Nearest Neighbor Search in High Dimensional Space

by

Mingjie Li

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Centre for Artificial Intelligence
Faculty of Engineering and Information Technology
University of Technology Sydney

December, 2020

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 at any other academic institution 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.

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

Signature: Production Note: Signature removed prior to publication.

Date: 10/12/2020

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Ying Zhang for his continuous support and guidance throughout my PhD career. Ying is professional, patient and kind. He introduced me to the research area of nearest neighbor search and further provided constant motivation which kept me going. His valuable ideas and suggestions always guided me and broadened my knowledge and horizon in the related areas. I am very thankful to him for his patience and confidence during my PhD study. Even experiencing failures and difficulties in research and life, his constant encouragement and support always kept me optimistic and positive. Additionally, Ying is a good mentor and friend for me. He gave me many useful advices on planing career development.

I would like to thank Prof. Wei Wang for his constructive ideas and invaluable suggestions on the works in this thesis. The insightful discussions with Prof. Wang always gave me many inspirations. In addition, I would like to thank Prof. Ivor W. Tsang and A/Prof. Lu Qin for the discussions and suggestions on the related topic.

I would like to thank Prof. Xuemin Lin and A/Prof. Wenjie Zhang for supporting the works in this thesis, as most of the works were conducted

in collaboration with them. I thank Prof. Lin for offering a rigorous while interesting research environment.

I am thankful to the faculties and staffs at the school of computer science at University of Technology Sydney. They were very helpful and supportive throughout my PhD study. It was indeed a pleasure to be a part of such an exciting community.

I would like to thank Dr. Xin Cao, Dr. Lijun Chang, Dr. Xiaoyang Wang, Dr. Shiyu Yang, Dr. Zengfeng Huang, and Dr. Yixiang Fang for sharing the ideas and experiences. Thanks to the members from database groups at UNSW and UTS, including Dr. Long Yuan, Dr. Longbin Lai, Dr. Xiang Wang, Dr. Xing Feng, Dr. Xubo Wang, Dr. Haida Zhang, Dr. Yang Yang, Mr. Xuefeng Chen, Dr. You Peng, Dr. Boge Liu, Ms. Xiaoshuang Chen, Mr. Yuren Mao, Dr. Fan Zhang, Dr. Dong Wen, Dr. Dian Ouyang, Ms. Wanqi Liu, Mr. Hanchen Wang, and Mr. Yuanhang Yu, for creating a dynamic and vibrant atmosphere in the labs and in life. I would also like to thank Mr. Daniel Ouyang, Mr. Peng Zhang, and Mr. Xunxiang Yao for their selfless help and care during my PhD life.

Last but not least, I would like to thank my father Mr. Zheng Li, my mother Mrs. Fengying Zhu, my brother Mr. Mingchang Li and my sister Mrs. Jenly Li for their continuous support, encouragement and love during my entire PhD journey and in my life. I am greatly indebted to them.

PUBLICATIONS

- **Mingjie Li**, Ying Zhang, Yifang Sun, Wei Wang, Ivor W. Tsang, Xuemin Lin. An Efficient Exact Nearest Neighbor Search by Compounded Embedding. DASFAA 2018. (Chapter 4)
- **Mingjie Li**, Ying Zhang, Yifang Sun, Wei Wang, Ivor W. Tsang, Xuemin Lin. I/O Efficient Approximate Nearest Neighbour Search based on Learned Functions. ICDE 2020. (Chapter 5)
- Wen Li, Ying Zhang, Yifang Sun, Wei Wang, **Mingjie Li***, Wenjie Zhang, Xuemin Lin. Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement. TKDE 2019. (*Corresponding Author)(Chapter 6)

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Nearest neighbor search (NNS) in high dimensional space is a fundamental and essential operation in applications from many domains, such as machine learning, databases, multimedia and computer vision, to name a few. In this thesis, we investigate both exact and approximate NNS in high dimensional space.

For the exact NNS, we propose an efficient technique which can have a significant speedup over the state-of-the-art exact solutions. Specifically, we first propose a novel compounded embedding technique, by which we achieve a tight distance lower bound for Euclidean distance. Then each point in a high dimensional space can be embedded into a low dimensional space such that the distance between two embedded points lower bounds their distance in the original space. Following the *filter-and-verify* paradigm, we develop an efficient exact NNS algorithm by pruning a large number of candidates using the new lower bounding technique. Comprehensive experiments on many real-world data demonstrate the effectiveness and efficiency of our new algorithm.

In terms of the approximate NNS, we propose an external memory-based approximate NNS algorithm by learning to hash. Specifically, we introduce a novel data-sensitive indexing and query processing framework for approximate NNS with an emphasis on optimizing the I/O efficiency, especially, the sequential

I/Os. The proposed index consists of several lists of point IDs, ordered by values that are obtained by learned hashing functions on each corresponding data point. The functions are learned from the data and approximately preserve the order in the high-dimensional space. We consider two instantiations of the functions (linear and non-linear), both learned from the data with novel objective functions. Comprehensive experiments on six large scale high dimensional datasets show that our proposed methods with learned index structure perform much better than the state-of-the-art external memory-based approximate NNS methods in terms of I/O efficiency and search accuracy.

Although lots of approximate NNS algorithms have been continuously proposed in the literature each year, there is no comprehensive evaluation and analysis of their performance. Therefore, we conduct a comprehensive and systematic experimental evaluation for the state-of-the-art approximate methods. Our study (1) is cross-disciplinary (i.e., including 19 algorithms in different domains, and from practitioners) and (2) has evaluated a diverse range of settings, including 20 datasets, several evaluation metrics, and different query workloads. The experimental results are carefully reported and analyzed to understand the performance results. Furthermore, we propose a new method that achieves both high query efficiency and high recall empirically on majority of the datasets under a wide range of settings.