DETECTING ACCESS PATTERNS THROUGH ANALYSIS OF WEB LOGS

Nilani Algiriyage

This dissertation submitted in partial fulfilment of the requirements for the Degree of Master of Science

Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka

March 2015

DECLARATION

I declare that this is my own work and this thesis 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 thesis, 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: .....

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

....
Signature of the supervisor: Date

ABSTRACT

With the evolution of the Internet and continuous growth of the global information infrastructure, the amount of data collected online from transactions and events has been drastically increased. Web server access log files collect substantial data about web visitor access patterns. Data mining techniques can be applied on such data (which is known as Web Mining) to reveal lot of useful information about navigational patterns.

In this research we analyze the patterns of web crawlers and human visitors through web server access log files. The objectives of this research are to detect web crawlers, identify suspicious crawlers, detect Googlebot impersonation and profile human visitors. During human visitor profiling we group similar web visitors into clusters based on their browsing patterns and profile them.

We show that web crawlers can be identified and successfully classified using heuristics. We evaluated our proposed methodology using seven test crawler scenarios. We found that approximately 53.25% of web crawler sessions were from âĂIJknownâĂİ crawlers and 34.16% exhibit suspicious behavior.

We present an effective methodology to detect fake Googlebot crawlers by analyzing web access logs. We propose using Markov chain models to learn profiles of real and fake Googlebots based on their patterns of web resource access sequences. We have calculated log-odds ratios for a given set of crawler sessions and our results show that the higher the log-odds score, the higher the probability that a given sequence comes from the real Googlebot. Experimental results show, at a threshold log-odds score we can distinguish the real Googlebot from the fake.

For the purpose of human visitor profiling, an improved similarity measure is proposed and it is used as the distance measure in an agglomerative hierarchical clustering for a data set from an e-commerce web site. To generate profiles, frequent item set mining is applied over the clusters. Our results show that proper visitor clustering can be achieved with the improved similarity measure.

Keywords: access logs, crawlers, web users, web usage mining

ACKNOWLEDGEMENTS

I would like to thank all the people who gave me a tremendous support in completing my masters research project successfully. My special thanks go to the academic supervisors Prof.Sanath Jayasena and Prof.Gihan Dias for their guidance, ideas, generous support and encouragement throughout the duration of my research.

Then my heartiest gratitude goes to Mr.Kushan Sharma and Mr.Amila Perera for their ideas and support. And also I remember all the staff members at LK Domain Registry and Techcert who gave me support and encouragement.

I would like to thank the Department of Computer Science and Engineering of University of Moratuwa for giving me the opportunity to carry out this research and providing necessary resources.

I greatly appreciate the valuable comments of all members at Research and Development department of LK Domain Registry.

Finally a big thank to my beloved parents, husband and brothers who were always with me providing support and confidence when I most needed it.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
ART2	Modified Adaptive Resonance Theory
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
RST	Rough Set Theory
SOFM	Self-Organizing Feature Map
SOM	Self Organizing Maps
UB	Usage Based
FB	Frequency Based
VTB	Viewing Time Based
VOB	Visiting Order Based
PC	Possible Crawler
RFC	Request For Comment