

# **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 "known" 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

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
1.1	Problem Definition . . . . .	2
1.2	Objectives . . . . .	3
1.3	Methodology . . . . .	3
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>5</b>
2.1	Web Server Access Log Files . . . . .	5
2.2	Web Crawler Detection . . . . .	6
2.2.1	Syntactical Log Analysis . . . . .	7
2.2.2	Traffic Pattern Analysis . . . . .	8
2.2.3	Analytical Learning . . . . .	8
2.2.4	Real-time Web Crawler Detection . . . . .	10
2.2.5	Suspicious Web Crawler Detection . . . . .	11
2.3	Web User Profiling . . . . .	12
2.4	Markov Chain Model . . . . .	15
2.5	Summary . . . . .	16
<b>3</b>	<b>METHODOLOGY</b>	<b>17</b>
3.1	Web Crawler Identification and Characterization . . . . .	17
3.1.1	Data Preparation-Web Crawler Detection . . . . .	17
3.1.2	Identifier . . . . .	19
3.1.3	Classifier . . . . .	21
3.1.4	Summary . . . . .	23
3.2	Detection of Googlebot Impersonation . . . . .	23
3.2.1	Data Preparation-Googlebot Impersonation . . . . .	24
3.2.2	Summary . . . . .	29
3.3	Human Visitor Profiling . . . . .	29
3.3.1	Data Preparation . . . . .	31
3.3.2	Comparison of Visitor Sessions . . . . .	34
3.3.3	Summary . . . . .	36
<b>4</b>	<b>Experimental Evaluation and Discussion</b>	<b>37</b>
4.1	Web Crawler Identification and Characterization . . . . .	37
4.1.1	Data Set . . . . .	37
4.1.2	Methodology Evaluation . . . . .	38
4.1.3	Experimental Results . . . . .	40
4.1.4	Crawler-Trap Tool . . . . .	42

4.2	Googlebot Impersonation . . . . .	43
4.2.1	Patterns of the Dataset . . . . .	43
4.2.2	Markov Chain Models . . . . .	44
4.2.3	Accuracy Evaluation . . . . .	45
4.3	Human Visitor Profiling . . . . .	47
4.3.1	Data Preperation . . . . .	47
4.3.2	Hierarchical Clustering . . . . .	47
4.3.3	User Profiling . . . . .	48
<b>5</b>	<b>CONCLUSIONS</b>	<b>50</b>
5.1	Characterization of Web Crawlers . . . . .	50
5.2	Detection of Googlebot Impersonation . . . . .	50
5.3	Human User Profiling . . . . .	51
5.4	Future Improvements . . . . .	51
5.4.1	Characterization of Web Crawlers . . . . .	51
5.4.2	Detection of Googlebot Impersonation . . . . .	51
5.4.3	Human User Profiling . . . . .	52
	<b>Appendix A Web Crawler Identification &amp; Characterization</b>	<b>57</b>
	<b>Appendix B Crawler-Trap Tool</b>	<b>59</b>
B.1	Introduction . . . . .	59
B.2	Screens . . . . .	59

## LIST OF FIGURES

1.1	PHP remote code execution vulnerability. . . . .	2
2.1	Web robot detection methodology hierarchy.[1] . . . . .	7
2.2	An Example of a CAPTCHA test. . . . .	11
2.3	Web mining taxonomy [2]. . . . .	12
3.1	Summary of web crawler characterization. . . . .	17
3.2	Methodology for identification & characterization of web crawlers. . .	18
3.3	Sessionalized log file example. . . . .	19
3.4	Flow chart of “IDENTIFIER“ module. . . . .	19
3.5	Flow chart of the “CLASSIFIER“ module. . . . .	22
3.6	Methodology for fake Googlebot detection. . . . .	26
3.7	Resource request pattern diagram . . . . .	29
3.8	Methodology for human visitor profiling. . . . .	32
3.9	Web site with navigational paths . . . . .	33
4.1	Number of HTTP requests per day. . . . .	38
4.2	Top most countries generating web crawlers. . . . .	40
4.3	Crawlers by total number of sessions generated. . . . .	41
4.4	HTTP requests per day for real-Googlebot . . . . .	44
4.5	HTTP requests per day for fake-Googlebot . . . . .	45
4.6	Log-odds ratio real Googlebot . . . . .	46
4.7	Log-odds ratio fake Googlebot . . . . .	46
4.8	Results of hierarchical clustering. . . . .	49
B.1	Upload log file. . . . .	59
B.2	Process log file. . . . .	59
B.3	Home page view I. . . . .	60
B.4	Home page view II. . . . .	60
B.5	Crawler analysis report view I. . . . .	60
B.6	Crawler analysis report view II. . . . .	61
B.7	Crawler profile view I. . . . .	61
B.8	Crawler profile view II. . . . .	61
B.9	Crawler profile view III. . . . .	62
B.10	Crawler profile view III. . . . .	62
B.11	IP lookup view I. . . . .	63
B.12	IP lookup view II. . . . .	64

B.13 Crawler list view I. . . . .	65
B.14 Crawler list view II. . . . .	66



## LIST OF TABLES

2.1	Apache combined log format. . . . .	6
2.2	Summary of attributes derived [3]. . . . .	9
3.1	Summary of the data set. . . . .	25
3.2	Real Googlebot “user-agent“ and %hitcount . . . . .	27
3.3	Fake Googlebot “user-agent“ and %hitcount . . . . .	28
3.4	Resource Classes . . . . .	28
3.5	Resource access sequence matrix . . . . .	29
3.6	Session-page matrix . . . . .	33
3.7	Session-time Matrix . . . . .	34
4.1	Summary of the log file. . . . .	37
4.2	Web crawlers found in the dataset. . . . .	39
4.3	Summary of crawler scenarios. . . . .	40
4.4	Summary of crawler sessions. . . . .	41
4.5	“Known“ crawler patterns. . . . .	41
4.6	“Suspicious“ crawler patterns. . . . .	42
4.7	“Other“ crawler patterns. . . . .	42
4.8	Summary of the data set. . . . .	43
4.9	Countries originating fake-Googlebot Academic web site . . . . .	43
4.10	Accuracy scores for different log-odds ratios (e-commerce web log) . . . . .	47
4.11	Summary of the log file. . . . .	47
4.12	Cluster results . . . . .	48
A.1	Web crawlers with originated country. . . . .	57
A.2	Examples for identified “known“ crawlers. . . . .	58
A.3	Examples for identified “suspicious“ crawlers. . . . .	58
A.4	Examples for identified “other“ crawlers. . . . .	58

## 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