MODELLING WEBSITE USER BEHAVIOR FROM WEB ACCESS LOGS

Ganihachchi Pathirannehelage Don Madhuka Udantha

(148016F)

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

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.

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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

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

LIST OF APPENDICES