Tree Model Guided (TMG) Enumeration as the Basis for Mining Frequent Patterns from XML Documents

Henry Tan Setiawan

9/28/2007

A Thesis presented for the degree of Doctor of Philosophy

Faculty of Information Technology
University of Technology, Sydney
Australia

Curriculum Vitae

Henry Tan was born in a small town, Sukabumi, Indonesia, on December 7th, 1979. He obtained his Bachelor of Computer System Engineering with first class honour from La Trobe University, VIC, Australia in 2003. During his undergraduate study, he was nominated as the most outstanding Honours Student in Computer Science. Additionally, he was the holder of 2003 ACS Student Award. After he finished his Honour year at La Trobe University, on August 2003, he continued his study pursuing his doctorate degree at UTS under supervision Prof. Tharam S. Dillon. His research interests include Data Mining, Computer Graphics, Game Programming, Neural Network, AI, and Software Development. On January 2006 he took the job offer from Microsoft Redmond, USA as a Software Design Engineer (SDE).

Dedications

My study and this thesis is about learning and presenting the truth. Therefore, first I would like to dedicate the thesis to the λογος through which the truth is manifested. I thank my wife Theresia Liu for consistently supporting me during the hard work and struggles of this thesis. Last but not least, I devote the thesis to my parents, Tan Djoe Jin and Toe Oey Jin who have been instrumental in guiding my life and encouraging me to succeed.

Certificate of Authorship

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

Production Note:
Signature removed prior to publication.

Acknowledgements

I would like to acknowledge a number of inspirational people for all the help, assistance, support and encouragement I have received during my PhD candidature.

Firstly and most especially, I would like to thank Prof. Tharam S. Dillon who has been the great mentor, teacher, and supervisor throughout my research life since my early years at La Trobe University completing my Bachelor of Computer Systems Engineering, Honours. Your guidance on work related issues, as well as your wise advice on matters of life and career, went beyond my expectations of a supervisor. Your time, attention and dedication to the research works carried out throughout my PhD have been outstanding and invaluable.

I would like to thank my wife who has been by my side throughout all the struggles and up and down cycles of my PhD research, also for presenting me with my cherished and much loved daughter, Enrica Tan, who was the main reason that this PhD thesis is completed. I would like to thank my parents who have given me all the necessary financial support, advice and moral guidance throughout my life and my study. Furthermore, I would like to thank my great colleague, Fedja Hadzic. He has been a great partner with a great, positive attitude and dedication through whom most of the important works in this thesis were conceived and published. Thank you to all of my colleagues from the Exel Research group for sharing great times. Last but not least, I would like to thank Dr. Ling Feng for her feedback during the initial stages of my study. Special thanks to Dr. Yun Chi who has always been very helpful providing the source code and related materials. Thanks also to Dr. Mohammed Zaki who has been very helpful providing the source code and answering a lot of questions throughout our discussions.

List of Publications

1. Tan, H, Dillon, TS, Feng, L, Chang, E & Hadzic, F 2005, 'X3-Miner: Mining patterns from XML database', in A Zanasi, CA Brebbia & NFF Ebecken (eds), *Proceedings of the 6th International Conference on Data Mining* (*Data Mining'05*), Skiathos, Greece, WIT Press, pp. 287-297.

2. Tan, H, Dillon, TS, Hadzic, F, Feng, L & Chang, E 2005, 'MB3-Miner: Mining eMBedded subTREEs using tree model guided candidate generation', *Proceedings of the 1st International Workshop on Mining Complex Data (MCD'05)*, Houston, TX, USA, pp. 103-110.

3. Tan, H, Dillon, TS, Hadzic, F, Chang, E & Feng, L 2006, 'IMB3-Miner: Mining induced/embedded subtrees by constraining the level of embedding', In WK Ng, M Kitsuregawa & J Li (eds), Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'06), Singapore, pp. 450-461.

4. Tan, H, Hadzic, F, Dillon, TS & Chang, E 2008, 'State of the art of data mining of tree structured information', *Computer System Science and Engineering, vol. 23, no. 4, July 2008 (pending publication).*

5. Tan, H, Dillon, TS & Hadzic, F 2006, 'Razor: Distance constrained mining of embedded subtrees', in Tsumota & Shusaku (eds), *Proceedings of the International Conference on Data Mining (ICDM'06)*, Hongkong, pp. 8-13.

6. Tan, H, Dillon, TS, Hadzic, F, Feng, L & Chang, E 2007, 'Tree model guided candidate generation for mining frequent subtrees from XML', Publication pending in *Transactions on Knowledge Discovery from Data (TKDD).*

7. Tan, H, Dillon, TS, Hadzic, F, Chang, E & Feng, L 2007, 'Mining induced /embedded subtrees using the level of embedding constraint', submitted to *Fundamenta Informaticae*.

8. Tan, H, Dillon, TS, Hadzic, F & Chang, E 2006, 'SEQUEST: Mining frequent subsequences using DMA strips', in A Zanasi, CA Brebbia & NFF Ebecken (eds), *Proceedings of the 7th International Conference on Data Mining and Information Engineering (Data Mining'06)*, Prague, Czech Republic, WIT Press, pp. 315-328.

9. Hadzic, F, Dillon, TS, Sidhu, AS, Chang, E & Tan, H 2006, 'Mining substructures in protein data', *Proceedings of the 6th International Conference on Data Mining Workshop* (*ICDMW'06*) - *Invited*, Hong Kong, pp. 213-217.

10. Hadzic, F, Tan, H & Dillon, TS 2007, 'UNI3 - efficient algorithm for mining unordered induced subtrees using TMG candidate generation', *Proceedings of the Computational Intelligence and Data Mining (CIDM'07)*, Hawaii, USA, pp. 568-575.

11. Hadzic, F, Tan, H, Dillon, TS & Chang, E 2007, 'U3: Unordered subtree mining using TMG candidate generation and the level of embedding constraint', submitted to ECML PKDD 2008.

12. Hadzic, F, Tan, H, Dillon, TS & Chang, E 2007, 'Implications of frequent subtree mining using hybrid support definition', in A Zanasi, CA Brebbia & NFF Ebecken (eds), *Proceedings of the 8th International Conference on Data Mining & Information Engineering (Data Mining'07)*, The New Forest, UK, WIT Press, pp. 13-24.

13. Hadzic, F, Dillon, TS & Tan, H 2007, 'Outlier detection strategy using the self-organizing map', in X Zhu & I Davidson (eds), *Knowledge Discovery and Data Mining: Challenges and Realities, Information Science Reference*, Hershey, PA, USA, pp. 224-243.

14. Hadzic, F, Dillon, TS, Tan, H, Feng, L & Chang, E 2007, 'Mining frequent patterns using self-organizing map', in D Taniar (ed.), *Research and Trends in*

Data Mining Technologies and Applications: Advances in Data Warehousing and Mining, IGI Global, Hershey, PA, USA, pp. 121-135.

15. Sidhu, AS, Dillon, TS & Setiawan, H 2004, 'XML based semantic protein map', in A Zanasi, NFF Ebecken & CA Brebbia (eds), *Proceedings of 5^{th} International Conference on Data Mining, Text Mining and their Business Applications (Data Mining'04)*, Malaga, Spain, WIT Press, pp. 51-60.

16. Sidhu, AS, Dillon, TS & Setiawan, H 2004, 'Comprehensive protein database representation', in A Gramada & PE Bourne (eds), *Proceedings of the 8^{th} International Conference on Research in Computational Biology (RECOMB'04)*, ACM Press, San Diego, CA, USA, pp. 427-429.

17. Sidhu, AS, Dillon, TS, Sidhu, BS & Setiawan, H 2004, 'Protein knowledge meta model', *Molecular & Cellular Proteomics*, pp. 262-263.

Table of Contents

Table of Figures

Abstract

Association mining consists of two important problems, namely frequent patterns discovery and rule construction. The former task is considered to be a more challenging problem to solve. Because of its importance and application in a number of data mining tasks, it has become the focus of many studies. A substantial amount of research has gone into the development of efficient algorithms for mining patterns from large structured or relational data. Compared with the fruitful achievements in mining structured data, mining in the semi-structured world still remains at a preliminary stage. The most popular representative of the semi-structured data is XML. Mining frequent patterns from XML poses more challenges in comparison to mining frequent patterns from relational data because XML is a tree-structured data and has an ordered data context. Moreover, XML data in general is larger in data size due to richer contents and more meta-data. Dealing with XML, thus involves greater unprecedented complexity in comparison to mining relational data. Mining frequent patterns from XML can be recast as mining frequent tree structures from a database of XML documents. The increase of XML data and the need for mining semi-structured data has sparked a lot of interest in finding frequent rooted trees in forests.

In this thesis, we aim to develop a framework to mine frequent patterns from XML documents. The framework utilizes a structure-guided enumeration approach, *Tree Model Guided* (*TMG*), for efficient enumeration of tree structure and it makes use of novel structures for fast enumeration and frequency counting. By utilizing a novel array-based structure, an *embedded list* (*EL*), the framework offers a simple sequence-like tree enumeration technique. The effectiveness and extendibility of the framework is demonstrated in that it can be utilized not only for enumerating ordered subtrees but also for enumerating unordered subtrees and subsequences. Furthermore, the framework tackles the unprecedented complexity in mining frequent tree-structured patterns by generating only valid candidates with non-zero frequency count and employing a constraint-driven approach. Our experimental studies comparing the proposed framework with the state-of-the-art algorithms demonstrate the effectiveness and the efficiency of the proposed framework.