Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

MASSEY UNIVERSITY

MAXIMISING THE EFFECTIVENESS OF THREAT RESPONSES USING DATA MINING: A PIRACY CASE STUDY

Seung Jun Lee

Master of Information Sciences

Information Technology

© Copyright 2015

By

Seung Jun Lee

To my

Beloved grandmother and uncle who left me during the writing of this thesis,

I dedicate this thesis to them and may their souls rest in peace…

Dedication also goes to my beloved parents and my wife who have always been

very supportive with much love and providing their sincere support to me

during the entire research of this thesis.

I also dedicate this thesis to my beloved grandfather who left us

a few years ago who always gave lots of support and love to my beloved family…

MAXIMISING THE EFFECTIVENESS OF THREAT RESPONSES USING DATA MINING: A PIRACY CASE STUDY

This thesis is presented in partial fulfillment of the requirements for the degree of

Master of Information Sciences in
Information Technology

School of Engineering and Advanced Technology
At Massey University Albany, Auckland, New Zealand

By

Seung Jun Lee, PG Diploma in Information Sciences and
B Information Sciences (Massey University)

August 2015

Acknowledgements

I would like to thank my supervisor, Professor Paul Watters for his sincere support, clear guidance, encouragement and strong belief in me during the entire process and research of this thesis. It was much privileged to have an invaluable opportunity working on this thesis project with his continuous support and helpful insight throughout the past year.

I am also very grateful to my dearest family, especially my parents who have always been very supportive to me with much love, endless support and patience. I also would like to thank my wife, Sily for providing me with much support, dedication and love to our family. My gratitude also goes to all the staff and faculty members of the School of Engineering and Advanced Technology at Massey University, Albany for their positive support and assistance over the past few years. I also would like to thank all my friends at Massey University who were shared lots of invaluable moments and great ideas that helped me to enhance and build more motivation during the past few years of studies.

Abstract

Companies with limited budgets must decide how best to defend against threats. This thesis presents and develops a robust approach to grouping together threats which present the highest (and lowest) risk, using film piracy as a case study. Techniques like cluster analysis can be used effectively to group together sites based on a wide range of attributes, such as income earned per day and estimated worth. The attributes of high earning and low earning websites could also give some useful insight into policy options which might be effective in reducing earnings by pirate websites. For instance, are all low value sites based in a country with effective internet controls? One of the practical data mining techniques such as a decision tree or classification tree could help rightsholders to interpret these attributes.

The purpose of analysing the data in this thesis was to answer three main research questions in this thesis. It was found that, as predicted, there were two natural clusters of the most complained about sites (high income and low income). This means that rightsholders should focus their efforts and resources on only high income sites, and ignore the others.

It was also found that the main significant factors or key critical variables for separating high-income vs low-income rogue websites included daily page-views, number of internal and external links, social media shares (i.e. social network engagement) and element of the page structure, including HTML page and JavaScript sizes. Further research should investigate why these factors were important in driving website revenue higher. For instance, why is high revenue associated with smaller HTML pages and less JavaScript? Is it because the pages are simply faster to load? A similar pattern is observed with the number of links. These results could form a study looking into what attributes make e-commerce successful more broadly.

It is important to note that this was a preliminary study only looking at the Top 20 rogue websites basically suggested by Google Transparency Report (2015). Whilst these account for the majority of complaints, a different picture may emerge if we analysed more sites, and/or selected them based on different sets of criteria, such the time period, geographic location, content category (software versus movies, for example), and so on. Future research should also extend the clustering technique to other security domains.

Table of Contents

Acknowledgements .... **iv**
Abstract .... **v**
Table of Contents .... **vi**
List of Figures .... **viii**
List of Tables .... **xi**
List of Abbreviations .... **xiii**
1 Introduction **1**
1.1 The Cost of Online Piracy and Cyber Security 1
1.2 Advertising and Risk 5
1.2.1 The Digital Millennium Copyright Act (DMCA) 5
1.2.2 Chilling Effects Database 6
1.2.3 Google Transparency Report 6
1.2.4 Mainstream Advertising and How Piracy is funded 9
1.2.5 High-Risk Advertising and Their Links to Piracy Websites 10
1.2.5.1 High-Risk Advertising: Case Studies in Canada 10
1.2.5.2 High-Risk Advertising: Case Studies in Australia 11
1.2.5.3 High-Risk Advertising: Case Studies in New Zealand 12
1.3 Research Scope and Research Objectives 14
1.4 Research Questions and Hypotheses 14
1.5 Structure of the Thesis 15
1.6 Summary 16
2 Literature Review **17**
2.1 Background of Online Advertising 17
2.1.1 Online Advertising and Behavioral Targeting (BT) 17
2.1.2 Online Advertising and Intrusiveness 19
2.1.3 Online Advertising and Privacy 21
2.1.4 Online Advertising and Obtrusiveness 23
2.1.5 Online Advertising and Economic Factors 25
2.1.5.1 Economic Factor of Online Ad: Pricing of Keywords 27
2.1.5.2 Economic Factor of Online Ad:

The Role of Indirect Network Effects 27
2.1.6 Online Advertising and Social Factors 29
2.1.7 Online Advertising and Positioning 34
2.2 High-Risk Advertising 37
2.3 Summary 42
3 Research Methodology 43
3.1 Background 43
3.2 Data Collection 44
3.3 Cluster Analysis 54
3.3.1 Theoretical Framework of Cluster Analysis 55
3.3.2 A Simple K-means Clustering Algorithm 58
3.3.3 Project Implementation by using K-means Clustering Analysis 60
3.4 Descriptive Statistics 69
3.5 Summary 74
4 Experiments and Results 75
4.1 Descriptive Statistics 75
4.2 Cluster Analysis 83
4.2.1 Income Per Day and Estimated Worth 88
4.2.2 Daily Unique Visitors and Daily Page-views 89
4.2.3 Search Engine Backlinks 89
4.2.4 Website Ranks and Scores 90
4.2.5 Location Latitude and Location Longitude 91
4.2.6 Social Network Engagement 92
4.2.7 Search Engine Indexes 93
4.2.8 Page Resources Breakdown 93
4.2.9 Homepage Links Analysis 95
4.2.10 Online Safety Information 97
4.3 Summary 97
5 Conclusion and Discussion 99
5.1 Conclusion and Discussion 99
5.2 Summary 101
6 Future Work 103
References 104

List of Figures

Figure 1: A trend graph showing URLs requested to be removed from search per week as of 15/08/2015 2
Figure 2: A trend graph showing the total removal requests received from Google in regard to government requests around the world to remove content since 2009 8
Figure 3: A pie-chart distribution of High-Risk Advertisings 11
Figure 4: Perceived intrusiveness depending on various forms of web advertisements 20
Figure 5: Intentions to return or revisit the website containing the various forms of web advertisements 21
Figure 6: An Overall Illustration of Parameter Estimates for Final Structural Model 31
Figure 7: An Overall Structure of Hypotheses Testing and Modified Model in regard to watching OVA 34
Figure 8: An Example of Website Statistics and Website Valuation from CuteStat.com 46
Figure 9: An Example of Website Statistics and Website Valuation from TriPHP 47
Figure 10: A Screenshot of Search Field with specified domain name 48
Figure 11: An Example Output of Domain Statistics and Domain Valuation retrieved from filestube.com as of 28/04/2015 48
Figure 12: An Example of General Clustering 56
Figure 13: An Example of a Simple K-means clustering 59
Figure 14: A Screenshot of the basic raw data file in a Microsoft Excel worksheet 61
Figure 15: A Screenshot of Find and Replace window in Data.xlsx 61
Figure 16: A Screenshot of Replace menu containing the two required input text fields 62
Figure 17: A Screenshot of Replace menu containing the two required input text fields 62
Figure 18: A Screenshot of saving the basic raw data file as CSV (Comma Delimited) type in an Microsoft Excel worksheet environment 63
Figure 19: A Screenshot of WEKA 3.6 Data Mining Program GUI Chooser 63
Figure 20: A Screenshot of opening the main data file “Data.csv” in WEKA 3.6 Explorer 64
Figure 21: A Screenshot of WEKA 3.6 Explorer after opening the main data file 65

Figure 22: A Screenshot of WEKA 3.6 filtering algorithm after implementing and applying this particular function successfully 66
Figure 23: A Screenshot of WEKA 3.6 filtering algorithm in WEKA Explorer 66
Figure 24: A Screenshot of selecting a simple K-means clustering algorithm in WEKA 3.6 Explorer 67
Figure 25: A Screenshot of the clustering output after the implementation of a simple K-means clustering algorithm is successfully processed 68
Figure 26: A Screenshot of SAS Enterprise Miner 13.1 Main Welcome Screen 71
Figure 27: A Screenshot of Opening the Project in SAS Enterprise Miner 13.1...................... 71
Figure 28: A Screenshot of selecting and opening the main project diagram in the Project Panel of SAS Enterprise Miner 13.1 environment 72
Figure 29: A Screenshot of selecting and displaying the option called “Results” from StatExplore Node in SAS Enterprise Miner 13.1 project diagram workspace 73
Figure 30: A Screenshot of selecting and displaying the option called “Results” from StatExplore Node in SAS Enterprise Miner 13.1 project diagram workspace 76
Figure 31: A Screenshot of inserting the function arguments for calculating the third quartile value of an existing variable 80
Figure 32: A bar-graph information of income per day and estimated valuation in terms of the HIC sites and the LIC sites 88
Figure 33: A bar-graph information of daily unique visitors and daily page-views in terms of the HIC sites and the LIC sites 89
Figure 34: A bar-graph information of Google Backlinks, Alexa Backlinks and Bing Backlinks in terms of the HIC sites and the LIC sites 90
Figure 35: A bar-graph information of Google Page-rank and Alexa Rank in terms of the HIC sites and the LIC sites 91
Figure 36: A bar-graph information of Location Latitude and Location Longitude in terms of the HIC sites and the LIC sites 92
Figure 37: A comprehensive bar-graph information of social network engagement in terms of HIC sites and LIC sites 93

Figure 38: A comprehensive bar-graph information of various page resources breakdown in Kilobytes in terms of the HIC sites and the LIC sites 94

Figure 39: A comprehensive bar-graph information of homepage links analysis in terms of the HIC sites and the LIC sites 96

List of Tables

Table 1: A numerical information about the most recent copyright removal requests received for search in the past month as of 15/08/2015 2

Table 2: Frequency by ad category – High-Risk ads 10

Table 3: Frequency by ad category – High-Risk ads 12

Table 4: An overall summary of 9 main hypothesises used in this particular study 30

Table 5: An overall result of relative risks for describing illicit drug me in the exposure and comparison advertisements 40

Table 6: Top 20 Specified Domains of DMCA Reported Complaints by Copyright Owners as of 28/04/2015 45

Table 7: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 49

Table 8: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 49

Table 9: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/201 continued 50

Table 10: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 50

Table 11: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 51

Table 12: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 51

Table 13: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 52

Table 14: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 52

Table 15: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 53

Table 16: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 continued 53

Table 17: A comprehensive information of Interval Variable Summary Statistics generated from the implementation of SAS Enterprise Miner 13.1 project 77

Table 18: The calculated results of the third quartile, the first quartile and the IQR for an existing variable 81

Table 19: Comprehensive Five-Number Summary Statistics and Interquartile Range (IQR) 82

Table 20: Clustering Output of a Simple K-means clustering algorithm for dataset (Data.csv) 83

Table 21: Clustering Output of a Simple K-means clustering algorithm for dataset (Data.csv) continued 84

Table 22: A table information of pre-defined criteria 84

Table 23: Final Clustering Output of a Simple K-means clustering algorithm for dataset (Data.csv) including the results of percentage for Cluster Number 1 over Cluster Number 0 based on the pre-defined criteria....................85

List of Abbreviations

ARFF	Attribute-Relation File Format
ASCII	American Standard Code for Information Interchange
BT	Behavioral Targeting
CDA	Communications Decency Act
COPA	Children Online Protection Act
COPPA	Children Online Privacy Protection Act
CPC	Cost-Per-Click
CPM	Cost-Per-Thousand-Impressions
CSS	Cascading Style Sheet
CSV	Comma Separated Values
CTR	Click-Through-Rate
DMCA	Digital Millennium Copyright Act
DMOZ	Directory Mozilla
EU	European Union
FTC	Federal Trade Commission
GUI	Graphical User Interface
HIC	High Income Cluster
HTML	Hypertext Markup Language
IP	Internet Protocol
IQR	Interquartile Range
IT	Information Technology
LIC	Low Income Cluster
LVF	Lower Visual Field
NA	Not-Applicable
OVA	Online Video Advertising

P/E	Price-To-Earnings ratio
PPA	Pay-Per-Auction
SAS	Statistical Analysis System
SEO	Search Engine Optimisation
TRA	Theory of Reasoned Action
URL	Uniform Resource Locator
US	United States
USC	University of Southern California
UVF	Upper Visual Field
WEKA	Waikato Environment for Knowledge Analysis
WOT	Web of Trust