

CONTENT BASED DATA MINING AND ANALYSIS FOR WEATHER RELATED WEB DOCUMENTS

Ms. Thushika Nishatharan
169338 U



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

February 2019

DECLARATION

We declare that, this thesis is our 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 Information derived from the published or unpublished work of others has been acknowledged. Due references have been provided on all supporting literatures and resources.

Thushika Nishatharan

Date:

Signature of Student:

Supervised By:

Mr. S. C. Premaratne

Senior Lecturer

Faculty of Information Technology

University of Moratuwa

Date:

Signature of Supervisor:

ACKNOWLEDGEMENT

I want to acknowledge and express my appreciation to those efforts and support in one way or other contributed to the successful completion of this research.

First, I am deeply grateful for the supervision throughout the one year of my research and the helps received from my supervisor Mr. S. C. Premaratne, Department of Information Technology, Faculty of Information Technology, University of Moratuwa. I have learned so much from our project discussions. He is willing to motivate me to contributed tremendously to do my research.

Besides I would like to thank all academic staff from the Faculty of Information and technology, who shared their vast knowledge throughout these two years by providing me a good environment which influenced a lot to achieve this goal.

It is with great pleasure to thank the University of Moratuwa, Sri Lanka, for all its efforts and facilities that it has taken to contribute towards this postgraduate programme. Especially, I would like to thank my colleagues for their support to complete this research in many ways.

I am as ever, especially indebted to my parents, husband, and brother for their love and support throughout my life to improve my career

ABSTRACT

More than two decades, there is a number of weather-related websites are available which approximately predict the weather and climate. By extracting important data from the websites, a predictive data pattern can be produced to show the next day's weather is with rain or not.

By applying different types of web mining and analyzing techniques those extracted weather-related data can be visualized to a common pattern for weather forecasting with the main deciding factors of weather. With the use of these approaches, reasonably precise forecasts can be made up to about four to five days in advance. For the weather prediction analysis, we need to discover deciding factors of the next day's weather. Particularly, common weather dependent factors and the relationship of the prediction to the particular phenomenon

The solution proposed by this research can be used to analyze a large amount of weather data which are in different forms in each source. By using predictive mining task our solution allows us to make predictions for future instances according to the model what we have created. Evaluation measurements for the selected data mining technique such as accuracy percentage, TP & FP Rate, Precision, F-Measure, ROC area, SSE, and loglikelihood for classification and clustering leads to create a high quality model of prediction.

Knowledge flow interface provides the data flow to show the processing and analyzing data with precise association rules. In order to evaluate the model, SSE values and time to build the model, are considered in an effective manner.

TABLE OF CONTENTS

DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
TABLE OF CONTENTS	iv
ABBREVIATIONS	vii
LIST OF FIGURES	viii
LIST OF TABLES	ix
CHAPTER 01	1
Introduction	1
1.1 Prolegomena	1
1.2 Background & Motivation	1
1.3 Research Problem Statement	3
1.4 Aim and Objective	3
1.4.1 Aim	3
1.4.2 Objectives	3
1.5 Overview of the Report	4
1.6 Summary	4
CHAPTER 02	5
State of the art of exploring issues in weather related web content mining	5
2.1 Introduction	5
2.2 Web mining techniques	5
2.3 Approach for weather forecasting analysis	9
2.4 Summary	9
CHAPTER 03	10
Technology adapted in weather data analysis	10
3.1 Introduction	10
3.2 Involvement of Data Mining	10
3.3 Major reasons for using data mining for weather prediction analysis	12
3.4 Involvement of data mining for weather prediction analysis	13
3.5 Summary	13

CHAPTER 04	14
A novel approach for analyzing weather predictions	14
4.1 Introduction	14
4.2 Hypothesis	14
4.3 Input	14
4.4 Output	14
4.5 Process	15
4.5.1 Data Selection	15
4.5.2 Data Preprocessing	15
4.5.3 Data Mining	16
4.5.5 Evaluation/ Interpretation	16
4.6 Features	16
4.8 Summary	16
Implementation of the solution	17
5.1 Introduction	17
5.2 Weka	17
5.3 Data Collection and Preprocessing	18
5.3.1 Replace Missing Values	19
5.3.2 Data Transformation	19
5.3.3 Data Normalization	19
5.4 Weather Prediction Analysis	20
5.4.1 Using suitable classifier for classification	20
5.4.2 Applying Appropriate Clustering Method	21
5.5 Summary	23
CHAPTER 06	24
Evaluation	24
6.1 Introduction	24
6.2 Evaluation for classification	24
6.3 Evaluation for Clustering	27
6.3 Information gain from the dataset	29
6.4 Generating Association Rules	30
6.5 Summary	30

CHAPTER 07	31
Conclusion and Further Work	31
7.1 Introduction	31
7.2 Overview of the research	31
7.3 Major findings in our research	32
7.4 Limitations	32
7.5 Further Developments	33
7.6 Summary	33
CHAPTER 08	34
REFERENCES	34
APPENDIX A	36
APPENDIX B	38
APPENDIX C	40
APPENDIX D	41
APPENDIX E	43

ABBREVIATIONS

KDD	Knowledge Discovery in Databases
WEKA	Waikato Environment for Knowledge Analysis
GUI	Graphical User Interface
TP	True Positive
FP	False Positive
SSE	Sum of Squared Error
ROC	Receiver Operating Characteristics

LIST OF FIGURES

	Page
Figure 1.1: Web mining categories	2
Figure 2.1: Overview of KDD process	11
Figure 5.1: Weka GUI Chooser	17
Figure 5.2: Weka Explorer	18
Figure 5.3 Visualization of attributes before preprocessing	19
Figure 5.4 Visualization of attributes after preprocessing	20
Figure 6.1 Knowledge flow for the Naïve Bayes classification	26
Figure 6.2 Graph of SSE within clusters versus different clusters	27
Figure 6.2 Graph of log likelihood versus different clusters	28
Figure 6.3 Information Gain Evaluation	29
Figure 6.4 Rules generated from Predictive Apriori algorithm	30
Figure 7.1 Cluster visualization with Rainfall factor	31
Figure 7.2 Graph of SSE values versus ignored attributes	32

LIST OF TABLES

	Page
Table 2.1: Web content mining using different algorithms	7
Table 2.2: Techniques for web content mining	8
Table 5.1: Cluster instances in K-Means Clustering	22
Table 5.2: Cluster instances in EM Clustering	22
Table 5.3: Cluster instances in Density-Based Clustering	23
Table 6.1. Evaluation measurements for classifiers	24
Table 6.2. Comparison of different classification methods for rainy day	25
Table 6.3. Comparison of different classification methods for light rain shower Day	25
Table 6.4. Comparison of different classification methods for no rainy day	25
Table 6.5. Variation of SSE within clusters for different clusters	27
Table 6.6. Variation of log likelihood for different clusters	28
Table 6.7. Time taken for clustering by different clustering algorithms	29