UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Efficient Query Processing and Analytics on High Dimensional Data

by

Wanqi Liu

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2020

Certificate of Authorship/Originality

I, Wanqi Liu declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Candidate signature

Production Note:
Signature removed prior to publication.

Date: 28/03/2020

ABSTRACT

Efficient Query Processing and Analytics on High Dimensional Data

by
Wanqi Liu

As a fundamental problem in query processing, similarity search has been applied in many fields including multimedia, machine learning, database, recommendation systems and so on. Generally, it will be challengeable when it comes to the high-dimensional space due to "the curse of dimensionality". Since it would be too expensive to find exact results, approximate solutions have been studied in many papers. There are various distance metrics to evaluate the similarity between the query object and other points in a dataset. In this thesis, we focus on some well-known distance metrics including Euclidean distance and inner product. Except for Euclidean space, we also study query processing on graphs and propose a novel distance metric on graph. The thesis contains four similarity search problems regarding to different distance metrics, which are Approximate Nearest Neighbour, Approximate Inner Product Search, Approximate Furthest Neighbour and Skyline Nearest Neighbour. Given a query point, the four problems all focus on retrieving a set of "similar" points from the dataset.

Given a set of d-dimensional data points, and a query point q, Approximate Nearest Neighbour Search (ANNS) aims to find the approximate closest object to q in the set. More specifically, in this thesis, we focus on c-ANNS problem, which means given a constant c, the purpose is to find a result whose distance is not larger than c times of the exact smallest distance with a certain possibility. Even though this problem has been researched for a long time, there are still some shortage of current algorithms. We studied the existed works, and proposed a novel I/O efficient algorithm to solve c-approximate nearest neighbour problem in external memory, which can dramatically reduce I/O cost and provide rigorous proof of its correctness.

Maximum Inner Product Search (MIPS) is another valuable problem. It returns an object with maximum inner product value to query point q. There are hundreds of solutions for MIPS but still short of comprehensive evaluation and analysis of these methods' performance. In this thesis, we chose several state-of-the-art algorithms of MIPS using different techniques, and conducted a set of comprehensive experiments to evaluate their performance fairly.

Approximate Furthest Neighbour Search is an opposite problem of Nearest Neigh-

bour Search. It finds the furthest object to query point q in a dataset instead of the closest one. Since most recent works for approximate furthest neighbour search in external memory are only suitable for low-dimensional data, we proposed a new I/O efficient technique to achieve a better performance on I/O cost.

In addition to Euclidian space, similarity search is also a fundamental problem in other spaces like graphs. Considering real-world applications, the multi-layer graph model is extensively studied to reveal the multi-dimensional relations between the graph entities. In this thesis, we formulated a new problem called skyline nearest neighbor search on multi-layer graphs, and proposed a baseline algorithm, and two optimizations instead of naively adopting the traditional skyline procedure as a subroutine. We also investigated the rule to optimize search order in the algorithm so that further improve the algorithmic efficiency.

Acknowledgements

Firstly, I wish to express my deepest gratitude to my supervisor Prof. Ying Zhang who is professional, diligent and patient, for offering the opportunity to study at UTS, and for the continuous support, patient guidance and enthusiastic encouragement he has provided through the three years. Discussions with him always enlighten me to improve my works in this thesis. He also helped me to extend my knowledge and gave me many inspirations. It would never have been possible for me to complete this work without his incredible support and valuable ideas.

Secondly, I wish to express my sincere appreciation to Prof. Wei Wang for his advice, and for his brilliant ideas which inspired me deeply. Additionally, his valuable comments from his abundant research experience helped me on my research work. Acknowledgement also goes to Prof. Xumin Lin and my co-supervisor Dr. Lu Qin for their supporting and guidance.

I am also sincerely grateful to Mr.Hanchen Wang as most of the works in this thesis are conducted collaborating with him. I really appreciate for his patient to assist me in proofreading and revising the drafts. He is always able to find out grammar mistakes which I missed. My sincere thanks also goes to Dr. Dong Wen who shows advanced research skills and helps me to refine my work.

I would also like to thank Mr. Kai Wang, Mr. Mingjie Li, Mr. Bohua Yang, Ms. Conggai Li, Mr. Boge Liu, Mr. Wentao Li, Dr. Yang Yang, Dr. Haida Zhang, Dr. Fan Zhang, Dr. Ouyang Dian, Mr. Junda Lu, Mr. Yufei Wang for giving me a pleasant time when working with them.

Last but not least, I would like to express my great appreciation to my parents for raising me with care and love, providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis, and to my grandparents for their understanding and love. Thanks to my boyfriend Chenhao Pan for his love, understanding and continuing support to complete the research work.

Wanqi Liu
Sydney, Australia, March 2020.

List of Publications

- **Wanqi Liu**, Hanchen Wang, Ying Zhang, Wei Wang and Lu Qin, I-LSH: I/O efficient c-Approximate Nearest Neighbor Search in High-dimensional Space, published in ICDE 2019
- **Wanqi Liu**, Dong Wen, Hanchen Wang, Fan Zhang and Xubo Wang, Skyline Nearest Neighbor Search on Multi-Layer Graphs, published in ICDE 2019 workshop
- **Wanqi Liu**, Hanchen Wang, Ying Zhang, Luqin and Wenjie Zhang, I/O efficient algorithm for c-approximate furthest neighbor search in high-dimensional space, will appear in DASFAA 2020
- **Wanqi Liu**, Hanchen Wang, Ying Zhang, Wei Wang, Lu Qin and Xuemin Lin, EI-LSH: An Early-Termination Driven I/O Efficient Incremental c-Approximate Nearest Neighbor Search, Received by VLDBJ

Contents

List of Figures

Abbreviation

NN: Nearest Neighbour
ANN: Approximate Nearest Neighbour
c-ANN: c-Approximate Nearest Neighbour
MIPS: Maximum Inner Product Search
FN: Furthest Neighbour
AFN: Approximate Furthest Neighbour
c-AFN: c-Approximate Furthest Neighbour
ET: Early-Termination
NT: Normal-Termination
SNNS: Skyline Nearest Neighbour Search