

MODELLING WEBSITE USER BEHAVIOR FROM WEB ACCESS LOGS

Ganihachchi Pathirannehelage Don Madhuka Udantha

(148016F)



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Dissertation submitted in partial fulfillment of the requirements for the degree Master
of Science

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

March 2016

DECLARATION

“I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

Name: G P D M Udantha



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

The above candidate has carried out research for the Masters Dissertation under my supervision.

Signature of the supervisor:

Date:

Name of the supervisor: Dr. Surangika Ranathunga

Signature of the co-supervisor:

Date:

Name of the supervisor: Prof. Gihan Dias

ABSTRACT

Mining web access log data is a popular technique to identify frequent access patterns of website users. Web logs can provide a wealth of information on the user access patterns of the corresponding website, if and when they are properly analyzed. However, finding interesting patterns hidden in the low-level log data is non-trivial due to large log volumes, and the distribution of the log files in cluster environments.

Existing clustering techniques have not focused on identifying infrequent patterns and most of the clustering techniques suffer from cluster parameter issues, when it comes to web usage mining. This thesis presents the application of Density Based Spatial Clustering of Applications with Noise (DBSCAN) and Expectation Maximization (EM) algorithms in an iterative manner for clustering, which is not a technique that has been used before. Each cluster corresponds to one or more web user activities. For clusters that did not have a unique access pattern, frequent pattern mining and sequence pattern mining techniques were used to identify the unique user access patterns.

Secondly, this thesis solves another issue in web usage mining – detecting slight changes between web user access sessions. This thesis introduces a method to identify these access patterns at a much lower level than what is provided by traditional clustering techniques, such as nearest neighbor based techniques and classification techniques. This technique makes use of the concept of episodes to represent web sessions. These episodes are expressed in the form of regular expressions. To the best of our knowledge, this is the first time that the concept of regular expressions are applied to identify user access patterns in web server log data.

We demonstrate that the implemented system is capable of not only identifying common user behaviors, but also in identify anomalous user behavior.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

ACKNOWLEDGEMENTS

I would like to dedicate my sincere thanks to my supervisors Dr. Surangika Ranathunga and Prof. Gihan Dias for their dedicated support for the success of this research. This would not have been a success without your support from the initial stage to the final phase of the research.

This research was supported by the LK Domain Registry, Sri Lanka. I thank our colleagues from the Research Division of LK Domain Registry who provided insight and expertise that greatly assisted the research.

I would like to thank the entire academic and non-academic staff of the Department of Computer Science and Engineering for their kindness extended to me in every aspect.

Last but not least, I thank my parents, my wife and all my friends who supported me for the success of this piece of work. Your support was very precious.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

TABLE OF CONTENTS

DECLARATION	i
ABSTRACT.....	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS.....	iv
LIST OF FIGURES	vii
LIST OF TABLES	ix
LIST OF ABBREVIATIONS	x
LIST OF APPENDICES	xi
1. INTRODUCTION	1
1.1. Motivation	2
1.2. Objectives	3
1.3. Contributions.....	4
1.3.1. Refereed Articles.....	5
1.4. Organization of the Thesis	5
2. LITERATURE SURVEY.....	7
2.1. Overview	7
2.2. Web Mining.....	7
2.3. Web Usage Mining and Applications	8
2.4. Data Preprocessing in Web Usage Mining.....	11
2.4.1. Data Selection	12
2.4.2. Web Log Data Cleaning.....	12
2.4.3. User Identification and Session Identification	14
2.5. Pattern Discovery	15
2.5.1. Association Rule Mining	16
2.5.2. Clustering	18
2.5.3. Sequential Patterns Discovery	21
2.5.4. Statistical Analysis for Web Usage Mining	24

2.6.	Anomaly Detection Techniques	24
2.6.1.	Anomaly Detection using Clustering Techniques	25
2.6.2.	Anomaly Detection using Classification Techniques	26
2.6.3.	Statistical Anomaly Detection Techniques	26
2.6.4.	Anomaly Detection using Association Rule Mining	27
2.6.5.	Other Techniques for Anomaly Detection	27
2.7.	Using Episodes for Web Usage Mining	27
2.8.	Suffix Array to Locate the Substring Pattern	30
2.9.	Regular Expressions	31
2.9.1.	Regular Expression Engines	31
2.10.	Discussion.....	32
3.	USING HYBRID CLUSTERING TO IDENTIFY WEBSITE USER ACCESS PATTERNS	33
3.1.	Overview	33
3.2.	Terminology and the Data Model.....	34
3.3.	Preprocessing Engine.....	36
3.3.1.	Implementation of Preprocessing Engine.....	36
3.4.	Hybrid Clustering for Web Usage Mining	39
3.4.1.	Justification for EM+DBSCAN.....	39
3.4.2.	EM+DBSCAN Algorithm Implementation	41
3.5.	The Signature Module	42
4.	EPISODE BASED APPROACH.....	44
4.1.	Overview	44
4.2.	Detecting Slight Changes	44
4.2.1.	Data Models for Detecting Slight Changes.....	44
4.2.2.	Slight Changes between Web User Sessions	46
4.2.3.	Design	46
4.3.	Episode	47
4.4.	Regular Expressions to Represent Episodes.....	49
4.5.	Regular Expression-Based Episode Representation.....	53
4.6.	Episode Clustering	53
4.7.	Summary	54
5.	EVALUATION AND DEMONSTRATION	55

5.1.	Overview	55
5.2.	Data Set for Evaluation	55
5.3.	Evaluation of the EM+DBSCAN Approach	56
5.3.1.	Evaluating Clustering Algorithms	56
5.3.2.	Evaluating Cluster Signature Uniqueness	58
5.3.3.	Evaluation of Effects of Temporal Website Changes	60
5.3.4.	Demonstrating Social Media Impact on Site Access	63
5.3.5.	Attack Detection.....	64
5.4.	Evaluation of the Episode based approach.....	65
5.4.1.	Improving Clustering with Episodes.....	65
5.4.2.	Evaluation of Memory Usage	68
5.4.3.	Identifying Attacks on a Website.....	69
5.4.4.	Common User Patterns	71
6.	Discussion.....	73
6.1.	Contributions	73
6.2.	Usability	74
6.3.	Scalability	74
6.4.	Accuracy	75
7.	CONCLUSION & FUTURE WORK.....	76
	APPENDIX A: SOURCE CODE	86



LIST OF FIGURES

Figure 1.1: Web mining categories	1
Figure 2.1: General Architecture for web usage mining (source [15])	9
Figure 2.2: The web usage mining steps (source [29])	11
Figure 2.3: Data preprocessing phases	11
Figure 2.4: Pattern discovery techniques and methods	16
Figure 2.5: Categorizing patterns for sequential mining algorithms.....	22
Figure 2.6: Sample query of WEBMINER	23
Figure.2.7: Clustering based anomaly detection techniques	25
Figure 2.8: Preprocessing steps with episode identification (source [29])	28
Figure 2.9: Detailed web usage mining process (source [1]).....	29
Figure 2.10: Two types of episodes (Source [29]).....	30
Figure 3.1: Hybrid clustering algorithm approach	33
Figure 3.2: Episode based approach.....	33
Figure 3.3: Log record in access log file	35
Figure 3.4: Functionality of the data preprocessing engine	36
Figure 3.5: sample access log file	37
Figure 3.6: Session file and Mapping file	38
Figure 3.7: Component architecture of the data preprocessing engine	39
Figure 3.8: EM +DBSCAN algorithm	41
Figure 3.9: Cluster matrix with session numbers and page occurrences	43
Figure 3.10: Cluster signature module	43
Figure 4.1: Data model types in web usage mining	45
Figure 4.2: An example of a slight change in web session	46
Figure 4.3: System architecture.....	47
Figure 4.4: Episode structures.....	49
Figure 4.5: Regular expressions generator	50
Figure 4.6: Sample session with page sequence	53
Figure 4.7: Session with episode representation	53
Figure 5.1: Evaluating cluster mechanisms and EM+DBSCAN for website N	58
Figure 5.2: Evaluating cluster mechanisms and EM+DBSCAN for website U	58
Figure 5.3: Evaluating cluster mechanisms and EM+DBSCAN for website F	58

Figure 5.4: Effect of training session addition to the non-profit organization site	60
Figure 5.5: Detecting major changes in the website using EM+DBSCAN clustering	61
Figure 5.6: Detecting major changes in the website using DBSCAN	61
Figure 5.7: Detecting major changes in the website using k-means with domain expert.....	62
Figure 5.8: Detecting major changes in the website using EM.....	62
Figure 5.9: Detecting major changes in the website using K-means with a cluster count of 15	63
Figure 5.10: Impact of social media on the user behavior model	64
Figure 5.11: Generated new cluster that represents the sessions of an attack.....	65
Figure 5.12: Comparing completeness of clustering algorithms with episodes.....	66
Figure 5.13: Intra-cluster distance of clusters (website N)	67
Figure 5.14: Nearest-cluster distance of clusters (website N).....	68
Figure 5.15: Suffix array length growth for the two suffix array versions in website N	69
Figure 5.16: Suffix array length growth for the two suffix array versions in websites U (left) and F (right).....	69
Figure 5.17: (a) Normal user pattern (b) Attacker pattern	70
Figure 5.18: Slight change in cluster	70
Figure 5.19: Sample web session attack	71
Figure 5.20: Search user access pattern	71
Figure 5.21: Article readers access patterns in website N	72



LIST OF TABLES

Table 2.1: Web log preprocessing techniques and algorithms.....	13
Table 3.1: Factors that affect user navigation in a website	40
Table 4.1: Suffix Array on a sample user session.....	51
Table 4.2: Sorted Suffix Array.....	51
Table 4.3: n-grams of the user session.....	51
Table 4.4: Sorted suffix array from n-gram	52
Table 5.1: Dataset for evaluation	56
Table 5.2: User behavior model count by domain experts and the system.....	57
Table 5.3: Cluster distribution and signature uniqueness	59



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF ABBREVIATIONS

API	Application Programme Interface
CEP	Complex event processing
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer protocol
HTTPS	HTTP over TLS
KDD	Knowledge Discovery and Data Mining
W3C	World Wide Web Consortium
WUM	Web Usage Mining



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF APPENDICES

Appendix	Description	Page
Appendix - A	Source Code	86



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk