high weight, unlike noise data, which is assigned low weight.

- The threshold support to determine the interestingness of a web page is defined based on the interest weight of a web page category (equation 9), as well as the recency adjustment weight (equation 10).

The recency adjustment measure is first defined, as in equation (12), then the standard deviation (α) of k^{th} web page for the j^{th} user.

$$\alpha = \sqrt{\frac{1}{N} \sum_{i=1}^K (W_{m_i}^j - Rec_m^j)^2} \quad (12)$$

where α is the standard deviation, K is the total number of web pages in i^{th} user session, $W_{m_i}^j$ is the interest weight of m^{th} category in i^{th} session for the j^{th} user, and Rec_m^j is the recency adjustment weight of m^{th} category for the j^{th} user. The threshold value is defined using equation (13).

$$Threshold = \alpha + \left(\frac{\sum_{i=1}^K W_{m_i}^j}{K} \right) \quad (13)$$

where $\sum_{i=1}^K W_{m_i}^j$ is the interest weight of m^{th} category in i^{th} session for the j^{th} user

The dynamic threshold value not only possesses the ability to manage a change of user interests, but also ensures useful information is not lost as a result of uniform threshold value. In view of this, dynamic threshold support is key to determining user interest level in a visited web page prior to noise data elimination. This is due to the fact that user interests evolve over a given period of time. Therefore, in order to determine a dynamic support threshold value, the evolving nature of web data, as well as user interest, should be considered.

Some critical issues justify considering dynamic threshold measure in the noise web data reduction process. (1) A web page with lower frequency and time of visit will be considered irrelevant to a given user profile. Where a threshold value is set too high, interest pages with lower threshold values will not be found. On the other hand, where the threshold value is set low, a lot of irrelevant web pages will be considered useful. (2) Interestingness of

information on the web varies. Seasonal web data tends to attract more attention than general information that is accessed on a daily basis. For example, the 2018 Football World Cup tournament will attract more traffic than the ongoing Brexit news. It is therefore important to understand the nature of web data and the interest a user expresses. This is because the threshold value set for these two types of information will vary.

In summary, dynamic threshold values in web data classification play a critical role in ensuring dynamic changes in user interests are considered during the noise web data reduction process. Where a standard threshold value is applied in the web usage mining process, it will be difficult to obtain results that conform to a user change in interests. For example, a high threshold support value will yield less useful information and a low threshold support value will yield too many results. Dynamic threshold values are defined by learning previous user interest in a requested web page. As user interests change over time, the dynamic values also change. Second, it is important to acknowledge that noisy data can be potentially useful in future. User interests change and therefore so does the interestingness of web data. Using static thresholds will impact the quality of information available to a user given the fact that current interest data can be noise in future and vice-versa. In the proposed research, dynamic threshold values are used in the following scenarios: (1) implicit learning of user interests, i.e. when users do not directly reflect the interestingness of a web page, but instead their activities on the web determine the importance of a web page; (2) during classification of a web page, it is subject to change of interest.

4.4. Learning Noise Web Data by Classification

Classification is one of the key processes in web usage mining and it has a significant impact on addressing problems with noise web data (Nanda et al., 2014). Classification takes an object and assigns a class label to it based on its attributes. Traditionally, classification aids in the creation of a user profile based on the user's level of interest in web pages requested. Tang et al. (2010) argue that user profiling is usually seen as a data classification problem because the interests of a user with regard to a given web page change over

time. The objective of carrying out classification is to determine the target class of each record in a web log file based on varying user interests. Existing researches have proposed various algorithms to address data classification problems. For example, Santra and Jayasudha (2012) proposed an algorithm to find interested and not interested users.

Web page classification can be divided into binary or multiclass classification (Qi and Davison, 2009; Waegeman et al., 2011). Binary classification defines data into two classes; based on user interest level, a web page can either be noise or useful. Multiple classification problems arise when data does not simply belong to one particular class, however. For example, a web page can be useful, noise or useful and noise. The proposed research recognises that determining the interestingness of a web page as either interest or noise opens up some critical issues in the web data reduction process. Given that user interest varies and the web evolves, there is a need to learn a web page taking into account user interest prior to determining its classification. When addressing classification problems, such as incorrect classification of web pages due to varying user interests, a set of web pages and a class label are provided. For example, if the weight of a web page requested by a user meets a specified threshold, it will be considered interesting and otherwise it is noise. The class label 'interesting' and 'noise' are specified to allow for the classification of web pages.

Defining class labels: The objective of defining a class is to learn whether each page visit meets the set criteria. For example, k^{th} web page in the j^{th} user profile is assigned to a class based on the level of user interest. A class is a label whose value can be described based on varying levels of user interest in visited web pages. For example, each page visit by a user can be of interest, potential noise or noise. For each of the weighted web pages, a classifier is defined that reflects the interest levels of a user in relation to the corresponding web pages, which is later used for the classification process. Even though all web pages in a user profile can be considered useful with varying interest levels, not all are of interest. Moreover, the user is not directly involved in determining the interestingness of a web page. The proposed research therefore considers the measures discussed in this chapter, for

example time and frequency of page visit, user visit depth and frequency of a web page category, to learn user interest levels. In this thesis, three classes are defined: interest, potential noise and noise data. Once classes have been defined, web pages are assigned to a class based on user interest level:

Interest Class: Web pages whose interest level meets the threshold value are assigned to the interest class. In order to ensure the interestingness of a web page reflects the user interest, the threshold value is dynamically defined so as to avoid low or high threshold value, as in the case of standard or uniform threshold values.

Potential Noise Class: As the web evolves, new information emerges that a user is likely to not have visited. With no interestingness identified, such web pages will be identified as noise and subsequently eliminated. In order to avoid this, a potential noise class is defined that will consider the interest category of a web page to learn its interestingness.

Noise Class: Noise web pages are determined by the interestingness of a web page taking into account all interest measures defined in this chapter; for example, duration and frequency of user visits to a web page. Given that the interests of a user change, what is noise today can be useful a different time, and for this reason dynamic threshold values protect against loss of useful information, as well as ensuring minimise noise levels.

Consider a class label $CL = (cl_1, cl_2, \dots, cl_n, \dots, cl_N)$ where N is the maximum number of predefined classes. For illustrative purposes, the following classes are considered: cl_1 = interest class, cl_2 = potential noise class and cl_3 = noise class.

Algorithm 4: Learn noise web data

Input: Weighted url_k for the j^{th} user

Output: Class label based on level of user interest

1. Define the j^{th} user profile
2. for each k^{th} web page in j^{th} user profile do
3. Determine the weight of k^{th} web page using eq.5
4. if url_k weight > threshold set then
5. assign to cl_1
6. else
7. assign to cl_2
8. end if
9. End for
10. for cl_2 do
11. Create a simple page link to the j^{th} user profile
12. Determine interest category using Eq.10
13. Determine interestingness of k^{th} web page using Eq.11
14. if $Freq_m^j < \text{threshold}$ set then
15. assign to cl_3
16. else
17. update cl_1
18. End else
19. End If
20. End for
21. End

In summary, the proposed research identifies and classifies a web page based on time, frequency and depth of a user visit to this web page. If the interest level cannot be determined for a specific web page, web page category interest is used to determine whether a user has previously shown interest in the related category. The holistic approach to learning noise in web data taking into account all specified measures is critical to finding useful information that is specific to a user.

4.5. Noise Web Data Learning: its Significance to Web Usage Mining

The noise web data learning approach identifies different types of web data and determines their interestingness to a user by taking into account a number of measures, i.e. time and frequency, depth of a user visit to a web page and interest category of a web page prior to elimination. Unlike existing research works where noise in web data is identified and eliminated based on the relationship with the main content, the proposed approach considers a web user a key character in determining the interestingness of web data. The

proposed research argues that the importance of web data is mainly dependent on what is interesting to a user at a given period of time. Therefore, web data can be irrelevant or noisy if it does not satisfy the interests of a user. A user profile that reflects the interests of a user subject to time is one of the key aspects that can significantly improve elimination of noise in web data with minimal loss of useful information. User interests evolve and so does web data, therefore, the evolution of a user profile and its adaptation to emerging data sources and associated web data reflects how user interests change over time (Mezghani et al., 2014). In order to ensure the user profile reflects changes in user interests, it is important to understand the interestingness of web data taking into account key measures, such as frequency and depth of a user visit to a web page.

Learning noise in web data also reflects the following critical aspects discussed above in this thesis:

1. Weighted web pages based on time and frequency of page visit – weighted k^{th} web pages are based on dynamic user session identification, which not only allows for dynamism in user interests, but also ensures all web pages visited by a user are considered.
2. Web page category interest – the proposed research mainly considers this approach to minimise the loss of useful information where the interests of a user with regard to a specific web page cannot be defined.
3. Dynamic threshold value – user interest change as web data evolves; where a uniform threshold value is set, it is difficult to determine the interestingness of a web page. Therefore, with dynamic threshold support, classification of web pages will vary as user interest change, thus minimising the loss of useful information, especially when threshold values are either set high or low.

4.6. Chapter Summary

In this chapter, a number of machine learning algorithms are proposed to address problems with noise in web data by learning user interests prior to elimination. One of the objectives the proposed research examines is how change in user interests influences the interestingness of web data. The

approach to learning noise web data proposed in this thesis is based on the fact that the interestingness of web data is influenced by the user. This is to ensure noise levels in web data are minimised without any loss of useful information. The proposed algorithms will contribute to addressing the problems with noise web data identified in chapter 1 of this thesis. Learning user interests and how changes in user interests influence the interestingness of web data helps to overcome challenges such as incorrect classification of web data, hence noisiness.

It is important to note that each proposed algorithm contributes towards the main research objective, i.e. learning noise in web data prior to elimination. For example:

1. Duration of page visit based on dynamic user session identification: ensures all web pages requested by a user are considered when determining user interest level, thus minimising loss of useful information.
2. Depth of a user visit: identifies web pages that lead a user to interesting pages, i.e. eliminates auxiliary pages from influencing interestingness of a web page. This is due to the fact that the time and frequency of a user visit to a web page are measures widely used to determine the level of user interest.
3. Interest category of a web page: the proposed research uses web page categories to learn the interests of a user and their dynamic change of web data. Web page category provides a clear picture of where a user passes through to find information of interest. As user interests change, the web evolves and new information emerges. Therefore, determining user interest based on web page category opens up the opportunity to identify useful information that would otherwise be classed as noise.

This chapter attempts to answer the second research question outlined in chapter 1:

What are the key indicators of learning user interests and how interests of a user could influence the identification of noise web data?

Findings: Where the interests of a user change, noise in web data also changes. Where user interests are seasonal, interestingness of available information varies as well. In essence, the web should be able to cope with dynamic changes in a user subject to time of interest. This, therefore, demonstrates that it is difficult to rely on existing noise web data patterns that are determined based on previous user activities. In order to identify and understand how dynamic changes in user interests impact the identification of noise in web data, the depth of user visit, as well as the use of dynamic threshold value, plays a critical role. This is in addition to measures such as time and frequency of user visit that are widely used by current research to learn the interestingness of web data.

In the following chapter, an experimental design setup is presented to validate the proposed algorithms in relation to the performance of the existing machine learning tools. A number of different experimental directions are introduced in order to evaluate the performance of the proposed algorithms compared to existing tools used to identify and eliminate noise web data.

Chapter 5: Experimental Design Setup

Chapter one of this thesis defines problems with noisy web data, explores the contribution made by current research to address such problems and more importantly the gap in research which the proposed research aims to address. A critical analysis and evaluation of current and relevant research justifies a need to propose a new approach that will attempt to bridge the gap. A user profiling approach introduced in chapter 3 identifies key aspects that contribute towards learning user interests and their changes over time, thus the ability to effectively learn noisy web data. In order to ensure noisy web data is eliminated with consideration to the user's change of interests, a number of machine learning algorithms are proposed in chapter 4. The proposed algorithms take into account key indicators such as duration, frequency and depth of user visit to measure the interestingness of a web data in relation to user interests. In order to find out if the proposed research contributes to addressing the defined research gap, it is important to validate the performance of the proposed algorithms against the existing tools. Performance is measured in terms of how well the proposed algorithms address problems with noisy web data, for example, identification and classification of noisy web data in relation to user needs and interests as they change over time.

This chapter is organised into the following sections: Section 5.1 revisits the proposed research aims as well as the question this chapter attempts to answer. Section 5.2 provides a description of the dataset used in conducting experiments. Section 5.3 introduces a number of experiments based on the proposed research objectives, the algorithms as well as the experimental data. Section 5.4 summarises the chapter.

5.1. Introduction

The proposed research is inspired by the gap in research defined after a crucial review and analysis of current research work that addresses problems with noise web data. Some of the key aspects the proposed research explore is ways in which existing research define and identify noise web data if user interest influences the process of identifying and eliminating noise web data. Even though existing research has made a significant contribution to addressing defined problems as discussed in chapter 1, a number of critical issues are still open to investigation. For instance, existing research proposes a number of machine learning tools to identify and eliminate noise web data, however, change of user interests over time prove to be critical towards interestingness of web data. Based on current literature, it is not clear how noise web data is currently addressed in relation to change o user interests. The proposed research seeks to find out how a change of user interests over a time influence the interestingness of web data. In essence, the viewpoint of this research is that the process of identifying and eliminating noise web data is dependent on the user's change of interests and not the structure or layout of the website. Subsequently, the research question which this chapter aims to answer is, **how learning noise web data could minimise the loss of useful information without affecting its quality and interestingness?**

A number of experiments are conducted in order to answer the above question. The key aspects discussed in this chapter are derived from the proposed algorithms. For example, (1) Demonstrate how page visit duration that considers time on the exit page impacts a user's level of interest on a web page. (2) To present using real-life data how a dynamic change of user interests influence the interestingness of web data. (3) To evaluate the overall performance of noise web data learning approach against current tools used to address noise web data problems.

5.2. Experimental Data Preparation

The main source of data used by the proposed research is anonymised web user access logs from a web server. Data extracted from web servers in form of web user logs is considered as the most suitable dataset used in the discovery of useful information based on user interests (Munka and Drlíka, 2011; Adeniyi et al., 2016). The experiments are conducted using two datasets as shown in **Appendix IV**, each of the datasets is user access logs extracted from a web server.

Dataset 1: User logs extracted from an e-commerce web server are used. The extracted web logs contain (*User_ID, Page_ID, Time_Stamp, Category,*). The IP address which identifies a user has been anonymized and User_ID introduced in its place. The extracted logs cover a period of 90 days, the date range considered is reasonable enough to provide an insight into user interests as well as demonstrate any dynamic changes within the specified period. The proposed tool keeps track of what a user is interested in, the level of interest in relation to time, frequency and depth of visit, as well as change of interest within the specified period.

Dataset 2: The second dataset is an extract of web visitor interests available at Kaggle dataset store <https://www.kaggle.com/uciml/identifying-interesting-web-pages>. Dataset 2 was extracted to generate a training and testing set. The training set is used to learn the proposed algorithms, which is then used to generate input to the proposed NWDL. The test set was retained as the validation data for testing, it was also used to evaluate the performance of proposed noise web data learning approach

The proposed research aims to demonstrate in a more simplified way how the proposed algorithms are implemented using real-life web data. A dimensional data model shown in **Figure 5.1** is considered for analysing raw web log data prior to using it to train the proposed algorithms. Moreover, it brings together data from different sources and creates a single and consistent user view, a sample dataset is shown in **Appendix IV**. Given the structure of the extracted web log, User, Page and Session Data are the main dimensional tables

considered in this research. *User dimension* simply identifies a unique user taking into account other attributes, such as browser and operating system. *Page dimension* describes the requested web page and the category the page belongs to. *Session dimension* defines the start and end of a user visit to a website, the time and day the user requested the page, the type of a visitor, i.e., new or returning. This dimension is critical given that it aids in learning user interest level on requested pages. The master table i.e., the measure of interest connects all dimensional tables.

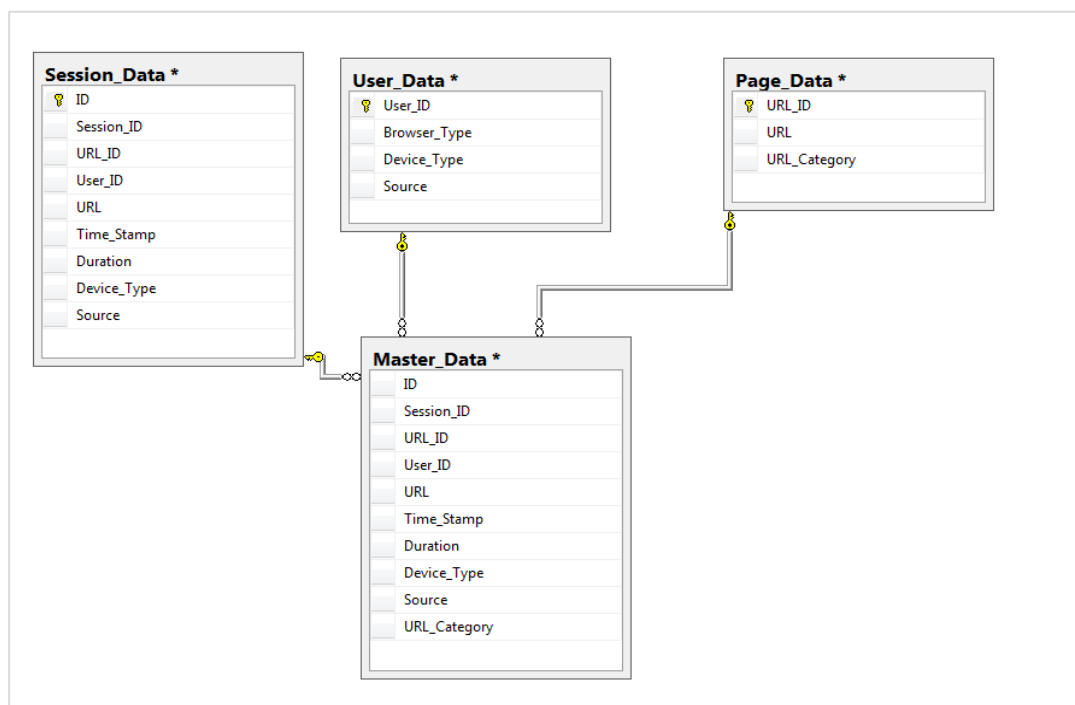


Figure 5.1: Database Model

5.3. Experimental Setup

Experiments conducted in this section are based on the dataset introduced in the previous section. In this section, a number of experiments are conducted which are based on the proposed algorithms defined in chapter 4. The objective is to determine if the proposed algorithms contribute to addressing problems with noise web data as defined and analysed in previous chapters. Some of the key aspects addressed in this chapter through a number of experiments include:

- To demonstrate by using real-life the influence of user time on the exit page in defining interestingness of a web page.

- The effect of dynamic change of user interests towards the classification of web pages visited by a user. The classification process is based on a user's level of interest thus defining web pages that are of user interest or noise.
- Overall evaluation of noise web data learning approach based on user interest.

The proposed algorithms are independent of a testing platform thus they can be implemented and test on any other platform as well as the dataset. However, for validation purposes RapidMiner studio and Orange are considered, this is open source machine learning platform which has been widely used by current research in data mining and knowledge discovery (Kasliwal and Katkar, 2015). They provide different machine learning algorithms which are used to solve data classification problems.

In order to initialize and validate the performance of the proposed noise web data learning approach, the following steps are considered.

Step 1: Parsing raw web user access logs – raw web log data is loaded into the SQL server after pre-processing.

Step 2: Attribute selection i.e. data attributes that are used to learn interestingness of web data based on user interest levels. Such attributes include a timestamp, page visit duration, URL_Category.

Step 3: We then execute our proposed algorithms to identify, determine the weights of visited web pages and learn user interest prior to the classification of a web page to associated class.

Step 4: Train and validate the proposed tool – the output is then evaluated against the input using various evaluation metrics.

5.3.1. Interestingness of web page based on exit page user visit duration

The aim of conducting this experiment is to demonstrate how user session identification process influence the interestingness of a web page. Two different user sessions are considered; user session defined by fixed time-out threshold of 30min and another with a dynamic time-out adjustment value of

which the time spent on the exit page is determined using missing value imputation technique as defined in equation (4). This experiment used web log data generated over 7 days, the choice of the specified time period to ensure that the results obtained reflect user interest over time considerable enough to learn their interest.

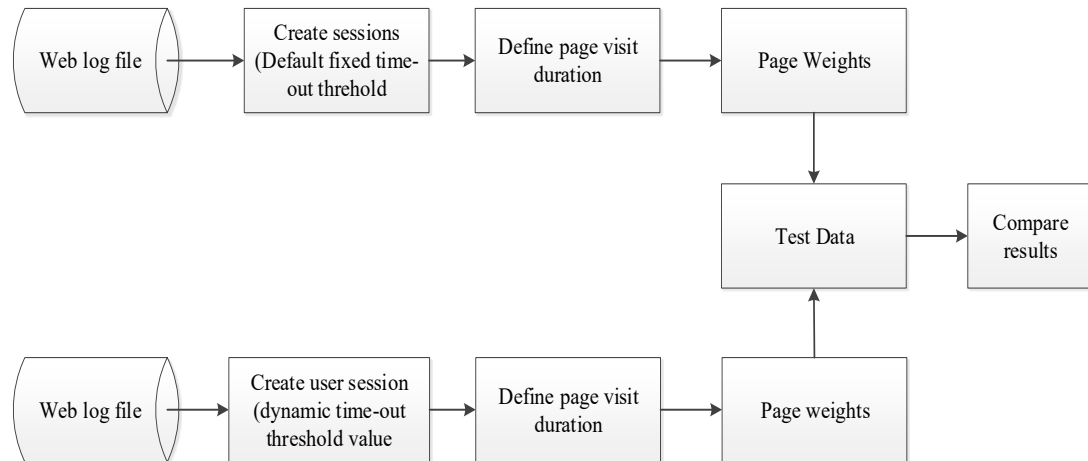


Figure 5.2: Page visit duration based on fixed vs dynamic time-out session identification

Section 4.2.2 defines interestingness of a web page based on duration and frequency of a user visit to a web page. Therefore interest of a web page is defined by the ratio of the average page visit duration to the number of times a page was visited by the j^{th} user in i^{th} session. The average page visit duration consider time spent on exit page as defined in equation (5). **Figure 5.3** shows the significance of page weights where time spent on exit page is considered when determining the interestingness of a web page in i^{th} session for the j^{th} user.

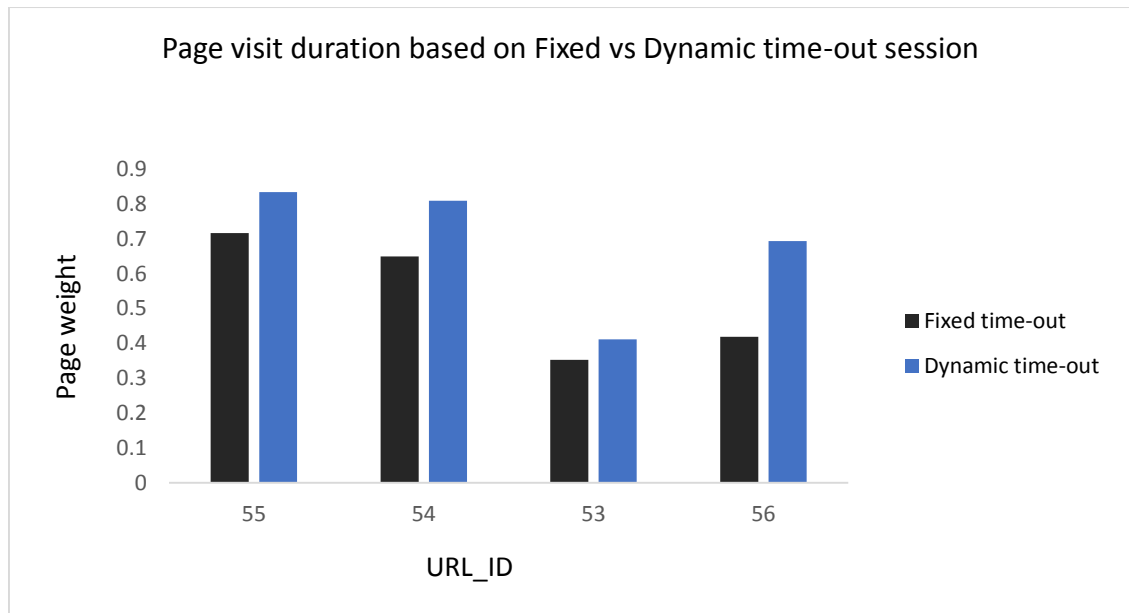


Figure 5.3: Comparing page weight based on fixed vs dynamic time-out session identification.

Discussion of Results

The weighted web page based on fixed time-out and dynamic time-out threshold value user session identification are presented in **Figure 5.3**. The average page visit duration determined by fixed time-out ignores user's varying time spent on a web page. A user might spend more or less based on time of access, how familiar they are to the website, layout and structure of the website, etc. For example, a new user might take longer to find useful information on a web page while it might take a returning user less time to find a web of his/her interest. Therefore, where a user session and interestingness of a web page is determined based on a fixed time-out threshold value, the output is likely to impact the process of identifying and eliminating noise in web data. It is therefore important to generate a user session that reflects the actual time a user spends on a web page. This is because if user sessions are not dynamic enough to capture user interest levels, the usefulness of a web page present in a web user profile will be compromised.

Unlike page weights determined by visit duration based on fixed time-out user sessions and frequency of visit, the exit page visit duration has some significant impact on interestingness of a web page. However, the proposed research argue that time and frequency measure alone cannot justify user

interest level on a web page. This research work further proposes learning interestingness of a web page based on the depth of a user visit to a web page, user interest category. The weight of a web page category which implements depth and user interest category signifies interestingness of a web page as defined in equation (10). It is important to understand how change of user interest influence interestingness of web data. The proposed research's viewpoint is that as user interest change, the process of identifying and eliminating noise web data is also affected. If not managed well, useful information can easily be eliminated as well as noise data would be suggested to a user. In the next section, the impact of the dynamic change of user interests is discussed using a wide range of user access log data.

5.3.2. Interestingness of Web Data Based On Dynamic Change of User Interest

The objective of this experiment is to investigate whether change of user interest influence identification and subsequent elimination of noise web data. To illustrate the significance of user's change of interest in noise web data learning process, the proposed research first defines a time period over which interest of a user on visited web page is learnt. The experiment conducted considered a 90 days' period. This is to ensure that there is enough gap for learning any change of user interest over the specified period. A user profile will contain web pages visited by a user within the specified period. If the user interests were to change over time, the weight of a web page category will be adjusted accordingly. This process takes into account the length of time interestingness of a web page is considered for the learning process.

A number of preliminary experiments were conducted to demonstrate how each measure, i.e., frequency and length of a user visit to a web page category performed on this type of data. The 90 days extracted web user access logs data was divided into two parts. The first-month data was used to build a user profile and subsequently learn their interest on visited web pages. The second part of the data was used to learn changes to user interests in order to define the interestingness of visited web pages prior to noise elimination.

The experiment is based on the recency adjustment measure defined by equation (11). Recency adjustment measure which is based on a forgetting function examines how a change of user interest influence interestingness of web pages in a user profile. The function does not just eliminate web pages as the user stops visiting the page but learns the user's interest change. Changes to user interest affect the weighting of a web page thus its classification status, i.e., either interesting or noise. Therefore, it is important to ensure a more dynamic and flexible approach to determine the interestingness of the page in line with the change of user interests. The goal is to ensure that the process of identifying and eliminating noise in web data is dynamic enough to reflect the change of user interest.

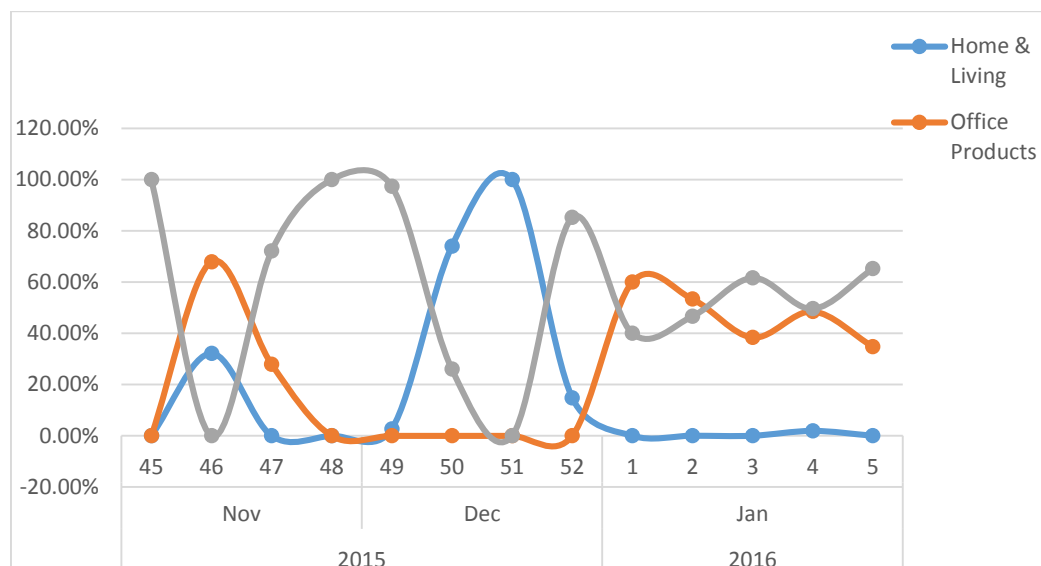


Figure 5.4: Dynamic change of user interest over 90 days' period

Discussion of Results

Figure 5.4, shows varying user interest on web page categories visited over 90 days period. It can be observed that the interestingness of a web page varies with time, for example, user interests in office products declined in a week from week 48- 52 but again increased after week 1. Subsequently, interest in home and living category decreased after week 52 with no visits for the next five weeks. When user interest level on a web page category is gradually decreasing, it signifies that a user is no longer interested in the category hence its noisiness. Therefore, its classification will be influenced by

the threshold value Equation (13) as user interests level change. On the other hand, the weight of web page category a user has recently shown interest will gradually increase, ultimately the classification will also be affected. The advantage of using this measure is that interestingness of a web page is not only determined by the duration and frequency of a visit to a web page but also over a time interests of a user might change. Without considering the dynamic change of user interest, the amount of useful information otherwise eliminated a noise is likely to be high. Therefore, the choice of using dynamic measures to learn noise web data improves the quality of information in a web user profile. As a result, the proposed measures supports the claim that noise web data learning approach is user-centric. This takes into account a number of factors which outperforms existing machine learning tools applied in the noise web data reduction process.

Dynamic threshold value: In chapter four of this thesis, the proposed research explores the impact of defining a threshold value that reflects the change of user interests. It is well recognised that interestingness of web page is dependent interest level of a user, which implies that the selection of a threshold value is a critical aspect in the process. For instance, if the threshold value is set low, then the output is likely to contain high noise levels. On the other hand, if the threshold is set high, then the chance of eliminating useful information is equally high. The dynamic threshold value considered by the proposed research take into account the change of user interest over time as defined in equation (13)

The results shown in figures 5.5a and 5.5b are a classification of web pages which reflect the change of user interests. The process considers a dynamic threshold value which is determined in accordance with the interestingness of a web page over a period of time. For example, results shown in figure 5.5a are based on user visits within a week, we observe that the visited web pages are 43% interest, 34% potential noise and 21% noise. In chapter four, the proposed research acknowledges the existing research's viewpoint that user interest reduces to half in a week. However, it is important to consider such changes over a period of time and this experiment consider interestingness of a web page over 7 weeks.

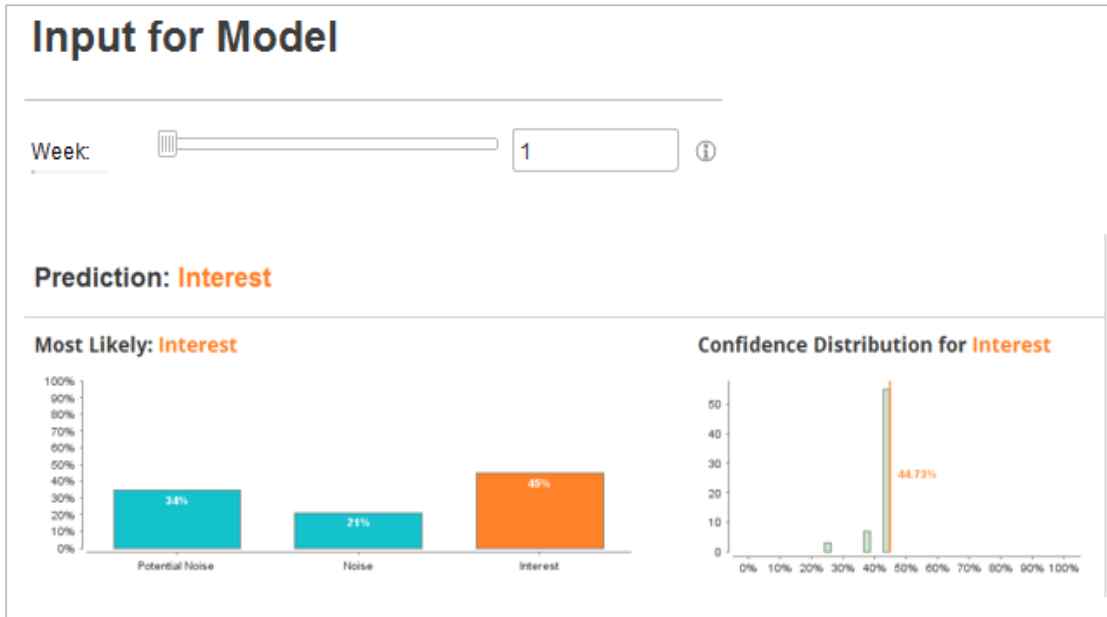


Figure 5.5a: User interest level after one week

The dynamic threshold value is determined based on the recency adjustment measure. As time since the last page visit increase, the interestingness of a web page decrease thus becoming noise to a specific user. Figure 5.5b the experimental results where user interest on visited web page is examined for 7 weeks, it can be observed that the 78% increase in noise class is due to the time since a user expressed interest on web pages previously visited.

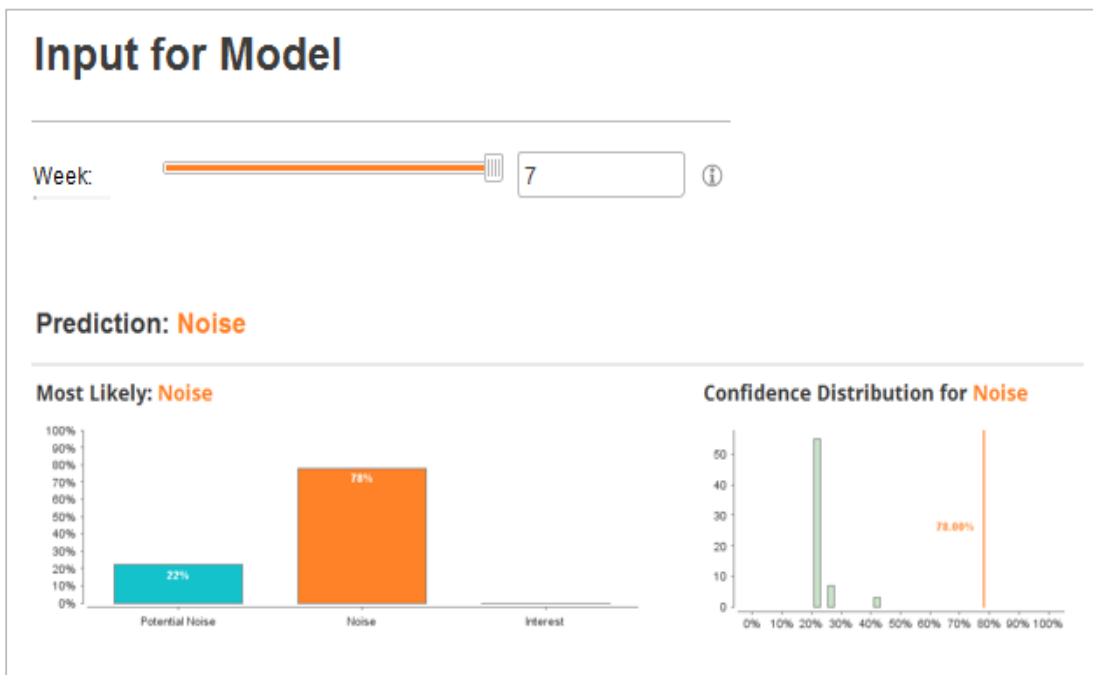


Figure 5.5b: User interest level after 7 weeks

In summary, the experiment reveals that as user interest change over time, interestingness of a web page is affected thus its classification. In essence, the level of nosiness in web data will be better managed if a change of user interest over time is considered prior to elimination. The next section present a noise web data learning approach taking into account the dynamic change of user interests discussed above. The performance of the proposed NWDL approach is measured by how well it defines and classify visited web pages in line with the level of interest.

5.3.3. The overall performance of the proposed noise web data learning approach

This experiment is based on noise web data learning approach which consists of a number of machine learning approach defined in the previous chapter. The previous sections of this chapter demonstrate the influence of various measures that are based on proposed algorithms. Further, this section focuses on the overall performance of the proposed noise web data learning approach, the objective is to examine its contribution towards addressing problems with noise web data. More particularly problems that lead to misclassification of web pages visited by a user hence noisiness. An approach to learning noise web data is based on web page classification as discussed in chapter 4, section 4.4. The experimental procedure is defined as follows:

Step 1: A classifier is built which describes a predefined class label. This is also referred to as a training phase where the proposed algorithms define the classifier based on various user interest measures

Step 2: Learn the user interest level. This is the learning process where the training data is examined by the proposed measures. This process includes (1) Depth of visit which identifies web pages that are of user interest. (2) User interest category which defines the level of user interest on categories associated with requested pages (3) Recency adjustment measure which learns dynamic change of user interests and adjusts page weights accordingly. (4) The dynamic threshold value which assigns a threshold value that reflects the interestingness of a web page.

Step 3: Identify to which of the defined class label a weighted web page belongs to. The dynamic threshold value is applied in relation to the interestingness of a web page.

Step 4: Validate the performance of proposed noise web data learning approach. The test data is used to evaluate how accurate the proposed approach is in terms of assigning web pages to a predefined class. Classification accuracy (CA) is considered a success when test data is correctly classified. Classification accuracy in this experiment is measured using a confusion matrix.

The above steps are presented in a process flow diagram shown in Figure 5.6

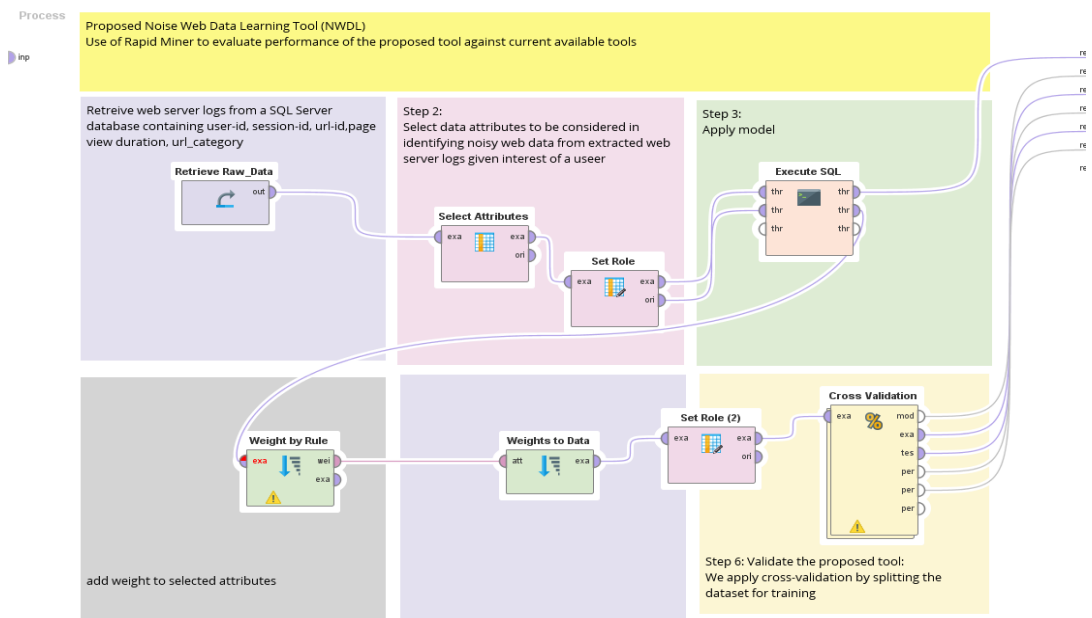


Figure 5.6: Noise Web Data Learning Process Flow Diagram

Experimental Results

The previous sections of this chapter present the performance of the proposed algorithms as well as their contributions to addressing defined problems. For example, the influence of the dynamic threshold value to the classification of web pages in a user profile prior to noise elimination process. In this section, the focus is to evaluate how noise web data learning approach perform on the test data provided. In this experiment, the overall performance of the proposed

NWDL which integrates the proposed algorithms is presented. The proposed NWDL demonstrates how dynamic change in user interest affects the elimination of noise in a web user profile. As shown in **Figure 5.7**, dynamic changes in user interests to a specific web page category are obtained using equation 3. We observe that the weight of a web page category which takes into account measures such as recency is considered a key to determining interestingness of a web page. Therefore, the classification accuracy is perceived dynamic and user-centric in the sense that the level of a user interest a web page affects its classification.



Figure 5.7a: Overall Performance of NWDL

Table View Plot View

accuracy: 60.00%

	true Potential Noise	true Interest	true Noise	class precision
pred. Potential Noise	6	0	2	75.00%
pred. Interest	13	31	10	57.41%
pred. Noise	1	0	2	66.67%
class recall	30.00%	100.00%	14.29%	

Figure 5.7a: Performance Evaluation - NWDL using Confusion Matrix

The results shown in figures 5.7a and 5.7b are based on user interest indicators explored in chapter 4 of this thesis. These include; duration, frequency and interest weight based on a web page category. Page visit duration presented in this thesis considers exit page which current research exclude in learning user interests. Moreover, this chapter demonstrates the significance of the exit page to the learning process and subsequent noise web data elimination process. It can be observed that the classification accuracy of NWDL is 60%. The potential noise class is introduced to learn interestingness of a web page considered noise, mainly due to use of a fixed threshold value in web page classification and lack of previous user visit to define interestingness of a web page. Therefore, the classification of the web page as interest and noise is better presented by learning noisiness using a potential class prior to determining if it is useful to a user or noise. The proposed research points out that addressing problems with noise web data in line with the user needs and interests improves web usage mining process. Furthermore, the process of NWDL is aimed at reducing noise levels in a user profile but also to minimise the loss of useful information otherwise considered as noise. To validate the results presented in this chapter, an evaluation of the proposed NWDL against current and relevant tools is presented in the next chapter.

5.4. Chapter Summary

This chapter presents a number of experiments that are conducted for the purposes of demonstrating how the proposed algorithms contribute to the defined research problem. The experimental results show a significant impact on finding useful information from the web when user interests are considered. One of the key aspects the proposed research consider critical is the dynamic change of user interests, the interest of a user on visited web pages is bound to change, but the process of managing such changes is challenging. The proposed algorithms through experiments conducted justify the need to consider identification and elimination of noise web data as a user-centric approach because the importance of web data is better defined when user interests are considered.

Chapter 6: Evaluating Performance of Proposed NWDL

Chapter five introduces a number of experiments to examine the performance of the proposed algorithm on the test data provided. However, it is equally important to find out how the proposed noise web data learning approach perform against existing tools applied in noise web data reduction process as critically evaluated in the literature. This chapter demonstrates the overall performance of noise web data learning approach with regard to the identification of noise web data as compared to existing tools. The evaluation process in this thesis involves testing the proposed algorithms using test data, this is to simulate the performance of proposed algorithms using a dataset that is not part of the training data. The aim of this process is to demonstrate that the proposed noise web data learning approach produces better results than existing tools. The outcome aims to respond to the research objectives defined in chapter one of the thesis, **how can learning noise web data better address problems with the noisy web in comparison to contributions made by the existing research?**

6.1. Introduction

Evaluating the performance of a machine learning algorithm is a fundamental aspect in examining how efficient the tool performance in addressing a defined problem (Amancio et al., 2011). Moreover, the evaluation process aid in understanding the how various measures incorporated within the proposed measure contribute to the problem domain, as well as refining parameters in the iterative process of learning and selecting most appropriate tool over the available options. Evaluation measures discussed in the previous chapter plays a critical role in comparing the performance used to learn a given dataset (Hossin M and Sulaiman M.N, 2015). Therefore, a selection of suitable evaluation metrics in relation to the defined problem is key to evaluating the performance of the proposed research work.

To ascertain if learning noise web data could improve web usage mining process without affecting the quality of user interest information, the following directions are considered during the evaluation process.

1. The first experimental consideration is an assumption that well-known analytics tools use machine learning tools to learn user interests in order to ensure only useful information is suggested to users. This direction is referred to as a black-box approach
2. Baseline model: First, a number of existing tools are trained using the same dataset. The objective is to find the best performance tool under the same conditions, which will be used to compare the performance of the proposed approach.
3. Finally, the performance of the proposed approach against the baseline model will be evaluated using a dataset which noise data has randomly been injected into the dataset. The objective is to find out how the two models will learn the dataset given the same constraints and if evaluate the output based on classification accuracy. In this scenario, precision, recall and F-measure are used.

The rest of this chapter is organised as follows; section 6.2 examines relevant metrics applied in evaluating the performance of a machine learning algorithm. The metrics selected and applied in this thesis are considered based on the proposed research problem defined by this thesis. Section 6.3. present a validation process based on a black-box approach, i.e., the machine learning tools considered for validation are not specifically defined but could be well known, a specific one, or a modification of existing tools. Section 6.4 presents a direction that evaluates the performance of the proposed NWDL against a baseline model. The baseline selects the best performing machine learning tool among the existing ones when trained by a specific dataset under same constraints. The selected tool is thereafter used as a baseline against which NWDL is evaluated. Section 6.5 evaluates the performance of proposed NWDL against the baseline model using a noise dataset. Finally, section 6.6. summarises the chapter.

6.2. Evaluation Metrics

Selecting a suitable evaluation metric is dependent on the problem to be addressed. Therefore, it is fair to say that it is difficult to state which metric is the most suitable to evaluate a machine learning algorithm without understanding the problem to be addressed by the algorithm. Existing research as evaluated in the literature argues that the usefulness of web data is based on its relationship with the main content of a web page, but the proposed research aims to justify that user interest influence importance of web data. Therefore, if the classification of web data is based on the layout and structure of the web, it can lead to misclassification problems thus increase in noise levels. It, therefore, leads this thesis to address a misclassification problem, thus evaluation metrics considered to measure the performance of proposed algorithms selected based on this concept.

When addressing web data classification problems, evaluation metrics are employed in two critical stages, i.e., training and testing stages (Hossin and Sulaiman, 2015). In training stage, evaluation metric is used to find identify the best tool with high classification accuracy while in the testing stage, the evaluation metric is used to measure the performance of the selected machine learning tool over the other. To understand the principles behind measuring the performance of a machine learning algorithm, the following key concepts should be taken into account:

- Model output: most of the classification models output a probability number for the dataset.
- Objective: create machine learning algorithms that can determine the class a web page is associated with based on its interestingness.
- Output: Web pages that are identified as interest, noise or potential noise.
- Testing: Comparing the output with actual results
- Dataset: Web log files with user access logs.

6.2.1. Confusion Matrix

Confusion matrix is one of the common metrics used in evaluating the performance of a machine learning algorithm in terms of its accuracy and

correctness. In general, is used for classification problems where the output from a machine learning tool can be assigned to one or more classes. A number of existing research consider it important to evaluate the performance of a machine learning tool applied in data classification problem using confusion matrix. This is due to its ability to easily identify when things go wrong when a machine learning algorithm is used in data classification. For instance, a new page with no previous user interest can easily be identified as either interestingness or noise. It is therefore important to consider a neutral class “potential noise” which be used to learn interestingness of a web page prior to assigning as noise or interest. Confusion matrix is widely recognised and applicable in data classification problems for very simple reasons, i.e., based on its ability to determine how well a machine learning algorithm performs. Representing outcome from confusion matrix in a more summarised way is considered efficient when comparing the performance of different algorithms (Ashari et al., 2013). The accuracy of the classifier is given by true positive rate, false positive rate, precision, recall and F-measures using RapidMiner Studio. The average measure from all the classes has been taken to give the overall measure for the classifier. For example, to give the overall precision for a classifier for a given dataset, average of precisions of both true/false classes is calculated.

6.2.2. Accuracy/Error rate

Accuracy is calculated as the number of instances predicted positively divided by the total number of instances. This means accuracy is the percentage of the accurately predicted classes among the total classes. This is the most common evaluation metric used in multi-class classification problems (Hossin and Sulaiman, 2011; Silva-Palacios et al., 2017; Statnikov, et al., 2004). Through accuracy, the performance of a machine learning algorithm is measured based on total corrections, i.e., the total number of web pages whose class is correctly determined by the algorithm when validated using test data.

$$accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{TruePositive} + \text{TrueNegative} + \text{FalsePositive} + \text{FalseNegative}}$$

Identifying and classifying a record to a specific class when it actually belongs to another class leads to classification error. Error rate is the measurement metric for accuracy which evaluates the output by its percentage of incorrect predictions.

$$Error\ rate = \frac{False\ Positive + False\ Negative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

The cost of misclassification error may be higher for one class than the other. For example, misclassification or noise page as interest may affect user's interestingness of a given web page category. The main advantage of error rate is as a result of its applicability to multi-class problems as well as easy to understand by end users (Hand and Till, 2011). On the other hand, evaluating a machine learning algorithm using an accuracy metric has its own limitations. For example, it leads to less discriminating when it comes to identifying and determining the optimal classifier. Moreover, it is considered powerless in terms of informativeness as well as its inclination to minority instances. This will be seen to impact the output in terms of classification accuracy thus impact noise web data reduction process.

6.2.3. Precision, Recall and F-Measure

Precision and recall measures are most common and widely used measures in web usage mining process (Aldekhail, 2016; Duwairi and Ammari, 2016). They are used to assess prediction capabilities of data classification to a predefined class. In the research work, **precision** evaluates the ability of the proposed approach to identify and classify web pages in a user profile based on user interest levels. It is measured by the fraction of extracted data instances i.e. web pages that are of a user interest while *Recall* is the fraction of relevant data instances that are present in the dataset. Therefore, high precision means that there were more interest results than noise instances while high recall means that the tool returned more user interest results.

Precision is the number of correct prediction divided by the number of total predictions made. Intuitively, a high precision for a class means that if our model predicts that class, it is very likely to be true. A high precision model will be useful in those situations where we need to have high confidence in our prediction. Precision can be calculated separately for each class. Graphically, for each row, we take the number on the diagonal, and divide it by the sum of all elements in the column.

Recall: is the number of correct predictions divided by the total number of elements present in that class. Graphically, it is the value on the diagonal, divided by the sum of the values in the row. If recall is high, it means that our model manages to recover most instances of that class. Obtaining high recall is very easy, it is sufficient to say that everything matches that class, and you can be sure that all the elements are retrieved. While it is widely acknowledged that precision and recall for binary classification of straightforward, it is confusing for multiclass classification problems (Hempstalk and Frank, 2008).

F-Measure: F1 score is a binary classification metric that considers both binary metrics precision and recall. It is the harmonic mean between precision and recall. The range is 0 to 1. A larger value indicates better predictive accuracy.

$$F1 = 2 (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) = 2TP / (2TP + FP + FN)$$

6.3. First Direction – Black-box Validation Process

The assumption made using this approach is that various data analytics platforms use machine learning tools to understand user needs and interests while on the web (Siemens, 2013). However, there is no explicit information on the type of machine learning tools used, as in the case of Google Analytics. This research considers a ‘black-box’ validation approach because the machine learning tools applied could be well known, a specific one, or a modification of existing tools. The objective is to find out how user interest level is determined based on visited web pages and the type of noise web data identified.

6.3.1. Evaluating Performance of NWDL using a ‘Black Box’ Approach

The output from the existing tool, i.e., interest and noise data will be used as input to the proposed tool for the validation process. The choice of this black-box approach presents a good platform for validation of the proposed algorithms which means that the performance of the proposed algorithms is compared with any current and practically used algorithms. Summary of the experimental design scheme is presented in **Figure 6.0**.

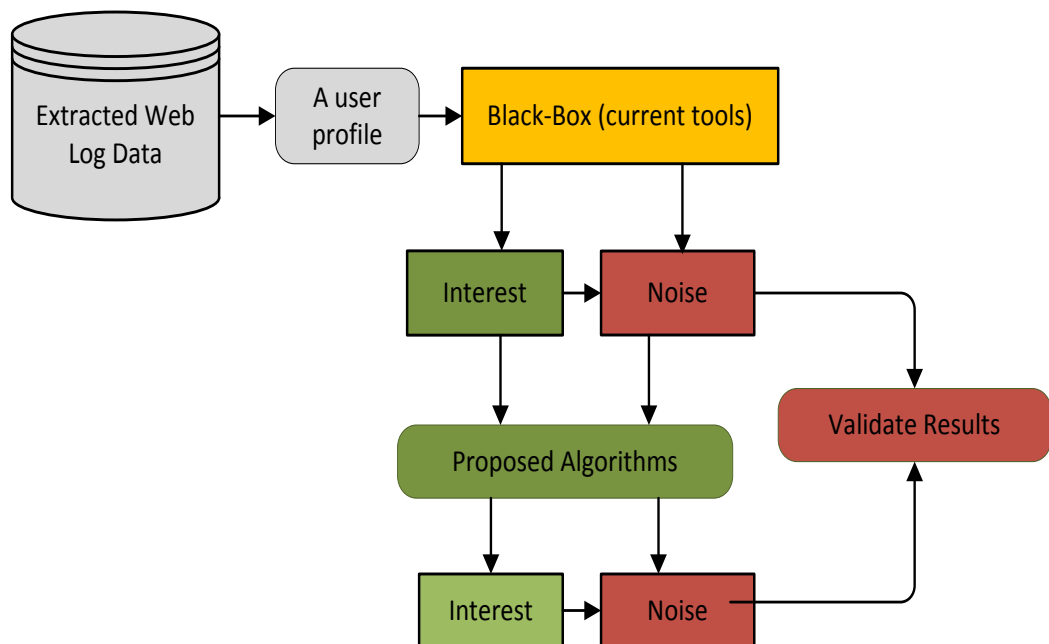


Figure 6.1: Black-Box Validation Process

6.3.2. Discussion of the results

One of the main objectives of the proposed research is to find out how current research identify noise in web data and if user interests are considered prior to noise elimination. In **Figure 6.2**, it can be observed that a web page is classified as noise or interest based on the level of user interest. However, the proposed research work attempts to understand if the user interest level of a web page influences its classification. For example, what is the threshold support value used to define interestingness of a web page, if a dynamic change of user interest influences the weight of a web page and how such changes impact classification of a web page to either noise or interest. Taking into account this critical issues, the proposed research uses the output from **Figure 15** in order to learn interestingness of a web page taking into account aspects such as dynamic change of user interest and threshold support value considered in web page classification.

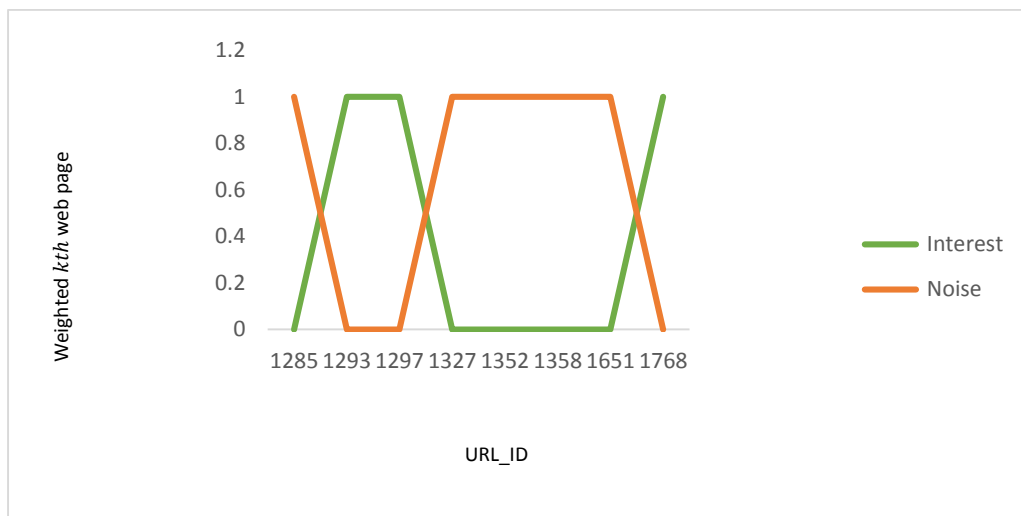


Figure 6.2: The output from black-box approach

Figure 6.2. The output from black-box approach - used as input to the proposed tool

Figure 6.2 shows the output obtained from Google Analytics (GA) account. The assumptions made is that GA uses various machine learning tools to determine the interestingness of a web page present in a web user profile. The output can be classified as either interest or noise data.

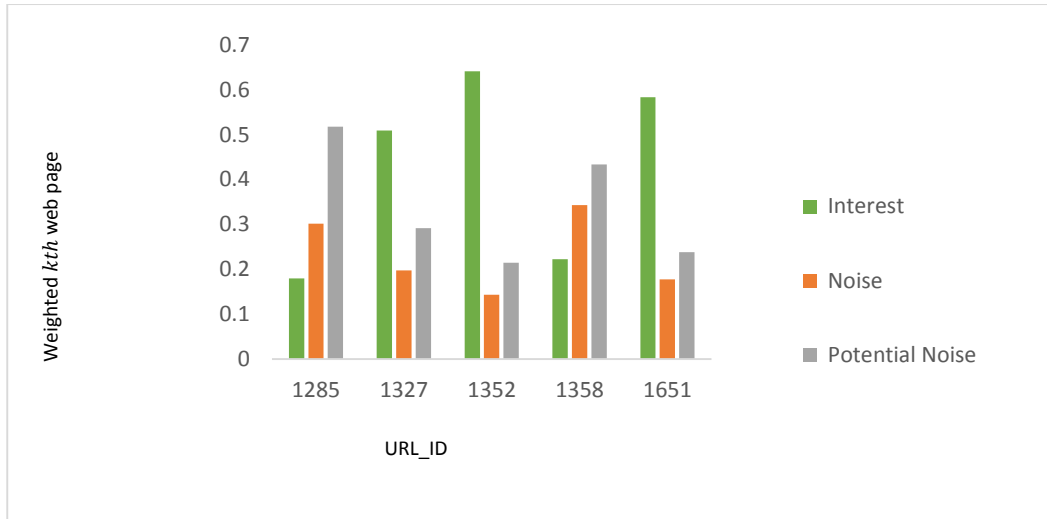


Figure 6.3: The output from the black-box approach

Figure 6.3. The output from the black-box approach- used as input to the proposed tool. It considers the output from **Figure 6.2** where interest and noise output results are used as input to the proposed noise web data learning approach. The objective is to find out what type of noise each process identifies taking into account various measures applied in the proposed algorithms.

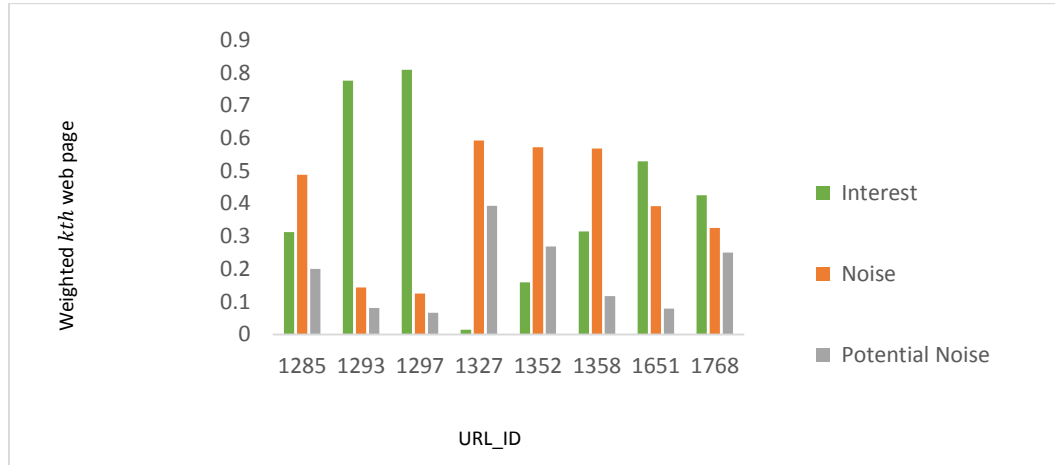


Figure 6.4: Noise output from black-box approach

Figure 6.4. Noise output from black-box approach is used as input to the proposed tool

The proposed NWDL learn noise out from the black box approach prior to classification, the objective is to ensure identification of web pages regarded as noise by current tools consider the interest of a user. In previous chapters of this thesis, the proposed research acknowledges that user interest influence the interestingness of a web page.

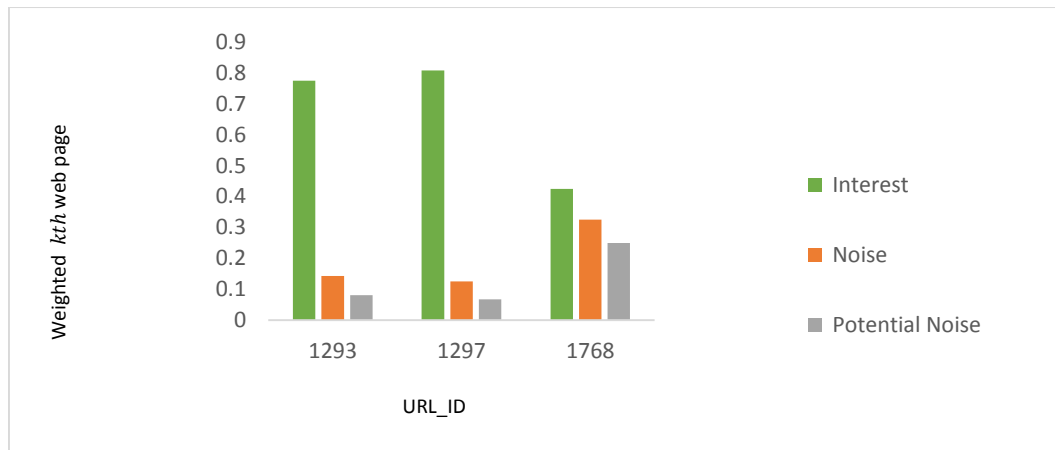


Figure 6.5: Interest output from black-box approach

Figure 6.5. Interest output from black-box approach is used as input to the proposed tool. This is the output from the proposed tool where weighted web page category and depth of interest are considered to determine user interest level.

An additional class this research work refer to as potential noise is considered. Potential noise class defines data instance whose characteristics can neither be classified as interest nor noise. For example, where a user interest is below the defined support threshold or user interest is unknown. Unknown interests are as a result of the web page to a specific web category with no user visits. Weighted web page category and depth of visit plays a significant role in the identification and classification of web data based on dynamic user interests.

In summary, **Figures 6.3-6.5** demonstrate that the proposed algorithms learn user interest levels in both noise and interest outputs. Unlike the results presented in **Figure 6.2**, levels of user interest on extracted web log data are determined. Therefore, the classification of web data cannot only rely on a standard threshold to determine a class, but the use of dynamic thresholds improves web data classification performance. Moreover, evaluating the performance of the proposed algorithms with existing tools is critical to achieving the outline research objectives. The proposed research considers factors that should be taken into consideration, for example, if the proposed tool can identify interestingness of web pages in a user profile in relation to dynamic change of user interests.

6.4. Second Direction- Baseline Model

It is difficult to choose a machine learning algorithm that performs well without comparing the performance of others (Doan and Kalita, 2015; Khosla et al., 2010; Lalor et al., 2017). In order to ensure the best performing tool is selected for the defined problem, a baseline model is considered. A baseline provides a point of reference from which to compare other machine learning algorithms is considered. Further, it defines a benchmark that all other machine learning algorithms must cross to demonstrate their contribution to the defined problem (Saad, 2014). Without a baseline, it is difficult to know how well the proposed algorithms perform in addressing the defined problem. Alligier et al (2015), Taylor and Fenner (2017) argue that it is important to try a number of different algorithms and determine what performs best on a specific problem. To evaluate the performance of the proposed NWDL approach, this thesis considers a baseline model which is based on a number of widely used existing tools. The selected tools are used to create predictions for the defined dataset, the output is thereafter used to measure the baseline's performance taking into account various evaluation metrics. The aim of using baseline is to determine the best performing tool among the existing ones when trained by a specific dataset under the same constraints. The following steps are considered when building a baseline model:

Step 1: Load the dataset

Step 2: Split data into training and test set

Step 3: Select a number of tools to compare

Step 4: Compare the models

Step 5: Evaluate performance with metrics that takes the model and testing data

Step 6: Select the best performing model to compare the performance of the proposed NWDL approach

The proposed research has considered an open source machine learning and data mining toolkit to define a baseline model. Orange is a machine learning tool-kit which house well-known algorithms applied in data mining (Demšar et al., 2004; Podpecan et al., 2012). In this thesis, it has been used mainly for validation purposes but the proposed algorithms can be validated using any

machine learning platform. **Figure 6.6** shows the process built to determine how existing tools perform in determining the interestingness of a web page prior to noise elimination. The best performing algorithm will then become the baseline during the evaluation of the proposed tool.

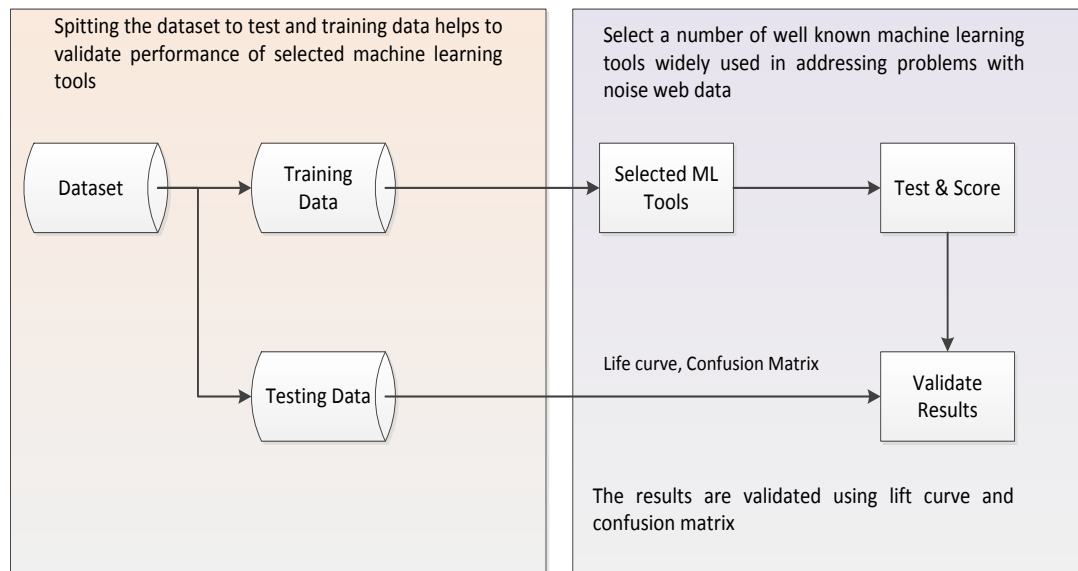


Figure 6.6: Defining a Baseline Model

The performance of machine learning tools selected in **Figure 6.6** is measured using the lift curve. It measures the performance of the selected machine learning tools in terms of classification accuracy. Classification accuracy is based on how correctly a tool will determine if a web page is interesting or noise based on user interest level. The web page is subsequently assigned to a predefined class, i.e., interest or noise class.

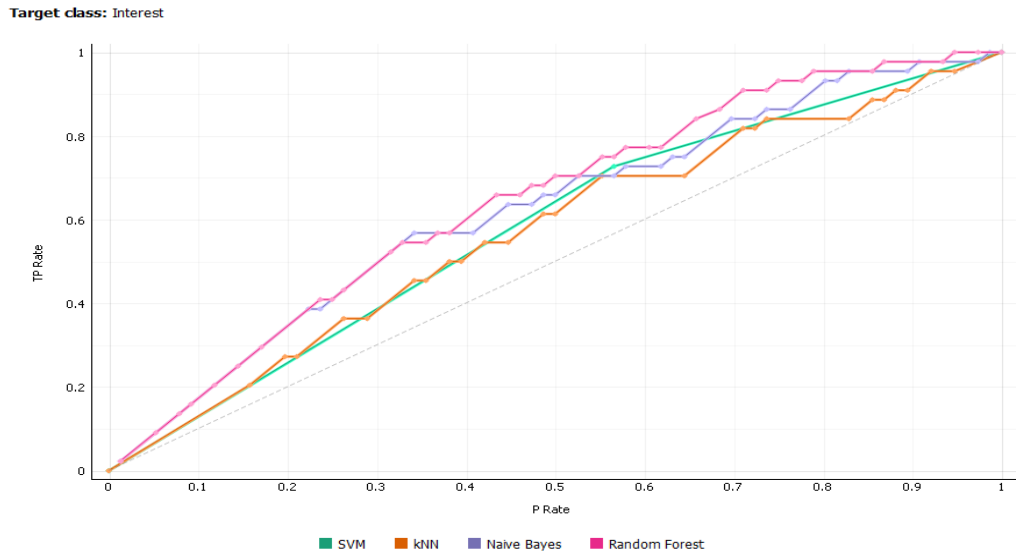


Figure 6.7a. Lift curve of the selected machine learning tools used in the classification of web pages in a user profile. The defined target class here is Interest class. It is observed that random forest has the highest classification accuracy.

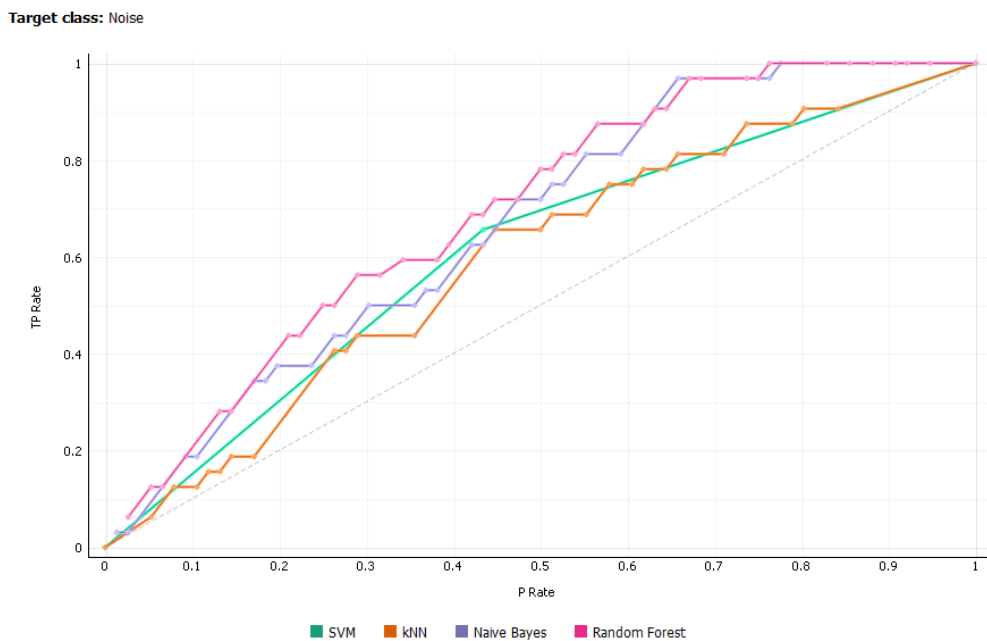


Figure 6.7b. The lift curve shows how the selected machine learning tools perform on the specified dataset in terms in terms of classification accuracy. The output results show that random forest has the highest classification accuracy. The target class is Noise class.

Based on output results shown in Figures 6.7a and 6.7b, it is observed that random forest performs better than rest. Further results based on the performance of the selected tools are presented in **Appendix V**. Therefore, the performance random forest is set as a baseline model to which NWDL is evaluated against.

6.4.1. Evaluating the performance of NWDL against the Baseline Model

The experimental results from the baseline model process form the basis for evaluating the proposed NWDL. The key aspects of this validation process aim to demonstrate include (1) using outcome from the evaluation process, the impact of the dynamic change of user interests towards defining interestingness of web data. (2) Influence of proposed noise web data learning approach in the classification of web pages prior to noise identification and elimination. These aspects are considered critical to the research contributions outlined in the thesis. Subsequently addressing noise in web data with a key focus on the user's change of interest over time justifies the need to learn noise web data prior to elimination. To demonstrate the performance of proposed NWDL over the existing tool (baseline model), a confusion matrix is used to demonstrate the classification accuracy of web pages is a user profile. In this section, we analyse 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. The following are the steps

1. Load data
2. Train the classifier with labelled data.
3. The training phase generates a model as output which will be used in the validation.
4. During validation use test data to evaluate the performance of the model.
5. After the evaluation of the classifier, the results will be displayed in a confusion matrix.

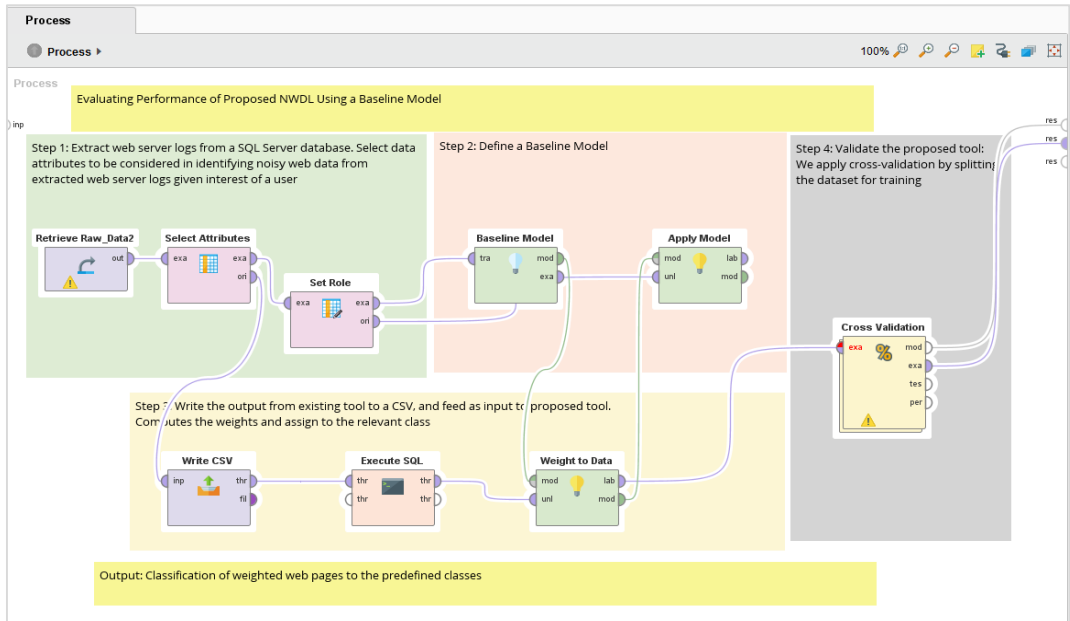


Figure 6.8: Baseline vs NWDL – A validation process.

The performance of the baseline model is measured using confusion matrix as shown in figure 6.9. The results are presented in form of a confusion matrix which shows the actual vs the predicted data instances for the defined classes. The figure shows the number of correct and incorrectly classified web pages in each predefined classes.

Table View Plot View

accuracy: 50.77%

	true Interest	true Potential Noise	true Noise	class precision
pred. Interest	30	18	9	52.63%
pred. Potential Noise	2	2	1	40.00%
pred. Noise	2	0	1	33.33%
class recall	88.24%	10.00%	9.09%	

Figure 6.9: Performance evaluation using Confusion Matrix

6.4.2. Discussion of the results

The overall performance of the proposed NWDL is presented in the previous chapter, Figure 5.6. This section now compares the performance of the

proposed NWDL against a *baseline* model whose results are shown in **Figure 6.9**. The confusion matrix shows different results when comparing the performance of the proposed algorithm against the baseline model. In **Table 7**, *Process 1* shows results from the baseline model while *process 2* shows results obtained from the proposed NWDL algorithm. Results from figure 5.6 show an increase in potential noise but a decrease in noise given the fact that web pages with no known user interest but whose category are known have been considered as potential noise. The accuracy is the sum of all the numbers on the diagonal divided by the sum of all numbers. The larger the number on the diagonal line, the better the classification performance. In this experiment, the results obtained are defined as the set of predefined classes with weighted web pages.

Table 7: Performance Comparison NWDL vs Baseline Model

Model	CA
Baseline Model	50.77%
NWDL	60%

This thesis notes that the proposed tool performs better than the baseline model with a high number of correctly predicted web pages that belong to the positive class. Moreover, the proposed tool is able to identify web pages that potentially interest which existing tool eliminate as noise. The proposed research considers a scenario where a standard threshold value is set to identify web pages in a user profile that are useful or noise. There is a high likelihood that a given a predefined class will have a high number of noise or interest based on the support threshold determined. This will result in high classification accuracy. For this reason, this experiment considers web pages with varying user interest levels. Using a dynamic threshold value, the results will be compared to actual test data. Generally, the evaluation measures used to identify noise web pages are defined from a matrix with a number of web pages correctly and incorrectly classified based on predefined classes. For example, this research considers confusion matrix to evaluate the performance of proposed noise web data learning process.

In summary, a key focus is to evaluate key measures that have been proposed by this research. For example, how dynamic threshold values influence the classification of web pages in a user profile and subsequent identification and reduction of noise. The results obtained suggest that the proposed tool perform better than existing tools. To address problems with noise web data as a result of misclassification, the performance of the proposed NWDL is measured in terms of error rate, this is the % of incorrectly classified web pages in a user profile. However, the proposed NWDL aim to learn noise web data prior to classification hence the consideration of a potential noise class to classify new web pages. The focus here is to measure the performance of the proposed NWDL on web pages regarded as noise by existing tools

6.5. Evaluating the performance of the proposed NWDL approach using a noise dataset

The objective of this experimental direction is to determine how existing tools perform in a noisy dataset. For example, what type of web data they identify as noise when compared to the proposed NWDL. In order to evaluate the proposed tool under different noise levels, noise data is introduced to the training dataset which is randomly generated. For every test data, the performance of the proposed tool is evaluated against currently available tools (Baseline Model).

6.5.1. Validation Process

In order to create a noisy dataset from the original one, the proposed research consider a number of aspects, such as, the type of noise which can either be web page categories which have not recorded any user visit before, the number of folds of the cross-validation used to validate the classifier as in this case. The general procedure adopted in this process is as follows:

1. A level of noise of either class noise or attribute noise is introduced into a copy of the full original dataset.
2. Both datasets, the original one and the noisy copy, are partitioned into K equivalent folds, that is, with the same example in each one.

3. The training partitions are usually built from the noise data, whereas the test partitions are from the actual dataset in the case interest class.

6.5.2. Discussion of the Results

From the results shown in **Figure 6.10**, the proposed tool does not consider web page category with no previous visit as noise, instead, they are considered potential noise, and the rationale is to give room to build user interest over a period of time. The confusion matrix shows different results when comparing the performance of the proposed algorithm against the Baseline model. *Process 1* results from the baseline model while process 2 shows results obtained from the proposed NWDL. Results from process 2 show an increase in potential noise but a decrease in noise given the fact that web pages with no known user interest but whose category are known have been considered as potential noise. The accuracy is the sum of all the numbers on the diagonal divided by the sum of all numbers. The larger the number on the diagonal line, the better the classification performance. In the previous chapters 4 and 5, the proposed research presents recency adjustment measure to learn interestingness of a web page. As discussed in chapter 4, user interest decline in half by a week based on half-life function. Therefore, using a potential class to learn noise web data justifies the fact the noise web data is defined by the level of a user interest over a period of time.


```

Process: 1
Performance:
PerformanceVector [
****accuracy: 53.08% +/- 4.37% (micro average: 53.08%)
ConfusionMatrix:
True: Potential Noise Interest Noise
Potential Noise: 8 1 5
Interest: 69 119 39
Noise: 3 5 11
----classification error: 46.92% +/- 4.37% (micro average: 46.92%)
ConfusionMatrix:
True: Potential Noise Interest Noise
Potential Noise: 8 1 5
Interest: 69 119 39
Noise: 3 5 11

Process: 2
Performance:
PerformanceVector [
****accuracy: 63.56% +/- 2.81% (micro average: 63.55%)
ConfusionMatrix:
True: Potential Noise Interest Noise
Potential Noise: 20 4 9
Interest: 35 99 18
Noise: 5 3 10
----classification error: 36.44% +/- 2.81% (micro average: 36.45%)
ConfusionMatrix:
True: Potential Noise Interest Noise
Potential Noise: 20 4 9
Interest: 35 99 18
Noise: 5 3 10

```

Figure 6.10: Performance evaluation in noise data set

6.6. Evaluating the performance of the proposed NWDL using Open Source Dataset

6.6.1. Data Description

The second dataset is an extract of web visitor interests available at Kaggle dataset store <https://www.kaggle.com/uciml/identifying-interesting-web-pages>. Dataset 2 was extracted to generate a training and testing set. The training set is used to learn the proposed algorithms, which is then used to generate input to the proposed to the proposed noise web data learning approach.

6.6.2. Validation Process

The performance of the proposed NWDL approach is validated using the process flow presented in Figure 6.8. Unlike the results presented in section 6.4, this section considers dataset 2. The results from this validation process aim to justify that the proposed NWDL is adaptable to different types of data and the size of data does not impact its classification performance. The output results are distinguished as process 1 which presents results from the

baseline model while results from the proposed NWDL are presented in process 2.

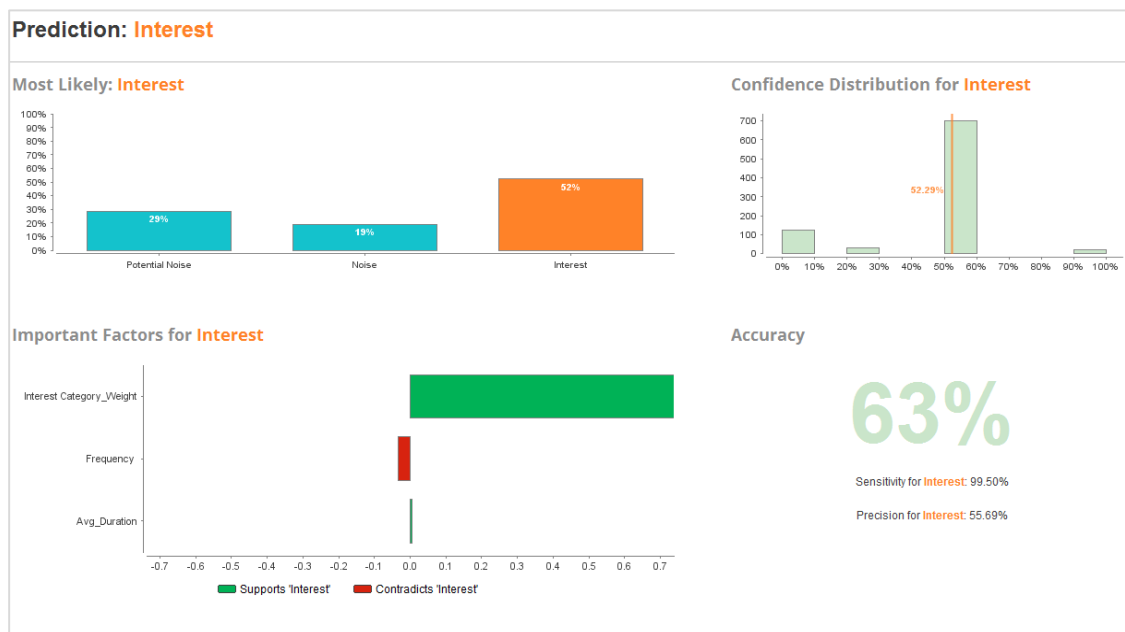


Figure 6.11: Classification Performance of the Baseline Model

Process 1: Baseline Model

accuracy: 63.10% +/- 3.34% (micro average: 63.10%)

ConfusionMatrix:

True:	Potential Noise	Interest	Noise
Potential Noise:	57	2	1
Interest:	142	289	86
Noise:	0	0	49

classification_error: 36.90% +/- 3.34% (micro average: 36.90%)

ConfusionMatrix:

True:	Potential Noise	Interest	Noise
Potential Noise:	57	2	1
Interest:	142	289	86
Noise:	0	0	49

Figure 6.11 presents the results obtained from a baseline model using dataset 2. Performance of the baseline model is evaluated using confusion matrix. The output from confusion matrix in a more summarised as shown above. The accuracy of the classifier is given by true positive rate, false positive rate, precision, recall and F-measures. The average of measure from all the classes has been taken to give the overall measure for the classifier which is 63.1%.

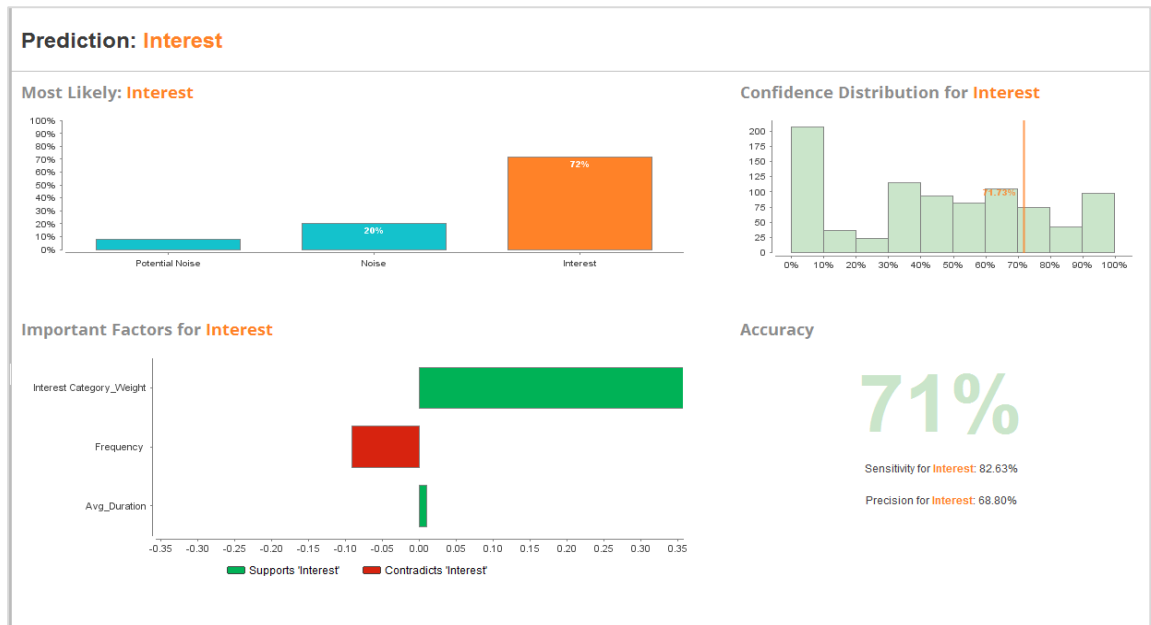


Figure 6.12: Classification Performance of the Proposed NWDL

Process 2: NWDL

accuracy: 71.88% +/- 2.86% (micro average: 71.88%)

Confusion Matrix:

True:	Potential Noise	Interest	Noise
Potential Noise:	134	29	16
Interest:	56	238	46
Noise:	4	25	78

classification_error: 28.12% +/- 2.86% (micro average: 28.12%)

Confusion Matrix:

True:	Potential Noise	Interest	Noise
Potential Noise:	134	29	16
Interest:	56	238	46
Noise:	4	25	78

6.6.3 Discussion of the Results

After training the classifier, the test data was fed to the selected classifier. The best performing classifier was determined based on classification accuracy. Figure 6.12 presents results obtained from NWDL using dataset 2, it is observed that the proposed approach obtained classification accuracy of 71.88%. The results presented in this section provide a detailed picture of why it is critical to determine the interestingness of web data from a user interest perspective. The test results demonstrate that even though any existing machine learning tool currently available is capable of addressing problems with data classification, noise web data learning approach appears to greatly contribute towards loss of useful information and reduction of noise data. The results presented in Figure 6.12 provides a benchmark of key measures

considered in the NWDL approach and their contributions to the needs and interests of web users. For example, using open source data, the proposed NWDL approach demonstrate the influence of dynamic user interests to the interestingness of web data. Further, it helps to identify latent information available on the web and derives a user profile that reflects ever-changing user interests

6.7. Discussion of critical aspects based on the performance of NWDL

As discussed in chapter 3, the user session identification process plays a critical role in learning the interestingness of web data. Even a number of existing approach have applied a dynamic time-out session identification approach, there is lack of sufficient evidence to demonstrate that the last page visit in a session is included. Excluding the last page visit from a user, session affect the weighting of a webpage thus a misclassification, either as noise or interestingness. In chapter 4, interestingness of a web page visited by a user does not only depend on frequency and duration of a user visit, instead, key measures such as change of interest over time based on the associated web page category are critical. Some existing research, for example, Ma et al., tend to overlook varying user interest and instead consider the total amount of time and number of visits to captured from a specific user. The proposed approach is able to learn user interest as they change over time thus ensuring useful information is available to a user only when it is interesting.

Based on the dataset considered in conducting experiments, it is evident that existing tools do not take into account the dynamic change of user interest during the noise web data reduction process. Learning of noise prior to elimination reduces the amount of useful information eliminated thus enriching a user profile. The results from these experiments contribute to the existing research work by understanding how user interest can influence the type of web data considered noise or useful. The proposed noise web data learning tool incorporates user interests and evolving web data, which means that the usefulness of data available on the web is determined by what the user is interested in and not the relationship with the website where the data resides.

In addition, this experimental direction offers a means of demonstrating a user-centric approach with respect to noise web data reduction process influenced by changes to user interests. It enables the process of learning noise in web data to consider the effects of evolving web data, and if certain aspects of time and location of user visit to a web page influence the interestingness of available web data. While presenting this direction as one way of enriching user profile through reduction of noise as well as decreasing the amount of useful information lost, it is also important to point out that the interestingness of web data does not only depend on its popularity, frequency and duration of visit but it's relevancy at the time and place of access.

6.8. Chapter Summary

This section evaluates how dynamic change of user interest impact the interestingness of web data. The experimental results are compared against the output from existing techniques using the same dataset. The experiments conducted in this section aim to justify the need to learn the interestingness of web data prior to noise elimination. The results from the proposed approach are compared against existing tools. Based on experiments conducted using existing tools, it is not clear how change or user interests are captured during the classification of web pages in line with user interest level. The proposed approach use weighted web page category based on a forgetting function approach to determine how change of user interest impact noise web data. The following are the algorithms with based on this approach, they include; duration of a user visit based on dynamic time-out adjustment, depth of a user visit to a web page and weighted web page category

Various measures that are used to analyse the degree of interestingness of the web page prior to the classification process are evaluated. More particularly using noise dataset, the aim was to compare the performance of the classifier learned with the original data set with the performance of the classifier based on a noise dataset. In addition, it has been observed from an experiment conducted in **Figure 5.6** that the widely used measures as such frequency and duration of a user visit have a very limited impact when it comes to learning change of user interests over time.

Chapter 7: Conclusion and Future Work

7.1. Introduction

This chapter presents a summary of the research this thesis proposes. The objective is to 'close the loop' i.e. it reflects on how the defined research objective has been achieved, evaluate with justification as to why the proposed noise web data learning approach is significant to address problems with noise web data. This chapter commences by evaluating research objectives against the proposed research outcomes.

7.2. Critical Discussions Based on Research Objectives

The proposed research commences by defining a problem in existing research that needs to be addressed. The defined problem is based on the contributions and limitations of existing research work that address problems with noise web data. Moreover, the proposed research is motivated by the increased usage of web data as well as the volume of data which is rapidly growing. As a result, problems with finding useful information that meets the interests of a user prove a challenge to the use of existing machine learning tools.

From the existing literature, this thesis establishes how relevant and most current research studies address problems associated with noise in web data. The proposed research argues that not all data that form part of the main web page is relevant to a user's interest and not every data which can be considered noise is actually noise to a given user. Therefore, without learning the interestingness of web data based on a user's interest, the process of eliminating noise web data is limited to simply recognising how web data is presented and not what users are interested in at any given time.

In summary, the proposed research's main focus is to learn to recognise noise in web data so as to reduce the loss of useful information otherwise considered as noise as well as to decrease noise levels. Rather than isolating the main web page content and relying on its layout and content, the proposed research aims to focus on how user interest can influence the type of noise present in web logs.

The key stages involved in defining a user profile are identified and explored. They include; user and session identification stages of pre-processing web log data play an important role in finding data on the web data defines a user and his/her interests. A session identification algorithm based on the dynamic threshold value is considered is used to determine user sessions. The rationale for using dynamic threshold value over fixed values in session identification is to ensure the interestingness of a web page reflects the level of user interests over time. This is due to the fact that user interest change over time, a fixed threshold value will not be ideal enough to ensure the classification process is dynamic to accommodate varying user interests. Subsequently, the process of identifying and elimination noise web data as user interest change over time is well managed.

Learning user interests plays a fundamental role in understanding how useful is web data to a user at a given time, thereby improving the process of noise web data reduction. It is therefore important to understand how the dynamic nature of the web and varying user interests influence the identification of noise in web data. In the proposed research work, the focus is on learning the interest of a user in relation to available web data with the aim of reducing the amount of useful information eliminated as noise.

Chapters 5-6 have introduced a number of experimental directions and discussions of results based on the application of proposed algorithms. From the results, it is observed that the proposed algorithms outperform existing tools applied in noise web data reduction process. This was measured using a number of criteria introduced and defined in previous chapters. This chapter evaluates the success of the proposed research work holistically against the research objectives as defined in chapter 1.

7.3. Key Findings Based on the Proposed Research Questions

This section revisits the research questions outlined in chapter one of this thesis. The main objective of this thesis is to identify current problems with noise in web data and proposed an approach capable of addressing defined problems in relation to web user interests. Therefore, there is a need to find

out if this thesis has managed to respond to the research questions outlined in the first chapter of this thesis.

Question 1: In what ways current research works define and address noise in web data?

At the onset, the criteria set to review and evaluate existing literature covered the following aspects; definition of noise in web data, tools and techniques proposed to identify and eliminate noise data, measures employed by existing research to evaluate performance of existing tools, contribution and limitations taking into account the defined problems and the current state of the art.

This thesis found out that although there are a number of tools and techniques that identify and eliminate noise web data, there are still unsolved issues critical to addressing problems with noise web data. For example, there are no tools currently applied to learning noise web data prior to elimination. There are no discussions on how existing tools used to eliminate noise in web data take into account evolving user interests. The current research work has not explicitly defined measures that will aid in understanding the influence of user interests and how a change to user interests are modelled to minimise loss of useful information.

Question 2: What are the key indicators for learning user interests and how could the interests of a user influence identification of noise web data?

Findings: Where the interests of a user change, noise in web data also change, further interestingness of available information vary as well if some interests are seasonal to a user. In essence, it is important to learn dynamic changes of a user subject to time of interest. This, therefore, demonstrates that it is difficult to rely on existing noise web data patterns which are determined based on previous user activities. In order to identify and understand how dynamic changes in user interests impact identification of noise in web data, the depth of user visit, as well as the use of dynamic threshold value, plays a critical role. Understanding the interest of users on the web and how users navigate through a web page will provide an insight into the type of users and their level of knowledge. This is because the level

of knowledge demonstrates if a user is interested in available information or he/she is actually struggling to find useful information.

The proposed research considers interest forgetting function to learn change in user interests which ultimately improves noise web data reduction process. Some of the key aspects the proposed research explores in relation to this approach include (1) use of a fixed time to determine if a web page is interest or noise to a user. (2) Learning interestingness of a web page gradually over time to determine the level at which a user is showing interest or losing interest on a given page. For instance, a user may have recently visited a specific category of a web page, but the frequency and time spent on the category are gradually decreasing. Even though website owners/developers will continue suggesting relevant information, the time will come when a user will no longer be interested, hence such information becomes noise. Chapter 4 underlines the need for learning interestingness of a web page gradually using forgetting function. In chapter 5 the approach is presented in form of an experiment in order to justify its contribution to the proposed NWDL.

Question 3: How can learning noise web data better address problems with the noisy web in comparison to contributions made by the existing research?

The process of learning noise web data prior to noise elimination is discussed entirely in this thesis. However, chapter 4 underlines a number of measures which are considered key to justifying the contribution of NWDL. Subsequently, chapter 5 and 6 present a number of experiments to justify how NWDL can address problems with noise web data better when compared to existing tools. In summary, the proposed NWDL consider dynamic changes of user interest while learning interestingness of a web page. The aim is to ensure that during the classification of web pages visited by a user, the level of user interest as well as the change of interest over time is considered to avoid misclassification. From the experiments conducted using real-life data, it is evident that existing tools do not take into account the dynamic change of user interest during the noise web data reduction process.

Learning of noise prior to elimination reduces the amount of useful information eliminated thus enriching a user profile. The results from these experiments contribute to the existing research work by understanding how user interest can influence the type of web data considered noise or useful. The proposed noise web data learning tool incorporates user interests and evolving web data, which means that the usefulness of data available on the web is determined by what the user is interested in and not the relationship with the website where the data resides. In addition, this experimental direction offers a means of demonstrating a user-centric approach with respect to noise web data reduction process influenced by changes to user interests. It enables the process of learning noise in web data to consider the effects of evolving web data, and if certain aspects of time and location of user visit to a web page influence the interestingness of available web data. While presenting this direction as one way of enriching user profile through reduction of noise as well as decreasing the amount of useful information lost, it is also important to point out that the interestingness of web data does not only depend on its popularity, frequency and duration of visit but it's relevancy at the time and place of access.

7.4. Research Contribution

In this thesis, a number of machine learning algorithms have been proposed to address problems with noise in web data. They include; session identification based on dynamic time-out value, determining user interest level on a web page based on depth and category of user visit to a web page, Web page classification based on the dynamic threshold value and noise web data learning algorithm

Determining user session based on dynamic threshold values overcome the limitations of web page classification based on the duration and frequency of user visit. As discussed in chapter 3, the majority of existing research consider user session identification based on a fixed time-out value. One of the main limitations of this approach is that some web pages a user visit are excluded from a session. For example, the time a user exit from a web page cannot be determined and therefore unable to determine if the last page is

interestingness or noise to a user. In most cases, web pages with no user interest or where a user spends less time are considered irrelevant or noise and thereby eliminated. With a dynamic session time-out approach, the level of noisiness and loss of useful information is managed effectively, simply because the time spent on the last page is determined rather than being excluded from a user session.

Dynamic change of user interests plays a fundamental role in learning interestingness of a web page. Based on a number of experiments conducted, the proposed research justifies the need to consider change of user interest over time prior to determining if information on the web is useful or noise. As user interests change over time, some information becomes noise as new interests emerge, subsequently, previous noise data can be useful as well. Therefore, learning user interests based on the dynamic change of user interests as proposed in this research makes some contributions to the research domain: (1) Improves web usage mining process by acknowledging and accommodating the dynamic change of user interests. (2) Minimises loss of useful information otherwise considered as noise when interests of a user are overlooked. (3) Given that the web is regarded as noisy, the NWDL approach contributes to the reduction of noise web data present in a user profile. This is achieved by taking into account the aforementioned measures, i.e., learning change of user interest prior to noise identification and elimination. Overall, the practical application of the proposed NWDL contributes towards creating web user profiles that are dynamic enough to adapt to the evolving web, thus ensuring only useful information is available to a user when required.

7.5. Future Work

This proposed research work in this thesis underlines and presents a number of machine learning models in an attempt to address problems with noise web data. The proposed approach shows the influence change of user interests has on the interestingness of web data. It further points out the need to learn noise web data with a close consideration of user's change of interests. Despite the results achieved from various experiments conducted as well as

the contribution made, this section highlights possible future research directions:

Learning noisiness without prior knowledge of user interests: Even though a number of measures have been proposed in this thesis to learn noise web data, there are emerging problems as a result of rapid increase of web data. For example, predicting user interests without any previous information to build an initial profile. In order to address noise web data and minimise information loss, there is a need to extend the proposed research taking into account existing geo-spatial data to benefit web usage mining process. More importantly, future work should accommodate machine learning tools that are dependent on the user and geo-spatial data. The logical connection between the two research domains can play a significant role in building dynamic websites that give web user more power to control what they view on the web. This includes dynamic web content that is user-driven as well as machine learning models that utilise noise data to learn dynamic change of user interests.

Fighting Fake News: Currently, everyone has joined the battle of combating fake news. Fake news is misleading information which can be used to deceive web users (Shu et al., 2017). For example, websites generate a significant portion of their revenue through clicks on contents regardless of its legitimacy, i.e., fake news or useful information. As defined in chapter 1 of this thesis, noise data is irrelevant or meaningless data. Based on these definitions and from web data pre-processing perspective, there is a correlation between fake news and noise data and the associated current research challenges. The main challenges the current research attempts to address are ways in which we can employ machine learning to fight fake news, minimise the amount of misleading information suggested to users who are searching for useful information on the web. Addressing problems with fake news is not only data-driven but also involves the user. This research proposes a noise web data learning approaches that identifies and learn noise data on the web. However, there is a need to extend this research using web user profiling approach presented in the thesis to identify features most commonly associated with fake news. It is important to focus on learning the source of

fake news prior to addressing its impact to the society. In order to achieve this, a further in-depth analysis of learning from social media networks what qualifies as fake news in relation to user's perception and interests over a time.

References

- Adeniyi, D.A., Wei, Z., and Yongquan, Y. (2016), "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Inform.* 12, 90–108
- Ahmed, A., Low, Y., Aly, M., Josifovski, V., and Smola, A.J. (2011), "Scalable Distributed Inference of Dynamic User Interests for Behavioral Targeting," *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY, USA: ACM), pp. 114–122.
- Ajabshir, Z.F., (2014), "The Effect of Implicit and Explicit Types of Feedback on Learners' Pragmatic Development," *International Conference on Current Trends in ELT*, Vol.98 pp.463-471
- Akpınar, M.E., and Yesilada, Y. (2013), "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 – Vol. 8295*, (New York, NY, USA: Springer-Verlag New York, Inc.), pp. 238–252.
- Akuma, S., Iqbal, R., Jayne, C., and Doctor, F. (2016), "Comparative analysis of relevance feedback methods based on two user studies," *Computers in Human Behavior*, Vol 60, pp.138–146.
- Aldekhail, M. (2016), "Application and Significance of Web Usage Mining in the 21st Century: A Literature Review," *Int. J. Comput. Theory Eng.* Vol. 8, 41–47.
- Aljuaid, T. and S. Sasi, S. (2016), "Intelligent imputation technique for missing values," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, 2016, pp. 2441-2445
- Alligier, R., Gianazza, D., and Durand, N. (2015). Machine Learning and Mass Estimation Methods for Ground-Based Aircraft Climb Prediction. *IEEE Trans. Intell. Transp. Syst.* Vol.16, 3138–3149.
- Alphy, A., and Prabakaran, S. (2015), "A Dynamic Recommender System for Improved Web Usage Mining and CRM Using Swarm Intelligence," *The Scientific World Journal, Volume 2015*, Article ID 193631
- Amancio, D.R., Nunes, M.G.V., Oliveira, O.N., Pardo, T.A.S., Antigueira, L., and da F. Costa, L. (2011), "Using metrics from complex networks to evaluate machine translation," *Phys. Stat. Mech. Its Appl.* 390, 131–142.
- Amato, G., and Straccia, U. (1999), "User Profile Modeling and Applications to Digital Libraries," *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, (Springer-Verlag), pp. 184–197.
- Ansari, Z., Azeem, M.F., Babu, A.V., and Ahmed, W. (2015), "A fuzzy clustering based approach for mining usage profiles from web log data," *International Journal of Computer Science and Information Security*, Vol. 9, No. 6, pp. 70-79.
- Aye, T.T. (2011). Web log cleaning for mining of web usage patterns. *In 2011 3rd International Conference on Computer Research and Development*, pp. 490–494.

Azad, H.K., Raj, R., Kumar, R., Ranjan, H., Abhishek, K., and Singh, M.P. (2014), "Removal of Noisy Information in Web Pages," *ICTCS '14 Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*, Article No. 88 (ACM Press), pp. 1–5.

Azimpour-Kivi, M., and Azmi, R. (2011), "A webpage similarity measure for web sessions clustering using sequence alignment," *In 2011 International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp. 20–24.

Bamshad Mobasher and Olfa Nasraoui, (2011) "CHAPTER 12: Web Usage Mining", invited book chapter in "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)," Second Edition

Bhamare, S.S., and Pawar, B.V. (2013), "Survey on Web Page Noise Cleaning for Web Mining," *Int. J. Comput. Sci. Inf. Technol.* Vol.4, pp.766–770.

Bhargava, P., Brdiczka, O., Roberts, M. (2015), "Unsupervised Modeling of Users' Interests from their Facebook Profiles and Activities," *Proceedings of the 20th International Conference on Intelligent User Interfaces*, Pages 191-201

Bhattacharjee, S., Das, A., Bhattacharya, U., Parui, S. K., Roy, S., (2015), "Sentiment analysis using cosine similarity measure," *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, Kolkata, pp. 27-32

Booth, D., and Jansen, B.J. (2009), "A review of methodologies for analysing websites," *IGI Glob*, pp.141–62.

Borzemski, L., (2007), "Internet Path Behavior Prediction via Data Mining: Conceptual Framework and Case Study," *Journal of Universal Computer Science*, vol. 13, no. 2, pp. 287-316

Cai, L. & Zhu, Y., (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*. 14, p.2.

Castellano, G., Fanelli, A.M., and Torsello, M.A. (2013), "Web Usage Mining: Discovering Usage Patterns for Web Applications," *Adv. Tech. Web Intel*, 1.-2 75.

Chakraborty, A., Ghosh, S., Ganguly, N., and Gummadi, K.P. (2017), "Optimizing the Recency-Relevancy Trade-off in Online News Recommendations," *In Proceedings of the 26th International Conference on World Wide Web - WWW '17*, (Perth, Australia: ACM Press), pp. 837–846.

Chen, L., and Su, Q. (2013), "Discovering user's interest at E-commerce site using clickstream data," *In 2013 10th International Conference on Service Systems and Service Management*, pp. 124–129.

Cheng, J., Liu, Y., Zhang, H., Wu, X., Fuzhen Chen, F., (2015), "A New Recommendation Algorithm Based on User's Dynamic Information in Complex Social Network," *Mathematical Problems in Engineering*, vol. 2015, Article ID 281629, 6 pages

Chitraa,V and Thanamani, A. S.(2014), "Web Log Data Analysis by Enhanced Fuzzy C Means Clustering," *Int. J. Comput. Sci. Appl.*, vol. 4, no. 2, pp. 81–95.

Cooley, R., Tan, P., Srivastava, J., (1999), "Discovery of Interesting Usage Patterns from Web Data," *Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, p.163-182

Das, S.N., Mathew, M., and Vijayaraghavan, P.K. (2012), "An Efficient Approach for Finding Near Duplicate Web Pages Using Minimum Weight Overlapping Method," *In 2012 Ninth International Conference on Information Technology - New Generations*, pp. 121–126.

Demšar, J., Zupan, B., Leban, G., and Curk, T. (2004), "Orange: From Experimental Machine Learning to Interactive Data Mining," *In Knowledge Discovery in Databases: PKDD 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 537–539.

Dhandi, M., and Chakrawarti, R.K. (2016), "A comprehensive study of web usage mining," *In 2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, pp. 1–5.

Dias, J.P. and Ferreira, H.S., (2017), "Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites," *The 8th International Conference on Ambient Systems, Networks and Technologies (ANT 2017)*, pp 297-304

Dimitrijevic, M., Subic, N., Bosnjak, Z., (2014) "Improving the interestingness of web usage association rules containing common web site menu items," *Online Journal of Applied Knowledge Management*, Volume 2, Issue 1, pp82-92

Doan, T., and Kalita, J. (2015), "Selecting Machine Learning Algorithms Using Regression Models," *In 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, (Atlantic City, NJ, USA: IEEE), pp. 1498–1505.

Dohare, M.P.S., Arya, P., and Bajpai, A. (2012), "Novel web usage mining for web mining techniques," *Int. J. Emerg. Technol. Adv. Eng.* 2, 253–262.

Dutta, A., Paria, S., Golui, T., and Kole, D.K. (2014a), "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)*, pp. 1445–1451.

Dutta, A., Paria, S., Golui, T., and Kole, D.K. (2014b), "Noise Elimination from Web Page Based on Regular Expressions for Web Content Mining," *Adv. Comput. Netw. Inform.* Vol. 1 545.

Duwairi, R., and Ammari, H. (2016), "An enhanced CBAR algorithm for improving recommendation systems accuracy," *Simul. Model. Pract. Theory* 60, 54–68.

Dwivedi, S.K., and Rawat, B. (2015), "A review paper on data preprocessing: A critical phase in web usage mining process," *In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 506–510.

Fan, X.-X., Chow, K.-P., and Xu, F. (2014), "Web User Profiling Based on Browsing Behavior Analysis," *In Advances in Digital Forensics X*, (Springer, Berlin, Heidelberg), pp. 57–71.

- Fatima, B., Ramzan, H., Asghar, S. (2016), "Session identification techniques used in web usage mining: A systematic mapping of scholarly literature," *Online Information Review*, Vol. 40 Issue: 7, pp.1033-1053
- Frenay, B. and Verleysen, M. (2014) "Classification in the Presence of Label Noise: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol.25, no.5, pp.845-869
- Forsati, R., and Meybodi, M.R. (2010), "Effective page recommendation algorithms based on distributed learning automata and weighted association rules," *Expert Syst. Appl.* 37, 1316–1330.
- Garcia, L. P. F., Lorena, A. C., Carvalho, A. C. P.L.F., (2012) "A study on class noise detection and elimination", *IEEE Proc 2012 Braz. Symp. Neural Networks*, pp. 13-18, 2012.
- García-Gil, D., Luengo, J., García, S., and Herrera, F. (2017), "Enabling Smart Data: Noise filtering in Big Data classification. ArXiv170401770 Cs.
- Garg, A., and Kaur, B. (2014), "Enhancing Performance of Web Page by Removing Noises using LRU," *Int. J. Comput. Appl.* 103.
- Gasparetti, F., Micarelli, A., and Sansonetti, G. (2014), "Mining navigation histories for user need recognition," *In International Conference on Human-Computer Interaction*, (Springer), pp. 169–173.
- Gao, M., Lim, E., Lo, D., Prasetyo, P.K., (2016), "On detecting maximal quasi antagonistic communities in signed graphs," *Data Mining and Knowledge Discovery*, v.30 n.1, p.99-146, January 2016
- Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A., (2007), "User profiles for personalized information access," *The adaptive web: methods and strategies of web personalization*, Springer-Verlag, Berlin, Heidelberg, 2007
- Goel, R. (2014), "Enhanced Web Mining Technique To Clean Web Log File," *International Journal of Computer Applications* (0975 – 8887), Vol 96– No.16, pp. 25-29
- Grčar, M., Mladenič, D., and Grobelnik, M. (2005), "User profiling for interest-focused browsing history," *In Proceedings of the Workshop on End User Aspects of the Semantic Web*, pp. 99–109.
- Gu, W., Dong, S., Zeng, Z., and He, J. (2014), "An Effective News Recommendation Method for Microblog User," *The Scientific World Journal*, vol. 2014, Article ID 907515
- Guebas, A., Addam, O., Zaarour, O., Nagi, M., Elhajj, A., Ridley, M., and Alhajj, R. (2013), "Effective web log mining and online navigational pattern prediction," *Knowledge-Based Syst.* Vol.49, 50–62.
- Gunduz-Oguducu, S. (2010), "Web Page Recommendation Models Theory and Algorithms," *Synth. Lect. Data Manag.* Vol.2, 1–85
- Gupta, R., Shah, A., Thakkar, A., and Makvana, K. (2016), "A Survey on Various Web Page Ranking Algorithms," *An international journal of advanced computer technology*, 5 (1), (Volume-V, Issue-I) 8.

Halfaker, A., Keyes, O., Kluver, D., Thebault-Spieker, J., Nguyen, T., Shores, K., Uduwage, A., and Warncke-Wang, M. (2014), "User Session Identification Based on Strong Regularities in Inter-activity Time," *WWW '15 Proceedings of the 24th International Conference on World Wide Web*, pp. 410-418

Han, Y. and Xia, K., (2014) "Data Preprocessing Method Based on User Characteristic of Interests for Web Log Mining," *2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, Harbin, pp. 867-872

Hasan, O., Habegger, B., Brunie, L., Bennani, N., and Damiani, E. (2013), "A discussion of privacy challenges in user profiling with big data techniques: The eexcess use case," *In Big Data (BigData Congress), 2013 IEEE International Congress On*, (IEEE), pp. 25–30.

Hawalah, A., and Fasli, M. (2015), "Dynamic user profiles for web personalisation," *Expert Syst. Appl.* Vol. 42, pp. 2547–2569.

Hofgesang, P.I. (2006), "Relevance of time spent on web pages," *In Proceedings of KDD Workshop on Web Mining and Web Usage Analysis, in Conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.

Holub, M., Bielikova, M. (2010), "Estimation of user interest in visited web page," *Proceedings of the 19th international conference on World wide web*, pp.1111-1112

Hossin M, and Sulaiman M.N (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* 5, 01–11.

Htwe, T., and Kham, N.S.M. (2011). Extracting data region in web page by removing noise using DOM and neural network. *In 3rd International Conference on Information and Financial Engineering*, pp. 1-8

Hu, F., Li, M., Zhang, Y.N., Peng, T., and Lei, Y. (2013), "A Non-Template Approach to Purify Web Pages Based on Word Density," *In Proceedings of the International Conference on Information Engineering and Applications (IEA)*, (Springer, London), pp. 221–228.

Hu, J., Zeng, H.-J., Li, H., Niu, C., and Chen, Z. (2007), "Demographic prediction based on user's browsing behaviour," *In Proceedings of the 16th International Conference on World Wide Web, (ACM)*, pp. 151–160.

Huang, X., Cercone, N., & Aijun, A. (2002), "Comparison of interestingness functions for learning web usage patterns," *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management*, McLean, VA, USA, November 4-9, 2002, pp617-620

Huidrom, N., and Bagoria, N. (2013), "Clustering Techniques for the Identification of Web User Session," *International Journal of Scientific and Research Publications*, Volume 3, Issue 1, pp. 1- 8.

Hussain, T., Asghar, S., and Masood, N. (2010). Web usage mining: A survey on preprocessing of web log file. *In 2010 International Conference on Information and Emerging Technologies*, pp. 1–6.

Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A., (2015), "*Recommendation systems: Principles, methods and evaluation*," Egyptian Informatics Journal, Vol 16, Issue 3, pp261–273

Jafari, M., SoleymaniSabzchi, F., and Jamali, S. (2013), "Extracting Users' Navigational Behavior from Web Log Data: a Survey," *J. Comput. Sci. Appl. J. Comput. Sci. Appl.* 1, pp39–45.

Jansen, B.J., Booth, D.L., and Spink, A. (2009), "Patterns of query reformulation during Web searching," *Journal of the American Society for Information Science and Technology* Vol.60, pp.1358–1371.

Jawaheer, G., Weller, P., and Kostkova, P. (2014), "Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback," *ACM Trans. Interact. Intell. Syst.* Vol.4, pp.1–26.

Jiang, K., and Yang, Y. (2015), "Noise Reduction of Web Pages via Feature Analysis," *In 2015 2nd International Conference on Information Science and Control Engineering*, pp. 345–348.

John, M., and Jayasudha, J.S. (2016), "Methods for Removing Noise from Web Pages: A Review," *International Research Journal of Engineering and Technology*, Vol. 3, Issue 8, pp. 1908-1912

Joshila Grace, L.K., Maheswari, V., and Nagamalai, D. (2011), "Analysis of Web Logs And Web User In Web Mining," *Int. J. Netw. Secur. Its Appl.* 3, 99–110.

Kabir, S., Mudur, S.P., and Shiri, N. (2012), "Capturing Browsing Interests of Users into Web Usage Profiles," *In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp.18-25

Kaddu, M.R., and Kulkarni, R.B. (2016), "To extract informative content from online web pages by using hybrid approach," *In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 972–977.

Kakol, M., Nielek, R., Wierzbicki, A., (2017), "*Understanding and predicting Web content credibility using the Content Credibility Corpus*," *Information Processing and Management* Vol 53, Issue 5, pp1043–1061

Kanoje, S., Girase, S., Mukhopadhyay, D., (2014), "*User Profiling Trends, Techniques and Applications*," *International Journal of Advance Foundation and Research in Computer (IJAFRC)*, Volume 1, Issue 1

Kapusta, J., Munk, M., and k, M.D. (2012), "Cut-off time calculation for user session identification by reference length," *In 2012 6th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–6.

Kapusta, J., Munk, M., Svec, P., and Pilkova, A. (2014), "Determining the Time Window Threshold to Identify User Sessions of Stakeholders of a Commercial Bank Portal," *Procedia Comput. Sci.* 29, 1779–1790.

Kasliwal, A.D., and Katkar, D.G.S. (2015), "Web Usage mining for Predicting User Access Behaviour," *International Journal of Computer Science and Information Technologies*, Vol. 6 (1), pp. 201-204

- Kavitha, D., and Kalpana, B. (2017). Dynamic Log Session Identification Using A Novel Incremental Learning Approach For Database Trace Logs.
- Kellar, M., Watters, C., Duffy, J., and Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest. *Proc. Am. Soc. Inf. Sci. Technol.* 41, 168–175.
- Khasawneh, N. and Chan, C.-C., (2006), “Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining,” *In WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 325-328
- Khosla, A., Cao, Y., Lin, C.C.-Y., Chiu, H.-K., Hu, J., and Lee, H. (2010). An integrated machine learning approach to stroke prediction. *In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*, (Washington, DC, USA: ACM Press), p. 183.
- Kim, H., and Chan, P.K. (2005). Implicit indicators for interesting web pages, <https://dspace-test.lib.fit.edu/handle/11141/162>
- Ko Ko, S. and Jiamthapthaksin, R. "A Categorized Item Recommender System Coping with User Interest Changes," *International Journal of Machine Learning and Computing* vol. 4, no. 5, pp. 399-404
- Laber, E.S., de Souza, C.P., Jabour, I.V., de Amorim, E.C.F., Cardoso, E.T., Rentería, R.P., Tinoco, L.C., and Valentim, C.D. (2009), “A fast and simple method for extracting relevant content from news webpages,” *In Proceedings of the 18th ACM Conference on Information and Knowledge Management*, (ACM), pp. 1685–1688.
- Lagun, M., Lalmas, M., (2016), "Understanding User Attention and Engagement in Online News Reading," *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp.113-122
- Lalor, J.P., Wu, H., and Yu, H. (2017), “CIFT: Crowd-Informed Fine-Tuning to Improve Machine Learning Ability,” pp.1-10, ArXiv170208563 Cs.
- Lavanya, J., and Vardhini, P.A.A. (2014), “Extracting Users Interest From Web Log Files,” *International Journal of Computer Systems*, Vol-12, Issue.1, pp.1-7
- Lingwal, S. (2013). Noise Reduction and Content Retrieval from Web Pages. *International Journal of Computer Applications* (0975 – 8887), Volume 73– No.4, pp.24-30
- Liu B., Mobasher B., Nasraoui O. (2011), “Web Usage Mining. In: Web Data Mining”. *Data-Centric Systems and Applications*. 527-603, Springer, Berlin, Heidelberg
- Liu, H., and Kešelj, V. (2007), “Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users’ Future Requests,” *Data Knowl Eng* 61, pp.304–330.
- Liu, J., Zhang, S., and Yang, J. (2004), “Characterizing Web usage regularities with information foraging agents,” *IEEE Trans. Knowl. Data Eng.* 16, 566–584.

Lokeshkumar, R., and Sengottuvelan, P. (2015), "A Novel Approach to Improve Users Search Goal in Web Usage Mining," *International Journal of Computer, Electrical, Automation, Control and Information Engineering* Vol:9, No:2, pp. 624-628

Lopes, P. and Roy, B., (2015, "Dynamic Recommendation System Using Web Usage Mining for E-commerce Users," *International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Volume 45, pp60-69

Malarvizhi, S.P., and Sathiyabhama, B. (2014), "Enhanced reconfigurable weighted association rule mining for frequent patterns of web logs," *Int. J. Comput.* 13, 97–105.

Mayil, V.V., (2012), "Web Navigation Path Pattern Prediction using First Order Markov Model and Depth first Evaluation," *International Journal of Computer Applications* (0975 – 8887), Vol.45– No.16

Mezghani, M., Péninou, A., Zayani, C. A., Amous, I., & Sèdes, F. (2014). Analyzing tagged resources for social interests detection". *International Conference on Enterprise Information Systems (ICEIS)*, pp. 340–345.

Mehak, Mukesh Kumar, Naveen Aggarwal, (2013), "Web Usage Mining: An Analysis", *Journal of emerging technologies in web intelligence*, vol. 5, no. 3, pp.240-246, 2013

Mishra, S.N., Jaiswal, A., and Ambhaikar, A. (2012), "An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis," *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.2, Issue. 4, pp.278-282

Mobasher, B., Cooley, R., and Srivastava, J. (2000), "Automatic Personalization Based on Web Usage Mining," *Communication of the ACM*, 43, 142–151, doi>10.1145/345124.345169

Munk, M., Benko, L., Gangur, M., and Turcani, M. (2015), "Influence of Ratio of Auxiliary Pages on the Pre-Processing Phase of Web Usage Mining," *E M Ekon. Manag.* 18, 144–159.

Nanda, A., Omanwar, R., and Deshpande, B. (2014), "Implicitly Learning a User Interest Profile for Personalization of Web Search Using Collaborative Filtering." *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 54–62.

Narwal, N. (2013), "Improving web data extraction by noise removal," *Fifth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2013)*, pp. 388–395.

Neelima, G., and Rodda, S. (2016), "Predicting user behavior through sessions using the web log mining," *2016 International Conference on Advances in Human Machine Interaction (HMI)*, pp. 1–5.

Nithya, P., and Sumathi, P. (2012). Novel pre-processing technique for web log mining by removing global noise and web robots. *2012 National Conference on Computing And Communication Systems*, pp. 1–5.

Onyancha, J., Plekhanova, V., and Nelson, D. (2017). Noise Web Data Learning from a Web User Profile: Position Paper. *In Proceedings of the World Congress on Engineering*, pp. 608–611.

Ou, J.-C., Lee, C.-H., and Chen, M.-S. (2008), "Efficient algorithms for incremental Web log mining with dynamic thresholds," *The VLDB Journal* (2008) 17: pp.827-845. <https://doi.org/10.1007/s00778-006-0043-9>

Page 103.2 Qi, X., and Davison, B.D. (2009), "*Web page classification: Features and algorithms*", *ACM Computing Surveys*, Vol. 41, No. 2, Article 12, pp. 41, 1–31.

Pappas, N., Katsimpras, G., & Stamatatos, E. (2012), "Extracting informative textual parts from web pages containing user-generated content," *In Proceedings of the 12th international conference on knowledge management and knowledge technologies - i-know '12* (p. 1). New York, New York, USA: ACM

Pasi, G. (2014), "Implicit Feedback through User-system Interactions for Defining User Models in Personalized Search," 6th International conference on Intelligent Human Computer Interaction, 39, pp. 8–11.

Patel, P., and Parmar, M. (2014), "*Improve heuristics for user session identification through web server log in web usage mining*," *International Journal of Computer Science and Information Technologies*, Vol. 5 (3), pp.3562-3565

Patil,U.M and Patil,J. B., (2012), "Web data mining trends and techniques," *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pp961-965.

Pazzani , M and Billsus, D. (1997) Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning*, v.27 n.3, p.313-331

Pechenizkiy, M., Tsymbal, A., Puuronen, S., Pechenizkiy, O. (2010), "Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction," *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pp. 708-713

Podpecan, V., Zemenova, M., and Lavrac, N. (2012), "*Orange4WS Environment for Service-Oriented Data Mining*," *The Computer Journal*, Volume 55, Issue 1, Pp. 82–98, <https://doi.org/10.1093/comjnl/bxr077>

Poo, D., Chng, B., and Goh, J.-M. (2003), "A hybrid approach for user profiling," *HICSS '03 Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4 - Volume 4*. Pp.103.2

Qi, X., and Sun, J. (2011). Eliminating Noisy Information in Webpage through Heuristic Rules. *In 2011 International Conference on Computer Science and Information Technology*, pp. 137-141

Qiu, F. and Cho, J. (2006) Automatic Identification of User Interest for Personalized Search," *In Proceedings of the 15th International Conference on World Wide Web*, pp.727-736.

Rahman, G. and Islam, Z. (2011), "A Novel Framework Using Two Layers of Missing Value Imputation," *Proceedings of the 11-th Australasian Data Mining Conference (AusDM'13), Canberra, Australia*, pp.149-160

Ramya, C., Kavitha, G., and Shreedhara, D.K. (2011), "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process," *International Journal of Information and Electronics Engineering*, Vol. 3, No. 2, pp.196-199, ArXiv Prepr. ArXiv11050350.

Rao, K.S., Babu, D.A.R., and Krishnamurthy, D.M. (2017), "*Mining User Interests from User Search by Using Web Log Data*," *Journal of Web Development and Web Designing*, Volume 2 Issue 1, pp.1-4

Rebon, F., Ocariz, G., and Argandona, J. (2015). In *Information and Communication Technologies in Tourism 2015: Proceedings of the International Conference in Lugano, Switzerland, February 3 - 6, 2015*, Eds. Iis Tussyadiah, Alessandro Inversin, (Springer), pp. 101–109.

Reusens, M., Lemahieu, W., Baesens, B., and Sels, L. (2017), "A Note on Explicit Versus Implicit Information for Job Recommendation," *Decision Support System*, Vol. 98, 26–35.

Saad, F. (2014), "Baseline Evaluation: An Empirical Study of the Performance of Machine Learning Algorithms in Short Snippet Sentiment Analysis," *In Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business*, (New York, NY, USA: ACM), pp. 6:1–6:8.

Sáez, J.A., Galar, M., Luengo, J., and Herrera, F. (2016). INFFC: *An iterative class noise filter based on the fusion of classifiers with noise sensitivity control*. *Information Fusion* Vol.27, pp.19–32,

Sáez, J.A., Luengo, J., and Herrera, F. (2013), "*Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification*," *Pattern Recognition*. Vol. 46, 355–364, <https://doi.org/10.1016/j.patcog.2012.07.009>

Sáez, J.A., Luengo, J., Stefanowski, J., and Herrera, F. (2015), "*SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering*," *Information Science*. Vol.291, 184–203.

Sahar S (2010) Interestingness measures—on determining what is interesting. In: Maimon O, Rokach L (eds) *Data mining and knowledge discovery handbook*, 2nd edn. Springer, New York, pp 603–612.

Sambhanthan, A., and Good, A. (2013). *Implications for Improving Accessibility to E-Commerce Websites in Developing Countries: A Subjective Study of Sri Lankan Hotel Websites*.

Santra, A.K., and Jayasudha, S. (2012), "*Classification of web log data to identify interested users using Naïve Bayesian classification*," *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 1, pp.381–387.

Sengottuvelan, P., Lokeshkumar, R., and Gopalakrishnan, T. (2017), "*An Improved Session Identification Approach in Web Log Mining for Web Personalization*," *Journal of Internet Technology*, vol. 18, no. 4 , pp. 723-730

Shanab, A.A., Khoshgoftaar, T. M., Wald, R., Napolitano, A., (2012), "Impact of noise and data sampling on stability of feature ranking techniques for biological datasets", *2012 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 415-422

Sharma, N., and Makhija, P. (2015), "*Web usage mining: A Novel Approach for Web user Session Construction*," *Global Journal of Computer Science and Technology: (E), Web & Security*, Volume 15 Issue 3, pp. 14-18

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017) *Fake News Detection on Social Media: A Data Mining Perspective*, *ACM SIGKDD Explorations Newsletter*, v.19, Issue.1, Pages 22-36

Singla, A. and White, R.W., 2010, "Sampling high-quality clicks from noisy click data," *Proceedings of the 19th international conference on World Wide Web*, Pages 1187-1188

Sirsat, S. and Chavan, V., (2016), "Pattern matching for extraction of core contents from news web pages," *2016 Second International Conference on Web Research (ICWR)*, Tehran, pp. 13-18.

Sivakumar, P. (2015). Effectual Web Content Mining using Noise Removal from Web Pages. *Wireless Pers. Communication.* 84, 99–121, //doi.org/10.1007/s11277-015-2596-7

Sreedhar, G. (2016). *Design Solutions for Improving Website Quality and Effectiveness* 1st, IGI Publishing Hershey, PA, USA ©2016, ISBN: 1466697644 978146669764

Sripriya, J., and S. Samundeeswari, E. (2012). *Comparison of Neural Networks and Support Vector Machines using PCA and ICA for Feature Reduction*. *International Journal of Computer Applications* (0975 – 8887) Vol, 40– No.16. 40, 31–36.

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N. (2000). *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. *SIGKDD Explorations Newsletter*, Vol.1, Issue 2, pp.12–23.

Stoica, A. (2012), "Filtering Noisy Web Data by Identifying and Leveraging Users' Contributions," *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Trinity College in Dublin, Ireland, June 4–8, 2012*, pp.583-586

Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004), "Adaptive web search based on user profile constructed without any effort from users," *In Proceedings of the 13th Conference on World Wide Web - WWW '04, (New York, NY, USA: ACM Press)*, p. 675.

Suguna, R., and Sharmila, D. (2013), "*User interest level based preprocessing algorithms using web usage mining*," *International Journal on Computer Science and Engineering (IJCSE)*, Vol. 5, No. 09, pp.815-822.

Suksawatchon, U., Darapisut, S., and Suksawatchon, J. (2015), "Incremental session based collaborative filtering with forgetting mechanisms," *In 2015 International Computer Science and Engineering Conference (ICSEC)*, pp. 1–6.

Sunitha, L., Bal Raju, M., Sunil Srinivas, B., A Comparative Study between Noisy Data and Outlier Data in Data Mining," *International Journal of Current Engineering and Technology*, Vol.3, No.2, pp575-577

Swe Swe Nyein (2011) "Mining Contents in Web Page Using Cosine Similarity". *2011 3rd International Conference on Computer Research and Development*, pp. 472-475.

Tan,B., Lv,Y., Zhai, C. (2012), "Mining long-lasting exploratory user interests from search history," *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp.1477-1481

Tang, J., Yao, L., Zhang, D., Zhang, J., (2010) A Combination Approach to Web User Profiling, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, v.5 n.1, p.1-44

Tavakolian, R., Beheshti, M.T.H., and Charkari, N.M. (2012), "An Improved Recommender System Based on Forgetting Mechanism for User Interest-Drifting," *International Journal of Information & Communication Technology Research*, Vol.4, No. 4, pp. 69–77.

Taylor, J.C., and Fenner, J.W. (2017). *Comparison of machine learning and semi-quantification algorithms for (1123) FP-CIT classification: the beginning of the end for semi-quantification?* *European Journal of Nuclear Medicine and Molecular Imaging Physics*, 4 (29). ISSN 1619-7070

Ting, I.-H., and Wu, H.-J. (2009), "Web Mining Techniques for On-Line Social Networks Analysis: An Overview," In: Ting IH, Wu HJ. (eds) *Web Mining Applications in E-commerce and E-services*. *Studies in Computational Intelligence*, vol 172, pp.169-179 Springer, Berlin, Heidelberg

Tyagi, N., and Sharma, S. (2012), "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page," *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-2, Issue-3, pp.441-446

Varnagar,C. R., Madhak,N. N., Kodinariya,T. M. and Rathod,J. N., (2013) "Web usage mining: A review on process, methods and techniques," *International Conference on Information Communication and Embedded Systems (ICICES)*, Chennai, 2013, pp. 40-46

Velloso, R.P., and Dorneles, C.F. (2013), "Automatic Web Page Segmentation and Noise Removal for Structured Extraction using Tag Path Sequences," *Journal of Information and Data Management*, Vol. 4, No. 3, Pages 173–187.

Verma, P., and Kesswani, N. (2014), "Web Usage mining framework for Data Cleaning and IP address Identification," [2014arXiv1408.5460V](https://arxiv.org/abs/2014arXiv1408.5460V)

Vidyavathi, B.M. and Begum,H. (2016), "An Efficient Web Recommender System for Web Logs," *International Journal of Computer Applications* (0975 – 8887) Volume 152 – No.3, pp.9-12

Waegeman, W., Verwaeren, J., Slabbinck, B., and De Baets, B. (2011), "Supervised learning algorithms for multi-class classification problems with partial class memberships," *Fuzzy Sets Syst.* 184, 106–125.

- Wang, C., Lu, J., and Zhang, G. (2007), "Mining Key Information of Web Pages," *A Method and Its Application. Expert Syst Appl* 33, 425–433.
- Wang, H., Xu, Q., and Zhou, L. (2014), "Deep Web Search Interface Identification," *A Semi-Supervised Ensemble Approach. Information* 5, 634–651.
- Wang, X., Chen, B., and Chang, F. (2011), "A Classification Algorithm for Noisy Data Streams with Concept-Drifting," *Journal of Computational Information Systems* 7: 12, 4392-4399
- Wei, X., Wang, Y., Li, Z., Zou, T., and Yang, G. (2015), "Mining Users Interest Navigation Patterns Using Improved Ant Colony Optimization," *Intelligent Automation & Soft Computing*, 21:3, 445-454, DOI: 10.1080/10798587.2015.1015778
- Wiedmann, K.-P., Buxel, H., and Walsh, G. (2002), "Customer profiling in e-commerce: Methodological aspects and challenges," *J. Database Mark. Cust. Strategy Manag.* 9, 170–184.
- Wu, S., Xiaonan, Z., and Yannan, D. (2015), "A Collaborative Filtering Recommender System Integrated with Interest Drift Based on Forgetting Function," *Int. J. U- E-Serv. Sci. Technol.* 8, 247–264.
- Wu, X., Wang, P., and Liu, M. (2014), "A Method of Mining User's Interest in Intelligent e-Learning," *International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2014)*, pp. 74 – 76.
- Wu, X. and Zhu, X. (2008) "Mining with noise knowledge: Error-Aware Data Mining," *IEEE Transactions on Systems, Man, and Cybernetics*, 38, pp. 917–932
- Xinhua, H., and Qiong, W. (2011), "Dynamic timeout-based a session identification algorithm," *In 2011 International Conference on Electric Information and Control Engineering*, pp. 346–349.
- Xiong, H., Pandey, G., Steinbach, M., and Kumar, V. (2006), "Enhancing data analysis with noise removal," *in IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 304-319
- Yang H., Fong S. (2011) Moderated VFDT in Stream Mining Using Adaptive Tie Threshold and Incremental Pruning. In: Cuzzocrea A., Dayal U. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2011. Lecture Notes in Computer Science, vol 6862, pp. 471-483, Springer, Berlin, Heidelberg
- Yazidi, A., Granmo, O.C., Oommen, B.J., (2011), "A User-Centric Approach for Personalized Service Provisioning in Pervasive Environments," *Wireless Personal Communications*, Volume 61, Issue 3, pp 543–566
- Yi, L., Liu, B., and Li, X. (2003), "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: ACM)*, pp. 296–305.
- Ying, J., Chin, C., Tseng, V. S., (2012), "Mining web navigation patterns with dynamic thresholds for navigation prediction," *2012 IEEE International Conference on Granular Computing*, Hangzhou, pp. 614-619.

- Yu, X., Li, M., Ah Kim, K., Chung, J., Ho Ryu, K., (2016), "Emerging Pattern-Based Clustering of Web Users Utilizing a Simple Page-Linked Graph," *Sustainability* (2071-1050), Vol. 8 Issue 3, p1-18.
- Yuankang, F., and Huang, Z. (2010). "A session identification algorithm based on frame page and page threshold," *3rd International Conference on Computer Science and Information Technology*, Chengdu, 2010, pp. 645-647. doi: 10.1109/ICCSIT.2010.5564697.
- Zahoor, S., Bedekar, M., and Kosamkar, P.K. (2014). User Implicit Interest Indicators learned from the Browser on the Client Side. (ACM Press), pp. 1–4.
- Zeng, Y., Zhong, N., Ren, X., and Wang, Y. (2012). "User Interests Driven Web Personalization Based on Multiple Social Networks," *In Proceedings of the 4th International Workshop on Web Intelligence & Communities*, (New York, NY, USA: ACM), pp. 9:1–9:4.
- Zhang W., Chen T. (2012) Data Pre-processing for Web Data Mining. In: Jin D., Lin S. (eds) *Advances in Electronic Commerce, Web Application and Communication. Advances in Intelligent and Soft Computing*, vol 149. Pp. 303-307, Springer, Berlin, Heidelberg
- Zhang, J. and Ghorbani, A.A. (2004), "Familiarity and Trust: Measuring Familiarity with a Web Site," *In Proceedings of the 2nd Annual Conference on Privacy, Trust and Security (PST 2004)*, pages 23—28
- Zhang, Y., and Deng, K. (2010). "Algorithm of web page purification based on improved DOM and statistical learning," *2010 International Conference on Computer Design and Applications*, pp. V5-288-V5-291.
- Zhang, H., Song, Y., Song, H., (2007) Construction of Ontology-Based User Model for Web Personalization, *Proceedings of the 11th international conference on User Modelling*, pp.67-76
- Zhuang L., Kou Z., Zhang C. (2005) Session Identification Based on Time Interval in Web Log Mining. In: Shi Z., He Q. (eds) *Intelligent Information Processing II. IIP 2004*. IFIP International Federation for Information Processing, vol 163. Springer, Boston, MA
- Zubi, K.S. and El Raiani, M.S. (2014), "Using Web Logs Dataset via Web Mining for User Behavior Understanding," *International Journal of Computers and Communications*, Vol.4, pp.103-111
- Zhu, X and Wu, X, (2004) "Class noise vs. attribute noise: A quantitative study," *Artif. Intell. Rev.*, vol. 22, pp. 177–210

Appendices


Appendix I: Samples of noise web data

FREE/OPEN-SOURCE HARDWARE GOOGLE APPLE MICROSOFT HOW-TOS PROGRAMMING NEWS BOOKS MOVIES

5 of the Best Free and Open Source Data Mining Software

POSTED BY JUN AUZA ON 11/25/2010


The process of extracting patterns from data is called data mining. It is recognized as an essential tool by modern business since it is able to convert data into business intelligence thus giving an informational edge. At present, it is widely used in profiling practices, like surveillance, marketing, scientific discovery, and fraud detection.




There are four kinds of tasks that are normally involve in Data mining:


- * Classification - the task of generalizing familiar structure to employ to new data
- * Clustering - the task of finding groups and structures in the data that are in some way or another the same, without using noted structures in the data.
- * Association rule learning - Looks for relationships between variables.
- * Regression - Aims to find a function that models the data with the slightest error.

For those of you who are looking for some data mining tools, here are five of the best open-source data mining software that you could get for free:

Orange
 Orange is a component-based data mining and machine learning software suite that features friendly yet powerful, fast and versatile visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It contains complete set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is written in C++ and Python, and its graphical user interface is based on cross-platform Qt framework.

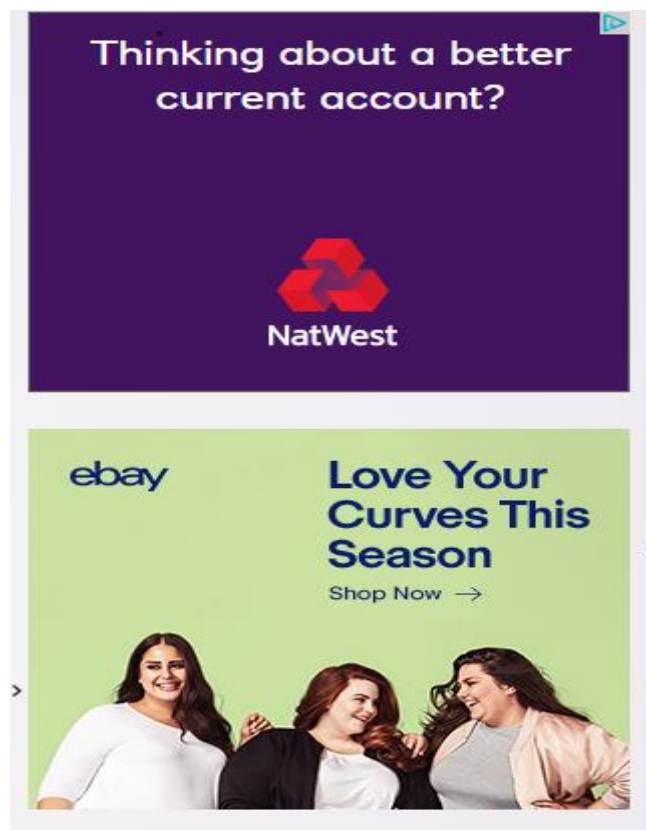
RapidMiner
 RapidMiner, formerly called YALE (Yet Another Learning Environment), is an environment for machine learning and data mining experiments that is utilized for both research and real-world data mining tasks. It enables experiments to be made up of

Search




Recent Featured Posts Categories

- 4 Cloud-based Applications that Work Perfectly on Linux
- Popular Hollywood Movies that Utilizes Linux
- Best Cross-Platform Note-taking Apps to Enhance Productivity
- Best Firefox Add-ons for a Better YouTube Experience
- Windows 10: Is it Really Worth Ditching Linux for?
- 5 Best Tools You Need To Create Your Next Big Android App
- 5 Best Calendar Apps for Google Chrome
- Using Android Apps for Keeping the Family Safe
- 7 Best Chrome Apps and Extensions for



Thinking about a better current account?




NatWest

ebay

Love Your Curves This Season

Shop Now →



Source: <http://www.msn.com/en-gb/?ocid=iehp> Accessed on 23rd July 2016

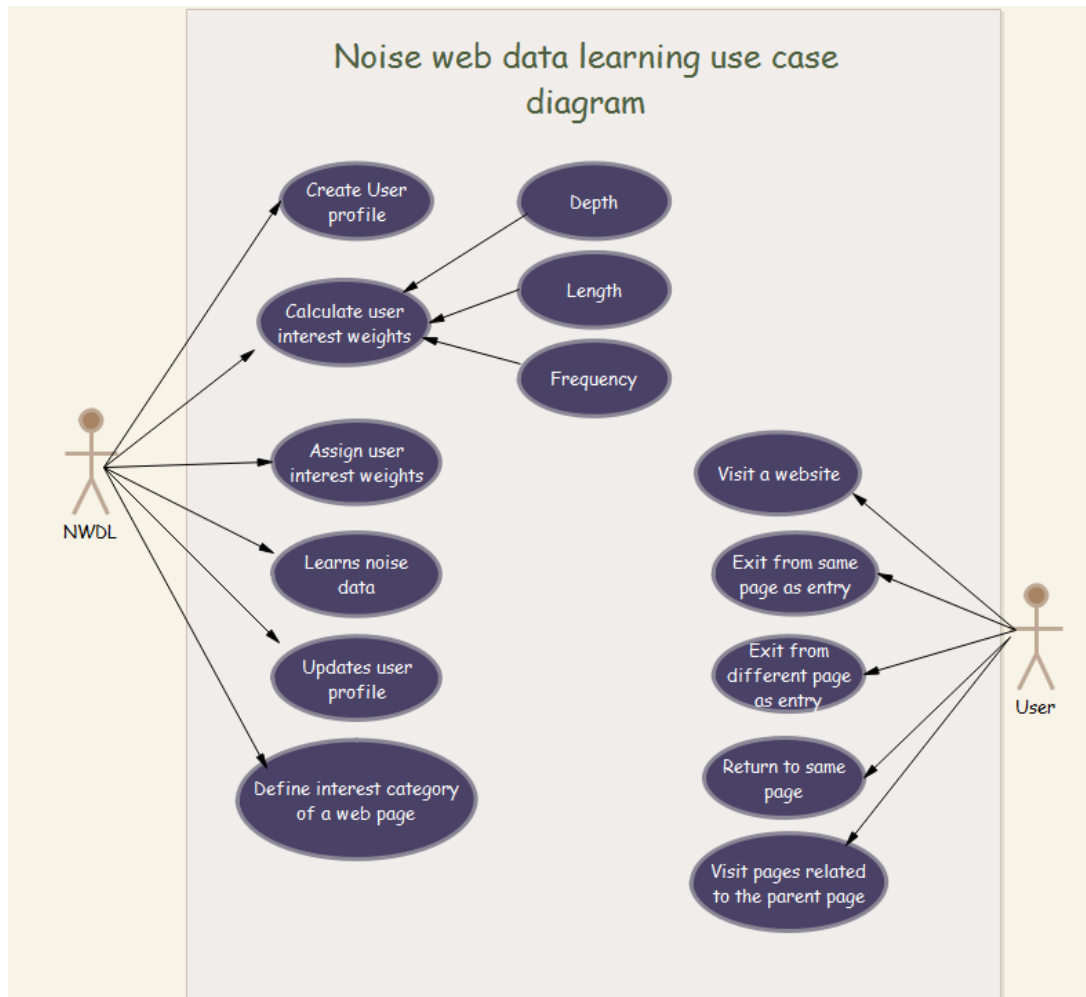
Appendix II: Comparative analysis of data mining techniques applied in noise web data reduction

KNN= K-Nearest Neighbours, **NB**= Naive Bayesian Classification, **SVM**= Support Vector Machines, **NN**=Neural Networks, **KM**= K-Means Clustering, **FCM**= Fuzzy C-Means Clustering, **ARM** = Association Rule Mining

Author	Technique	Objective	Input (Attributes)	Findings	Evaluation Criteria	Output
Adeniyi and Yongquan (2016)	KNN	To determine classes of unknown data instances during classification	News category, feed type, news daily name	All neighbouring data instance are encapsulated and assigned to the nearest class.	Evaluated against sample testing session data in terms of accurate recommendations	User interest data
Vidyapriya and Pushpa (2016)	KNN	To identify interest of user access pattern from web logs	User IP address, Session-id, page visits, source	The k- nearest neighbour algorithm shows a maximum accuracy and minimum error rate compared to NB	Evaluated WEKA and rapid miner.	Pattern set generated user access data
Santra and Jayasudha (2012)	NB	To classify interested and non-interested users	No of page views, time taken, URL page, User IP address	Consider many attributes irrelevant for classification	Results were evaluated based on time take to compute maximum likelihood of user interest against C4.5	The output has only two possible outcomes: interested users or not interested users.
Padmapriya and Hemalatha (2014)	NB	To Identify data instances whose classes are unknown	User IP address, Page URL, Session-id, page view duration	Classification attributes are independent with no relationship between them	Expectation-maximization	A pair of tag path occurrence user access patterns
Suchaka and Potempa (2015)	SVM	Classification of user session from web log file by eliminating irrelevant session and predicting buying sessions.	Session duration, No. of page views, session-id, user agent, IP address	SVM classifier are effective both in respect of the overall predictive accuracy and the ability to predict user interest i.e. buying session	Evaluated performance based on error rate, accuracy and sensitivity	Correctly and incorrectly classified web data logs.
Htwe et al., 2010	NN	To find noise pattern in current Web page by matching	Page URLs from news website	It is difficult to determine pure data region because data	Evaluated performance of their proposed tool by calculating levels of	Useful web data and Noisy web data

		similar noise pattern kept in Case-Based		regions of these sites are surrounded by noise.	noise eliminated from each data region.	
Kaur and Kaur (2013)	FCM	To reduce the amount of irrelevant data on the web and predict user interest	Web log data (User IP address, session duration, timestamp, User request)	User interest level depends on weight of each web page which is determined by session duration	Results are session oriented and page oriented in order to determine next page request	A tool capable of determining user next page request
Chandel et al., 2016	FCM	To discover patterns of user activities on a web page	IP address, session-id, timestamp	Efficiency of FCM is better than KM	Efficiency of the algorithm compared to K-Means	Web logs with effective usage pattern
Chitraa and Thanamani (2012)	KM	To determine only relevant logs that the user is interested in	User's IP address, Page duration, user's browser, operating system, No of page views	Defined clusters with similar intra objects are extracted while dissimilar inter objects are removed	K-means finds initial points and optimize for accurate results	Similar data cluster and Dissimilar data cluster
Ramya and Sajeev (2015)	KM	To suggest web pages to users based on their interest	IP address, requested page, visit duration, access method, timestamp	Classification of users is not only based on web access patterns but also on their dynamic interest.	Evaluated against SVM classifier to determine frequency of visit and time duration	Pattern set generated user access data
Langhnoja et al., 2013	ARM	To find user access patterns based on user interest	User IP address, timestamp, URL requested, URL referrer, user agent	All irrelevant entries are removed prior to applying ARM	Results are compared with clustered access patterns in terms of accuracy	Log Database with effective usage pattern
Malarvizhianan and Sathiyabhama (2014)	ARM	To find unwanted pages visited by the students during a suspected duration which obviously affects their progress	IP address, Session-id, Page URL, No. of pages views,	Pages visited only once by only one visitor are considered irrelevant.	Efficiency and execution time	Weighted association rules

Appendix III: Data Modelling: Use Case Diagram



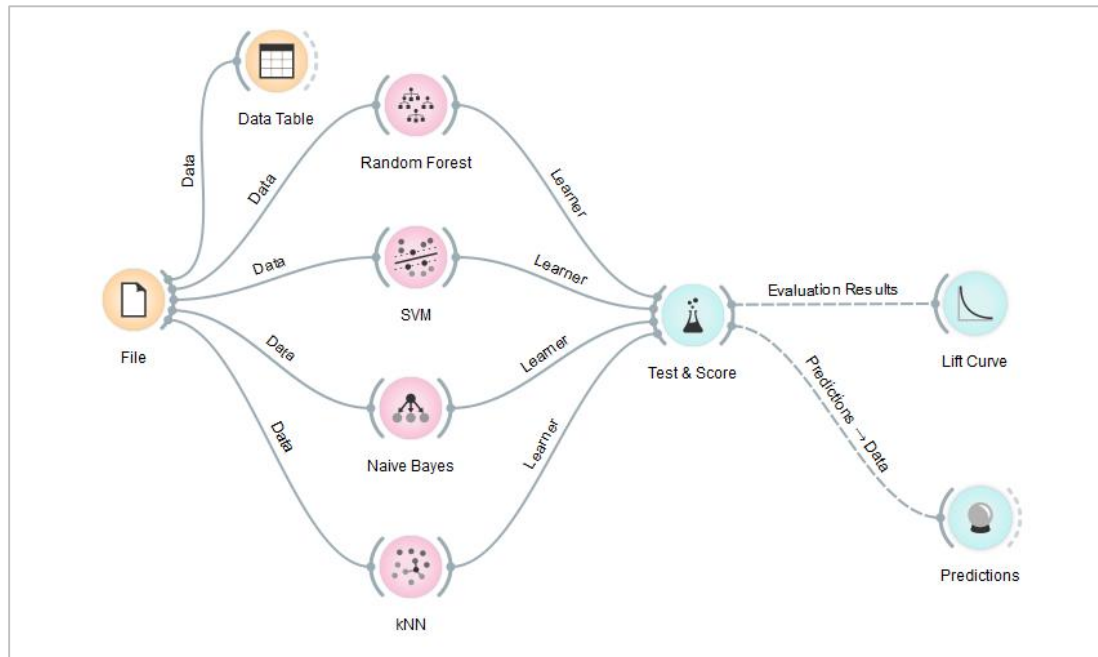
Appendix II: Noise Web Data Learning Approach – Use Case Diagram

Appendix IV: Sample raw datasets

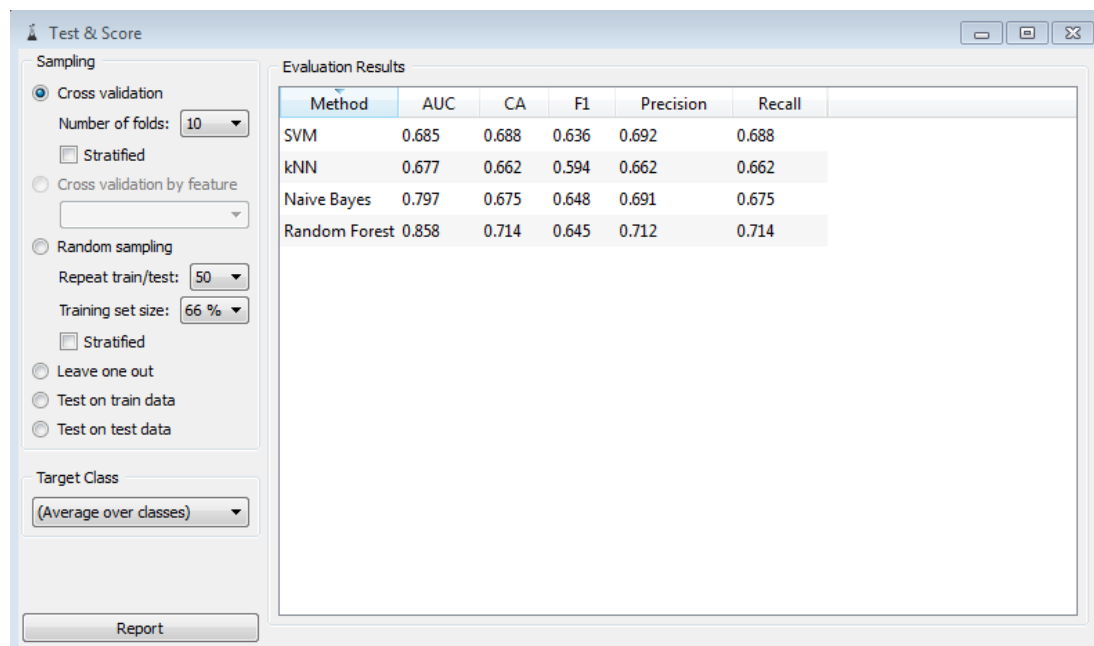
ID	Session_ID	URL_ID	User_ID	URL	Time_Stamp	D	Device_Type	Source	URL_Category
1	200	1057	1001	http://www.kili...	2016-04-06 00:2...	2...	mobile	referral	Phones and Ac...
2	201	1309	1001	http://www.kili...	2016-04-06 00:5...	1...	desktop	direct	Home and Living
3	533	1309	1001	http://www.kili...	2016-04-07 01:1...	1...	desktop	referral	Home and Living
4	921	1309	1001	http://www.kili...	2016-04-08 03:1...	2...	desktop	direct	Home and Living
5	201	1404	1001	http://www.kili...	2016-04-06 01:0...	5	desktop	referral	Electronic and ...
6	201	1544	1001	http://www.kili...	2016-04-06 01:0...	1	desktop	referral	Women's Cloth...
7	261	1306	1002	http://www.kili...	2016-04-06 00:2...	2...	desktop	direct	Computers and...
8	691	1306	1002	http://www.kili...	2016-04-07 07:4...	1...	desktop	referral	Computers and...
9	697	1306	1002	http://www.kili...	2016-04-07 07:5...	2...	mobile	referral	Electronic and ...
10	773	1306	1002	http://www.kili...	2016-04-08 08:4...	1...	desktop	direct	Electronic and ...
11	263	1369	1002	http://www.kili...	2016-04-06 01:1...	1...	desktop	organic	Home and Living
12	262	1384	1002	http://www.kili...	2016-04-06 00:4...	2...	desktop	referral	Home and Living
13	241	1077	1003	http://www.kili...	2016-04-06 00:2...	3	desktop	referral	Phones and Ac...
14	502	1077	1003	http://www.kili...	2016-04-07 07:4...	1...	desktop	referral	Phones and Ac...
15	241	1307	1003	http://www.kili...	2016-04-06 01:2...	2...	desktop	referral	Electronic and ...
16	502	1307	1003	http://www.kili...	2016-04-07 07:0...	1...	desktop	referral	Electronic and ...
17	518	1307	1003	http://www.kili...	2016-04-07 07:5...	2...	mobile	referral	Electronic and ...
18	881	1307	1003	http://www.kili...	2016-04-08 08:4...	9	desktop	referral	Electronic and ...
19	252	1328	1003	http://www.kili...	2016-04-06 01:4...	2...	desktop	direct	Home and Living
20	257	1329	1003	http://www.kili...	2016-04-06 02:1...	1...	desktop	direct	Home and Living
21	211	1078	1004	http://www.kili...	2016-04-06 00:2...	1	desktop	direct	Phones and Ac...
22	498	1078	1004	http://www.kili...	2016-04-07 07:4...	1...	desktop	referral	Phones and Ac...
23	217	1308	1004	http://www.kili...	2016-04-06 00:2...	3...	desktop	referral	Electronic and ...
24	496	1308	1004	http://www.kili...	2016-04-07 07:1...	2...	desktop	referral	Electronic and ...
25	872	1308	1004	http://www.kili...	2016-04-08 02:5...	8	desktop	referral	Electronic and ...
26	879	1308	1004	http://www.kili...	2016-04-08 08:5...	1...	mobile	referral	Home and Living
27	230	1079	1005	http://www.kili...	2016-04-06 00:3...	2	mobile	direct	Phones and Ac...
28	524	1079	1005	http://www.kili...	2016-04-07 07:4...	3...	desktop	referral	Phones and Ac...
29	230	1337	1005	http://www.kili...	2016-04-06 00:3...	4	desktop	referral	Home and Living
30	514	1337	1005	http://www.kili...	2016-04-07 07:0...	2...	desktop	referral	Home and Living
31	772	1337	1005	http://www.kili...	2016-04-08 08:4...	1...	desktop	referral	Home and Living
32	772	1337	1005	http://www.kili...	2016-04-08 08:5...	1...	desktop	organic	Home and Living
33	230	1383	1005	http://www.kili...	2016-04-06 00:4...	8	desktop	referral	Home and Living
34	519	1383	1005	http://www.kili...	2016-04-07 07:3...	1...	desktop	referral	Home and Living

Appendix IV: Raw Data imported to SQL Server

Appendix V: Experimental Design Procedures and Validation



Appendix V: Baseline Model Process



Appendix V: Output from Baseline Model Process