

MST(EC)
087
WANG

Research on Outlier Mining Method to Web Content

by

Pengcheng Wang

Master of Science in E-commerce Technology



2016



**Faculty of Science and Technology
University of Macau**

University of Macau

Abstract

RESEARCH ON OUTLIER MINING METHOD TO WEB
CONTENT

by PengCheng Wang

Thesis Supervisor:

Associate Professor, JingZhi Guo

Master of Science in E-commerce Technology

With the rapid development of cloud computing, big data and Internet technology, mass data has been widely used in the process of work, life and learning, such as E-government, E-commerce, online learning, etc., leading to the birth of a huge amount of data resource information. Mass data resource is a great convenience to people, because the scale of network data resources is increasing. It is becoming more and more complex, and it has brought obstacles to the use of information. Therefore, it is needed to build a fast Web content data mining algorithm, improve the filtering of data noise, more accurate access to valuable information.

In this paper, I analyze the background and significance of outlier data mining, elaborate the contents of data mining, outlier mining, Web outlier mining, and sum up the research status and the innovation of this paper. Secondly, this paper summarizes the concept and definition of outlier data mining, analyzes the application of outlier mining technology, the traditional outlier mining algorithm based on statistics, based on proximity, based on density and clustering based on four categories, and then analyzes the data mining content of outliers. Then, this paper studies and analyzes Web outlier data mining, describes the classification of web data mining, and then summarizes the framework and steps of web data mining. Also, I analyze in detail the existing problems of algorithm based on bottom-up clustering, and put forward a new

dual-path web outlier mining algorithm. Dual-path algorithm can analyze the relation of the data from multiple angles, such as mapping, according to each data that contains relevant information content, then choose a reasonable data analysis and evaluation mechanism. Compared with bottom-up clustering algorithm, dual-path algorithm can make use of the collaborative relationship between various data that obtain outliers more accurate.

Outlier detection can effectively improve the accuracy of web data mining analysis, and further improve the content of web data. It can provide a large amount of valuable information resources and improve the level and ability of people to use data.

Keywords: Data Mining; Outlier Mining; Web Mining; Web Outlier Mining;



Table of Contents

LIST OF FIGURES	III
LIST OF TABLES.....	IV
LIST OF ABBREVIATIONS	V
ACKNOWLEDGMENTS	VI
CHAPTER 1: INTRODUCTION.....	1
1.1 RESEARCH BACKGROUND	1
1.2 RESEARCH MOTIVATION.....	2
1.3 ORGANIZATION OF THE THESISI	3
1.4 LITERATURE REVIEW	4
1.4.1 DATA MINING.....	4
1.4.2 OUTLIER MINING.....	6
1.4.3 WEB MINING.....	7
1.4.4 WEB OUTLIER MINING.....	9
CHAPTER 2: OVERVIEW OF OUTLIER DATA MINING	10
2.1 INTRODUCTION	10
2.2 OUTLIER DETECTION	12
2.3 THE ALGORITHM OF OUTLIER DETECTION	14
2.3.1 STATISTICAL APPROACHES	14
2.3.2 PROXIMITY-BASED APPROACHES.....	16
2.3.3 CLUSTERING-BASE APPROACHES	18
2.3.4 CLASSIFICATION APPROACHES.....	19
2.4 OUTLIER ANALYSIS	20
2.5 CHAPTER SUMMARY	20
CHAPTER 3: FRAMEWORK FOR WEB OUTLIER MINING	21
3.1 INTRODUCTION	21
3.2 THE DATA SOURCE OF WEB MINING	21
3.3 TAXONOMY OF WEB OUTLIERS	22
3.4 WEB OUTLIER MINING FRAMEWORK.....	25
3.4.1 DATA EXTRACTION.....	26
3.4.2 DATA PREPROCESSING.....	27
3.4.3 OUTLIER DATA DETECTION	27
3.4.4 OUTLIER DATA ANALYSIS	28
3.5 CHAPTER SUMMARY	28
CHAPTER 4: OUTLIER DETECTION BASED ON BOTTOM-UP CLUSTERING ALGORITHM.....	29

4.1 DEFINITIONS.....	29
4.2 ALGORITHM DESIGN AND ANALYSIS	34
4.3 EXPERIMENTAL RESULTS AND EVALUATION	36
4.4 CHAPTER SUMMARY	38
CHAPTER 5: DUAL-PATH OUTLIER MINING ALGORITHM.....	39
5.1 EXISTING PROBLEMS OF TRADITIONAL ALGORITHMS	39
5.2 DEFINITIONS RELATED TO THE ALGORITHM	40
5.3 ANALYSIS OF DUAL-PATH OUTLIER DATA MINING ALGORITHM ..	41
5.4 EXPERIMENTAL ANALYSIS AND EVALUATION.....	46
5.5 EXPERIMENTAL RESULTS ANALYSIS	49
5.6 CHAPTER SUMMARY	50
CHAPTER 6: CONCLUSION AND FUTURE WORK.....	51
BIBLIOGRAPHY	53

