Hashing for Large-Scale Structured Data Classification

Lianhua Chi

A thesis submitted for the Degree of
Doctor of Philosophy

Faculty of Engineering and Information Technology
University of Technology, Sydney 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate:

Date:

ACKNOWLEDGEMENTS

I would like to express my earnest thanks to my principal supervisor, Dr. Ling Chen, co-supervisor, Professor Xingquan Zhu, and Dr. Bin Li who have provided tremendous support and guidance for my research. Their comprehensive guidance has covered all aspects of my PhD study, including research methodology, research topic selection, experiments, academic writing skills and thesis writing, and even the sentence structure and formulas. Their critical comments and suggestions have strengthened my study significantly. Their strict academic attitude and respectful personality have benefited my PhD study and will be a great treasure throughout my life. Without their excellent supervision and continuous encouragement, this research could not have been finished on time. Thanks to you all for your kind help.

I am grateful to all members of the centre for Quantum Computation and Intelligent Systems (QCIS) for their careful participation in my presentation and valuable comments for my research. I especially thank Professor Chengqi Zhang, and PhD students Mr Ting Guo, Mr Shirui Pan, Mr Meng Fang, Mr Zhibing Hong, Ms Hongshu Chen and Mr Junyu Xuan, and other students for their contributions and help during my PhD study.

I am grateful to FEIT Travel fund, Vice-Chancellor's Postgraduate Conference Fund, and the ARC Discovery Scholarship.

Last but not least, I would like also to thank my family members. Thanks to my mother and father for their constant encouragement and generous support. Thanks to my husband for his unconditional support.

TABLE OF CONTENTS

CERTIFICATE OF AUTHORSHIP/ORIGINALITY i

ACKNOWLEDGEMENTS ii

TABLE OF CONTENTS iii

LIST OF FIGURES vi

LIST OF TABLES viii

ABSTRACT ix

CHAPTER 1 Introduction 1

1.1 Motivation 1

1.2 Problem Statement and Solutions 2

1.2.1 Problem Statement and Solutions for Graph Structured Data 2

1.2.2 Problem Statement and Solutions for Text Structured Data 5

1.3 Contributions 6

1.4 Thesis structure 8

1.5 Publications 10

CHAPTER 2 Preliminary Concepts and Notations 11

2.1 Definitions 11

2.1.1 Definitions for Graph Structured Data 11

2.1.2 Definitions for Text Structured Data 13

2.2 Notations 15

CHAPTER 3 Literature Review 18

3.1 Hashing 18

3.1.1 Hashing Introduction 18

3.1.2 Hashing Overview 20

3.1.3 Hashing Methods 22

3.2 Large-Scale Structured Data Classification based on Hashing 33

3.2.1 Graph stream Classification based on Hashing 33

3.2.2 Text Classification based on Hashing 35

CHAPTER 4 Discriminative Clique Hashing for Fast Graph Stream Classification 38

4.1 Introduction 38

4.2 Definitions & Method Framework 40

4.2.1 Problem Definition 40

4.2.2 Method Framework 41

4.3 DICH: DIscriminative Clique Hashing 42

4.3.1 Graph Clique Detection 42

4.3.2 Graph Clique Hashing 45

4.3.3 Clique-based Classifier 46

4.4 Experiment 47

4.4.1 Effectiveness Evaluation 48

4.4.2 Efficiency Evaluation 53

4.5 Summary 55

CHAPTER 5 Adaptive hashing for Real-time Graph Stream Classification 56

5.1 Introduction 56

5.2 Definitions & System Overview 59

5.2.1 Problem Definition 59

5.2.2 System Overview 60

5.3 ARC-GS: Adaptive Real-time Classification for Graph Stream 61

5.3.1 Graph Clique Detection 62

5.3.2 Differential Graph Clique Hashing 62

5.3.3 Clique-based Chunk Classifier 64

5.3.4 Weighted Chunk Classifier Ensemble 66

5.4 Experiment 67

5.4.1 Data Sets 68

5.4.2 Effectiveness Evaluation 69

5.4.3 Efficiency Evaluation 83

5.4.4 Concept Drifts 86

5.5 Summary 95

CHAPTER 6 Context-Preserving Hashing for Fast Text Classification 97

6.1 Introduction 97

6.2 Preliminaries & Baselines 99

6.2.1 Preliminaries 99

6.2.2 Baselines..101

6.3 RMH: Recursive Min-wise Hashing 102

6.3.1 Multi-Level Exchangeable Representations 102

6.3.2 Recursive Min-wise Hashing 104

6.3.3 Time Complexity Analysis..107

6.4 Experiment 108

6.4.1 Data Sets 109

6.4.2 Compared Methods 109

6.4.3 Performance Comparison 110

6.4.4 Investigation of Min-hash Size 113

6.5 Summary 115

CHAPTER 7 Conclusions and Further Study 117

7.1 Conclusions 117

7.2 Further study..118

References..120

LIST OF FIGURES

Figure 1.1: Graph classification 3
Figure 1.2: Thesis structure..8
Figure 3.1: A basic hashing example 19
Figure 3.2: The Framework for Categorizing Hashing methods: the contents above the arrows represent the classification rules; the contents in the dotted box represent different classes of hashing methods...22
Figure 3.3: The classification of data-independent hashing methods: the contents above the arrows represent the classification rules; the contents in the dotted box represent different classes of hashing methods 23
Figure 3.4: An example of the Min-Hash 26
Figure 3.5: The classification for data-dependent hashing methods: the contents above the arrows represent the classification rules; the contents in the dotted box represent different classes of hashing methods 30
Figure 4.1: The framework of DICH for graph stream classification 42
Figure 4.2: Clique detection in a compressed graph 43
Figure 4.3: A toy example of frequent and discriminative clique-pattern mining.....................46
Figure 4.4: Effectiveness evaluation w.r.t. $\boldsymbol{\alpha}$ on the DBLP data set (left) and the IBM sensor data set (right), $\boldsymbol{N = 5000}$ and $\boldsymbol{\theta = 0.4}$ 49
Figure 4.5: Effectiveness evaluation w.r.t. $\boldsymbol{\theta}$ on the DBLP data set (left) and the IBM sensor data set (right), $\boldsymbol{N = 5000}$ and $\boldsymbol{\alpha = 0.05}$...50
Figure 4.6: Effectiveness evaluation w.r.t. the edges compression size $\boldsymbol{N}$ on the DBLP data set (left) and the IBM sensor data set (right), $\boldsymbol{\alpha = 0.06}$ and $\boldsymbol{\theta = 0.3}$ 52
Figure 4.7: Efficiency evaluation (1) w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{N = 5000}$ and $\boldsymbol{\theta = 0.4}$ (left); (2) w.r.t. $\boldsymbol{\theta}$, $\boldsymbol{N = 5000}$ and $\boldsymbol{\alpha = 0.05}$ (right); and (3) w.r.t. $\boldsymbol{N}$, $\boldsymbol{\alpha = 0.06}$, $\boldsymbol{\theta = 0.3}$ (down), on the DBLP data set 54
Figure 5.1: The framework of the proposed adaptive real-time hashing for graph stream classification method 60
Figure 5.2: Classification accuracy on the IBM (ensemble size *K* = 4), and the CNS (ensemble size *K* = 6) with different numbers of features *M* 74
Figure 5.3: Average accuracy on the IBM (left, ensemble size *K* = 4), and the CNS (right, ensemble size *K* = 6) with different numbers of features *M* 75

Figure 5.4: Classification accuracy on the IBM (number of features M = 1000), and the CNS (number of features M = 10000) with different ensemble size K..79
Figure 5.5: Average accuracy on the IBM (left, number of features M = 1000), and the CNS (right, number of features M = 10000) with different ensemble size K...80
Figure 5.6: Classification accuracy and average classification accuracy on the IBM (upper row, number of features M = 1000 and ensemble size K = 4), and the CNS (bottom row, number of features M = 10000 and ensemble size K = 6) with different hash ratio R..................................82
Figure 5.7: Average time on the IBM (left, ensemble size K = 4), and the CNS (right, ensemble size K = 6) with different numbers of features M...84
Figure 5.8: Average time on the IBM (left, number of features M = 1000), and the CNS (right, number of features M = 10000) with different ensemble size K...85
Figure 5.9: Classification accuracy on the IBM (ensemble size K = 4) and the GTGraph (ensemble size K = 4) with different numbers of features M...91
Figure 5.10: Classification accuracy on the IBM (number of features M = 1000) and the GTGraph (number of features M = 5000) with different ensemble size K..95
Figure 6.1: Motivation examples. The standard min-wise hashing on bags-of-words (flat-sets) gives sim (1, 2) > sim (2, 3) while our RMH on nested bags-of-words (nested-sets) gives sim (2, 3) > sim (1, 2)..98
Figure 6.2: An illustration of the proposed Recursive Min-wise Hashing (RMH) algorithm on a nested set...107
Figure 6.3: Classification accuracy and CPU time of the compared methods w.r.t. the length of output fingerprints..112
Figure 6.4: Classification accuracy and CPU time of the RMH algorithm with different min-hash sizes at different levels of the nested sets..115

LIST OF TABLES

Table 2.1: Notations used in the thesis..15

ABSTRACT

With the rapid development of the information society and the wide applications of networks, almost incredibly large numbers bytes of data are generated every day from the social networks, business transactions and so on. In such cases, hashing technology, if done successfully, would greatly improve the performance of data management. The goal of this thesis is to develop hashing methods for large-scale structured data classification.

First of all, this work focuses on categorizing and reviewing the current progress on hashing from a data classification perspective.

Secondly, new hashing schemes are proposed by considering different data characteristics and challenges, respectively. Due to the popularity and importance of graph and text data, this research mainly focuses on these two kinds of structured data:

1) The first method is a fast graph stream classification method using Discriminative Clique Hashing (DICH). The main idea is to employ a fast algorithm to decompose a compressed graph into a number of cliques to sequentially extract clique-patterns over the graph stream as features. Two random hashing schemes are employed to compress the original edge set of the graph stream and map the unlimitedly increasing clique-patterns onto a fixed-size feature space, respectively. DICH essentially speeds up the discriminative clique-pattern mining process and solves the unlimited clique-pattern expanding problem in graph stream mining;

2) The second method is an adaptive hashing for real-time graph stream classification (ARC-GS) based on DICH. In order to adapt to the concept drifts of the graph stream, we partition the whole graph stream into consecutive graph chunks. A differential hashing scheme is used to map unlimited increasing features (cliques) onto a fixed-size feature

space. At the final stage, a chunk level weighting mechanism is used to form an ensemble classifier for graph stream classification. Experiments demonstrate that our method significantly outperforms existing methods;

3) The last method is a Recursive Min-wise Hashing (RMH) for text structure. In this method, this study aims to quickly compute similarities between texts while also preserving context information. To take into account semantic hierarchy, this study considers a notion of “multi-level exchangeability”, and employs a nested-set to represent a multi-level exchangeable object. To fingerprint nested-sets for fast comparison, Recursive Min-wise Hashing (RMH) algorithm is proposed at the same computational cost of the standard min-wise hashing algorithm. Theoretical study and bound analysis confirm that RMH is a highly-concentrated estimator.