

LB/DON/106/2016

IT 01/136

Analysing Citizen Profiles with Data Mining

LIBRARY
UNIVERSITY OF MORATUWA, SRI LANKA
UNIVERSITY MORATUWA, SRI LANKA
MORATUWA

W. A. Mohotti

139172C

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology.

April 2016

004"16"
004(043)

University of Moratuwa



TH3171

TH 3171
+ 1 DVD ROM
(TH 3160 - TH3180)

TH3171

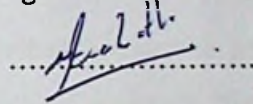
Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student

W. A Mohotti

Signature of Student



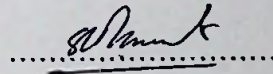
Date: 10 - 04 - 2016

Supervised by

Name of Supervisor

S. C. Premaratne

Signature of Supervisor



Date: 10/04/2016

Dedication

We dedicate the output of this research work and thesis to government policy setters who are trying to uplift lifestyles of Sri Lankans by addressing the citizens' problems. Also we specially dedicate this system to all those who generously contributed their valuable time, advising and helping in doing this research, specially to my supervisor Mr. S.C. Premaratne. In Sri Lanka, area of analyzing citizen profiles is not effectively done with appropriate techniques. It is with this thought in mind that we have done this research. I hope the research and the findings described below will provide a useful insight for analyzing lifestyle data to provide solutions to issues attach with citizens' life patterns.

Acknowledgement

First of all I would like to thank my project supervisor Mr. S.C. Premaratne who spent his valuable time for guiding this research to make it a success. Furthermore, my next big thank goes to Prof. Asoka Karunandha who taught us Research Methodology and Literature Review and thesis writing subjects which were the basis for this research.

Not only that my thanks should go to all the lecturers in M.Sc in Information Technology degree program of Faculty of IT, who gave their hands to sharpen our knowledge and ideas throughout these two years as they were the illumination which lit up our path ways to success.

Apart from the people who were directly involved, many more helped to make this project a success. Department of Census and Statistics contribute to this research by giving their HIES 2012/2013 data. So thank you all for your great support. Finally, I would like to thank all the batch mates of the M.Sc. in IT degree program who gave their valuable feedbacks to improve the results of the research.

Abstract

There is an exponential growth in issues attached with lifestyles of Sri Lankans over the past few decades. These may contribute to low down the life quality within citizens. In Sri Lanka, there are no adequate researches in the field of analyzing lifestyle data. Though there are few researches which have analyzed the causes for the socio-economic problems, such approaches are not capable of handling big data effectively and not efficient in predicting or describing the issues attach with lifestyle.

Hence, the research has been conducted to analyze citizen profiles in effective way to explore different lifestyle issues. It is hypothesized that analyzing citizen profiles can be done through data mining according to the output want to achieve through predictive or descriptive techniques. The solution takes HIES data set as the input and predict the factors attach with a particular lifestyle issue or describe specific lifestyle issue with its associative causes. Having received the input, this approach preprocessed the dataset to remove the anomalies. Then build data models to represent the lifestyle issue by extracting attributes from HIES data set. Then proceed with pattern recognition for the issues. The important patterns recognized through this approach will be useful for government and policy makers to set up appropriate government policies to uplift the life quality of citizen. The overall design of the research consists of two research question, one question used predictive mining based solution and other one is based on descriptive mining. Classification in data mining was used in finding the factors and their relationships that associated with no schooling and dropouts as those were predictive mining tasks. Clustering is used to explore the relationship between chronic diseases and family.

The overall research is designed using WEKA data mining tool and SPSS statistical tool. Finally, the data models build for citizen profile analysis using data mining techniques are evaluated for their performance using measurements such as value for accuracy, error rate, training time, TP rate, FP rate and ROC measurement.

Contents

	Page
DECLARATION	I
DEDICATION	II
ACKNOWLEDGEMENT	III
ABSTRACT	IV
CONTENTS	V
LIST OF FIGURES	IX
LIST OF TABLES	XI
CHAPTER 1 INTRODUCTION	1
1.1. Prolegomena	1
1.2. Background & Motivation	1
1.3. Aims & Objectives	3
1.3.1 Aim	3
1.3.2 Objectives	3
1.4 Proposed Solution	3
1.5 Resource Requirements	5
1.6 Summary	5
CHAPTER 2 STATE OF THE ART OF EXPLORING ISSUES IN CITIZEN PROFILES	6
2.1 Introduction	6
2.2 Lifestyle and household profiling	6
2.2.1 Lifestyle and household profiling in Europe	7
2.2.2 Lifestyle and household profiling in Asia	8
2.2.3 Lifestyle and household profiling in Sri Lanka	8
2.3 Methods for lifestyle data analysis	10
2.3.1 Currently existing methods in the world	11

2.4	Research Question	12
2.5	Summary	12
CHAPTER 3 TECHNOLOGY ADAPTED		14
3.1.	Introduction	14
3.2.	What is Data Mining?	14
3.3.	Reasons for using data mining for citizen profile analysis	16
3.4.	How to use data mining for lifestyle analysis	16
3.5.	Summary	17
CHAPTER 4 A NOVEL APPROACH FOR CITIZEN PROFILING ANALYSIS		18
4.1.	Introduction	18
4.2.	Hypothesis	18
4.3.	Input	18
4.4.	Output	18
4.5.	Process	18
4.2.1.	Data Selection	19
4.2.2.	Data Preprocessing	19
4.2.3.	Data Transformation	20
4.2.4.	Data mining	20
4.2.5.	Evaluation/Interpretation	21
4.6.	Users	21
4.7.	Features	22
4.8.	Summary	22
CHAPTER 5 RESEARCH DESIGN FOR ANALYSING CITIZEN		23
5.1.	Introduction	23
5.2.	Research Design	23
5.3.	Top Level Design	24
5.4.	Detailed Design of the Research	25
5.4.1.	Primary Research Question	26
5.4.2.	Sub Research Question 1	26

5.4.3. Sub Research Question 2	26
5.5. Summary	26
CHAPTER 6 IMPLEMENTATION	27
6.1. Introduction	27
6.2. Solution for Sub Research Question 1:	27
6.2.1. Existing work in this domain	27
6.2.2. Theoretical Framework using SPSS	28
6.2.3. Data Model using WEKA	29
6.2.3.1. Classification as the data mining technique	30
6.2.3.2. Decision Tree for School dropouts and no-schooling	31
6.2.3.3. Bayesian network classifier for School dropouts and no-schooling	32
6.2.3.4. K-nearest neighbours for School dropouts and no-schooling	32
6.3. Solution for Sub Research Question 2:	32
6.3.1. Existing work in this domain	32
6.3.2. Theoretical Framework	33
6.3.3. Data Modeling	33
6.3.3.1. Data Preprocessing	33
6.3.3.2. Clustering as the data mining technique	34
6.3.3.2.1. Clustering using Kmean Algorithm	34
6.3.3.2.2. Clustering using Expectation-Maximization	35
6.3.3.2.3. Clustering using MakeDensityBasedClusterer Algorithm	35
6.4. Summary	36
CHAPTER 7 EVALUATION	37
7.1. Introduction	37
7.2. Evaluation for Classification	37
7.3. Evaluation for Clustering	39
7.4. Summary	40
CHAPTER 8 CONCLUSION AND FURTHER WORK	41
8.1. Introduction	41
8.2. Overview of the research	41
8.3. Problem encountered & limitations	42
8.4. Further work	42
8.5. Summary	43



REFERENCE	44
APPENDIX A DATA PREPROCESSING	48
APPENDIX B ATTRIBUTE SELECTION USING SPSS	49
APPENDIX C PREPROCESSING WITH WEKA	50
APPENDIX D DATA MINING WITH WEKA-CLASSIFICATION	52
APPENDIX E DATA MINING WITH WEKA-CLUSTERING METHODS	61

List of Figures

	Page
Figure 1.1: Steps in Data mining process	4
Figure 2.1: Different data mining techniques used for lifestyle data	12
Figure 3.1: Cross industry standard process for data mining	15
Figure 3.2: Data mining Techniques classification	15
Figure 5.1: Analytical framework for citizen profiling in Sri Lanka	25
Figure 6.1 : Different categories in data models	30
Figure 7.4: Scattered graph of SSE vs. Number of clusters	40
Figure 9.1: Ignore tuples	48
Figure 9.2: data transformation	48
Figure 9.3: Likelihood Ratio Test considering both no-schooling and dropouts	49
Figure 9.4: Model fitting information to represent statistical significance of model	49
Figure 9.5: Preprocessing with filters	50
Figure 9.6: Binning with filters	50
Figure 9.7: Convert class label to nominal for KNN	51
Figure 9.8: Decision Tree for School Dropouts	52
Figure 9.9: Naïve Bayes Classifier for School Dropouts	55
Figure 9.10: K-nearest neighbours for	56

School Dropouts

Figure 9.11: Decision Tree for No schooling	57
Figure 9.12: Naïve Bayes Classifier for No Schooling	59
Figure 9.13: K-nearest neighbours for No schooling	60
Figure 9.14: KMean clustering Results Window	61
Figure 9.15: KMean clustering Visualization	61
Figure 9.16: EM clustering Results Window	62
Figure 9.17: EM clustering Visualization	62
Figure 9.18: MakeDensityBasedClusterer clustering Results Window	63
Figure 9.19: MakeDensityBasedClusterer clustering Visualization	63

List of Tables

	Page
Table 2.1: Comparison of methods for big data analysis	11
Table 2.2: Comparison of existing systems	13
Table 6.1: Attributes selected after multinomial logistic regression	29
Table 7.1: Evaluation measurements for classifiers	37
Table 7.2: comparison of different classification methods to determine school dropout	38
Table 7.3: comparison of different classification methods to determine no-schooling	39
Table 7.5: SSE vs. Number of Clusters	39
Table 7.5: Time taken by clustering algorithms to make clusters for given data set	40