Big Data Behaviour Modelling and Visual Analytics

Jinson Zhang

A thesis submitted for the degree of

Doctor of Philosophy in Computing Sciences

Supervisor by: A/Professor Mao Lin Huang

May 2017

Faculty of Engineering & Information Technology

University of Technology Sydney

Australia

Certificate of Authorship/Originality

Data: **October 2016**

Author: **Jinson Zhang**

Title: **Big Data Behaviour Modelling and Visual Analytics**

Degree: **Doctor of Philosophy in Computer Sciences**

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 STUDENT

_____

Acknowledgements

Firstly, I would like to thank my supervisor A/Professor Mao Lin Huang for giving me the opportunity to do my PhD within the Faculty of Engineering & Information Technology at the University of Technology Sydney, and for his enormous encouragement, advice, and academic guidance during my part-time research period. Dr. Mao Lin have given me not only academic advice and freedom in my research area, but had also provided me with teaching opportunities over the last 15 years across a broad range of different subjects, including in Internet Programming and Data Visualization.

Secondly, I would like to thank our Visual Analytics Team, consisting of Dr. Jie Liang, Dr. Jie Hua, Dr. Liang Fu Lu, and Dr. Wen Bo Wang, for their knowledge sharing and information exchange. Their focus and patience have encouraged me to finish my work, and I have greatly increased my skills and visualization techniques during our fortnightly team meetings. Their humor and trust have benefited me not only through academic partnerships, but also through friendships outside of research.

Thirdly, I would like to thank the Library for their support during my research period. As a full-time Library staff and part-time researcher, I have been given the ability to work flexibly in order in allowing me to attend conferences and other research/teaching activities. Without their support, I would not have been able to finish my research. The Library has also provided me with enormous online materials, and has helped me with referencing works.

Furthermore, I would like to thank the School of Software, Faculty of Engineering & Information Technology, and the Graduate Research School for their help in providing the Faculty Travel Fund and Vice-Chancellor's Conference Fund to attend national and

international conferences. This has been enormously beneficial for part-time PhD researchers such as myself.

Finally, I would like to take this opportunity to thank my wife Lisa and son Edmund for their enormous help and support. Having started my PhD as part-time researcher, their understanding and belief in my capability has provided enormous encouragement during my research period. They have provided continuous support and help in making this PhD thesis possible, and I am thankful for having my family behind me every step of the way.

Table Contents

List of Figures

List of Tables

List of Equations

Abstract

Big Data is composed of text, images, video, audio, mobile or other forms of data collected from multiple datasets, and is rapidly growing in both size and complexity. This has created a huge volume of multidimensional data within a very short time period. Big Data is therefore too big, too complex and moves too fast for us to analyze using traditional methods. Big Data behaviour is considered as a set of concepts and categories that descripts Big Data's acts towards others. The challenges facing Big Data analysis and visualization include: 1) how to classify Big Data across multiple datasets and different forms of data, 2) how to visualize structured and unstructured Big Data behaviour patterns for multidimensional data, 3) how to display Big Data behaviour patterns with very large volumes onto a normal-sized screen, 4) how to visualize Big Data behaviour patterns without the loss of information.

Big Data visualization normally requires optimized solutions through using different visual techniques for integrating display and exploration. To illustrate the huge amount of multidimensional data within a standard-size screen, visualization needs to find an efficient classification method for multiple datasets across any form of data. The current data interactive exploration has normally optimized data for visualization by excluding some pieces of information, resulting in missing information. Big Data visualization also suffers from visual cluttering and data overcrowding problems, whilst dealing with huge amounts of multidimensional data.

My approach includes two parts: Big Data behaviour modelling and Big Data visualization. I have firstly established the 5Ws dimensions for Big Data classification, based on data behaviour ontologies, that can be applied to multiple datasets and to any

form of data. Each data incident contains these 5Ws dimensions, which are posed as a set of concepts and categories that descripts Big Data acts for; **When** did the data occur, **Where** did the data come from, **What** did the data contain, **How** was the data transferred, **Why** did the data occur, and **Who** received the data. Secondly, I have introduced Pair-Density algorithms to measure Big Data behaviour patterns, which enables comparison and analysis between any two dimensions of behaviours. Two non-dimensional axes in parallel coordinates have then been created by using Pair-Density to measure and compare visual patterns for Big Data visualization. Finally, Shrunk Attributes has been deployed into Pair-Density parallel coordinates. This not only narrows down Big Data patterns for better understanding, but also dramatically reduces data cluttering and overcrowding in Big Data visualization.

Three different datasets with a combined total of more than 2.5 million data incidents have been implemented for measuring and visualizing different data patterns, including both numerical and non-numerical dimensions. The experimental results have shown that my new approach has significantly improved the accuracy of Big Data visualization, reduced data cluttering by more than 80% without the loss of information. The use of 5Ws dimensions and Pair-Density parallel coordinates therefore has large potential benefits and applications across both the business and research fields.

This thesis contains the research approach and implementation results obtained by the author during his Ph.D period. The majority of methods and results have been published in **Seventeen** research papers in journals and conference proceeding by May 2016.

Author's Seventeen published papers for PhD

1). J. Zhang, ML. Huang, (2016), "2D Approach Measuring Multidimensional Data Pattern in Big Data Visualization", IEEE first International Conference on Big Data Analysis (ICBDA2016), In Proceeding of 2016 IEEE International Conference on Big Data Analysis, IEEE Computer Society, pp. 194-199, March 2016, DOI: 1109/ICBDA.2016.7509823

2). LF Lu, ML. Huang, and J. Zhang, (2016), "Two Re-ordering Methods in Parallel Coordinates Plots", Journal of Visual Languages and Computing, Elsevier, Vol. 33, pp. 3-12, April 2016, DOI: 10.1016/j.jvlc.2015.12.001

3). J. Zhang and M.L. Huang, (2016), "Data Behaviours Model for Big Data Visual Analytics", International Journal of Big Data Intelligence, InderScience Publishers, Vol. 3, No. 1, pp. 1-17, Dec 2015, DOI: 10.1504/IJBDI.2016.073899

4). J. Zhang, M.L. Huang, and Z. Meng, (2015), "Visual Analytics for BigData Variety and Its Behaviours", International Journal of Computer Science and Information Systems, ComSIS Consortium, Vol. 12, No. 4, pp. 1171-1191, Nov 2015, DOI: 10.2298/CSIS141122050Z

5). WB. Wang, M.L. Huang, J. Zhang, and W. Lai, (2015), "Detecting Criminal Relationships Through SOM Visual Analytics", 19th International Conference on Information Visualization (IV2015), In Proceeding of 19th International Conference on Information Visualization, IEEE Computer Society, pp. 316-321, July 2015, DOI:10.1109/IV.2015.62

6). J. Zhang and M.L. Huang, (2015), "A New Analytics Model for Large Scale Multidimensional Data Visualization", Cloud Computing and Big Data (CloudCom-Asia 2015), Lecture Notes in Computer Science, Vol. 9106, pp. 55-71, June 2015, DOI: 10.1007/978-3-319-28430-9_5

7). J. Zhang, M.L. Huang, W.B. Wang, L.F. Lu, and Z.P. Meng, (2014), "Big Data Density Analytics using Parallel Coordinate Visualization", The 13th

International Symposium on Pervasive System, Algorithm, and Network (I-SPAN2014). In Proceeding of IEEE 17th International Conference on Computational Science and Engineering (CSE), IEEE Computer Society, pp. 1115-1120, Dec 2014, DOI:10.1109/CSE.2014.219

8). W.B. Wang, M.L. Huang, L.F. Lu, and J. Zhang, (2014), "Improving Performance of Forensics Investigation with Parallel Coordinates Visual Analytics", The 8th International Conference on Frontier of Computer Science and Technology (FCST2014), In Proceeding of IEEE 17th International Conference on Computational Science and Engineering (CSE), IEEE Computer Society, pp.1838-1843, Dec 2014, DOI:10.1109/CSE.2014.337

9). J. Zhang, M.L. Huang, and Z. Meng, (2014), "BigData Visualization: Parallel Coordinates using Density Approach", The 2nd International Conference on System and Informatics (ICSAI2014), In Proceeding of IEEE 2nd International Conference on System and Informatics, IEEE Computer Society, pp. 1056-1063, Nov 2014, DOI:10.1109/ICSAI.2014.7009441

10). J. Zhang and M.L. Huang, (2014), "Density approach: a new model for BigData analysis and visualization", Concurrency and Computation: Practice and Experience, Vol. 28, Issue. 3, pp 661-673, First publish online July 2014, DOI: 10.1002/cpe.3337

11). J. Zhang and M.L. Huang, (2013), "5Ws Model for BigData Analysis and Visualization", The 2nd International Conference on Big Data Science and Engineering (BDSE2013), In Proceeding of IEEE 16th International Conference on Computational Science and Engineering (CSE), IEEE Computer Society, pp. 1021-1028, Dec 2013, DOI:10.1109/CSE.2013.149

12). J. Zhang and M.L. Huang, (2013), "Detecting Flood Attack through New Density-Pattern Based Approach", The 15th International Conference on High Performance Computing and Communication (HPCC2013), In Proceeding of 15th IEEE International Conference on High Performance Computing and Communication, IEEE Computer Society, pp. 246-253, Nov 2013, DOI: 10.1109/HPCC.and.EUC.2013.44

13). M.L. Huang and J. Zhang, (2013), "Visual Analysis and Detection of Network Flood Attack through Two-Layer Density Approach", The 3rd International Conference on Computer Science and Network Technology (ICCSNT2013), In Proceeding of 3rd IEEE International Conference on Computer Science and Network Technology, IEEE Computer Society, pp. 625-629, Oct 2013, DOI: 10.1109/ICCSNT.2013.6967191

14). J. Zhang and M.L. Huang, (2013), "Visual Analytics Model for Intrusion Detection in Flood Attack", The 12th International Conference on Trust Security and Privacy in Computing and Communications (TrustCom2013), In Proceeding of 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, IEEE Computer Society, pp. 277-284, July 2013, DOI:10.1109/TrustCom.2013.38

15). J. Zhang, M.L. Huang, and D.B. Hoang, (2013), "Visual Analytics for Intrusion Detection in Spam Emails", International Journal of Grid and Utility Computing, InderScience Enterprises, Vol. 4, No. 2/3, pp. 178-186, Jan 2013, DOI:10.1504/IJGUC.2013.056254

16). J. Zhang, M.L, Huang, and D.B. Hoang, (2011), "Detecting DDoS Attack in Spam Emails using Density-Weight Model", The 2nd International Conference on Information Theory and Information Security, In Proceeding of 2nd IEEE International Conference on Information theory and Information Security, IEEE Press, Vol. II, pp. 344-352, Nov 2011, https://opus.lib.uts.edu.au/research/handle/10453/29567

17). M.L. Huang, J. Zhang, Q. Nguyen, and J. Wang, (2011), "Visual Clustering of Spam Emails for DDoS Analysis", The 15th International Conference on Information Visualization (IV2011), In Proceeding of 15th IEEE international Conference on Information Visualization, IEEE Computer Society, pp. 65-72, July 2011, DOI:10.1109/IV.2011.41