Research Statement

ZHENG Baihua
School of Computing & Information Systems, Singapore Management University
Tel: (65) 6828-0915; Email: bhzheng@smu.edu.sg
Updated on 1 December 2024

Background

With rapid advances of wireless technologies and explosive growth of smart devices, mobile computing has fundamentally transformed almost every single aspect of our daily lives, from grocery shopping to house hunting. In 2016, mobile internet use surpassed that of desktops for the first time and the trend is not likely to be reversed. In 2017, around 1.54 billion smartphones were sold to end users worldwide. However, the ubiquitous adoption of mobile applications also highlights issues and concerns from security, privacy, scalability to mobile devices' limited power resources and computation capability.

My research focuses on mobile computing while attacking data management problems from multiple angles. My major contributions are in the area of query processing, especially location-dependent queries, applicable to a wide range of mobile environments and queries. They can be divided into three areas, i) trajectory management; ii) efficient query processing, and iii) mobile data management.

Research Areas

Trajectory Management.

My research in this area focuses on four dimensions of car trajectory data, namely volume, variety, veracity and value.

The volume of the trajectory data generated by cars is huge and increases every day. We proposed novel compression techniques to drastically reduce the size of GPS trajectory data while maintaining support for queries without having to fully decompress the data. In addition, we were able to decompose the trajectory data into spatial paths and temporal sequences based on underlying road network. We are the first to explain the inefficiency of representing a trajectory in the format of ⟨xi,yi,ti⟩ sequence theoretically from the perspective of trajectory compression, and to design novel compression algorithms for spatial paths and temporal sequences respectively.

Variety refers to the many sources and types of data. Information captured by different data sources can be complementary. In our works related to the digital map auto-update, we considered both the existing digital map of a given city and the GPS trajectories generated by cars in that city, and matched the trajectories to the map. GPS trajectories could be decomposed into two subsets, the matched set and the unmatched set. Similarly, the road segments in the map could also be partitioned into with-trajectories segments and without-trajectories segments. In our current work, we utilize the unmatched trajectories to generate new road segments that are missing from the given map with both

the geometry properties and the topology features well preserved. Meanwhile, we are studying the without-trajectories segments in the map, as they are very likely to be outdated.

Trajectories are uncertain and imprecise, especially when the sampling rate is low or the localization techniques are not precise. Our strategy to improve veracity of trajectory is to adopt data-driven approaches to correct the errors and to leverage on historical data to resolve the uncertainty. For example, to recover the exact route of a low-sampled GPS trajectory, we proposed a novel system to incorporate both temporal and spatial dynamics while addressing the data sparsity issue with a probabilistic approach. In addition, to reduce the large scale errors (e.g., 100 meters) of the trajectory generated by cellular-based localization techniques, we proposed a general system based on theoretical study of the error property and all available information, including the road network, the position information (i.e., the coordinates), the temporal information (i.e., the timestamps) as well as the motion property of moving objects.

By studying large amount of trajectory data, we can enhance urban transportation systems. Useful information learned from the trajectory data can be used to maximize their value. For example, we proposed DeepTravel, an end-to-end training-based model that can predict the travel time of a whole path directly through learning from the historical trajectories.

Efficient Query Processing.

The ultimate goal is to make the search faster via novel indexing techniques or to make the access of relevant information easier via innovative queries. According to the nature of the underlying searches, my research of this area could be further clustered into three sub-areas, namely spatial queries, non-spatial queries, and searches that involve both spatial attributes and non-spatial attributes.

The answers to spatial queries depend on the spatial relationships between spatial objects, such as the distances between points, and the overlaps between polygons. Initially, I employed Euclidean distance as the metric to evaluate the proximity between queried objects, and formulated new types of spatial queries that were motivated by real-life applications, together with efficient index structures. Next, I explored the obstructed space and the road network. The former considers the existence of physical obstacles and the latter has an underlying network that constrains the movement of mobile objects. As spatial indexes built on top of Euclidean distances are no longer valid in either obstructed space or road networks, I designed new indexes to support complicated queries such as reversed nearest neighbour search and continuous nearest neighbour search. In addition to specific spatial queries, I also designed general frameworks that are able to support a variety of spatial queries, such as ROAD system. Recently, I furthered my work to consider metric space and uncertain data.

As not all the requests from mobile users are spatial queries, my research covers queries for non-spatial objects as well, such as skyline query, reverse skyline query, skyband

query, and top-*k* search based on a given scoring function. My main contributions are formulation of new types of queries and development of efficient index structures. My work can be applied to a variety of data objects, e.g., multi-dimensional objects stored in database, streaming data, string datasets, and multi-dimensional objects with missing attribute values.

To serve increasingly more complicated user requests, a new type of searches that considers both spatial attributes and non-spatial attributes becomes necessary. For example, we proposed IR-tree, an index structure that handles both the textual and the spatial aspects of documents simultaneously and allows searches to adjust weightage on textual relevance and spatial relevance of documents at the run time. It is one of the early works that consider spatial relevance in document search and has since inspired many follow-up studies.

Mobile Data Management.

To enable mobile clients to access the data more efficiently, unique constraints imposed by the underlying wireless networks or the limitations of mobile clients have to be considered. I mainly focused on the access requests of location-dependent data (e.g., the nearest ATM) and my research outputs in this area could be further clustered into three sub-areas, according to the techniques used, namely indexing, caching, and scheduling.

Initially, I focused on the dissemination of location-dependent data in wireless broadcast systems. The fact that the data broadcast on air are only periodically available renders most existing indexing structures inefficient if not useless, as they depend on random access and backtracking. Taking linearity requirement into account, I designed several new indexes to support various location-dependent queries, such as window queries, KNN queries, and continuous-KNN queries, in wireless broadcast systems. To extend my initial study, I started to explore the properties of wireless channels to ensure the search performance in real-world systems, such as the error-prone nature of wireless broadcast, and the privacy issue. Furthermore, I formulated and studied new types of queries motivated by real applications. For example, nearest surrounder queries consider not only the proximity but also the angular relationship between objects and query points.

In addition to wireless broadcast systems, I have also developed high performance index structures for location-dependent query processing in mobile systems of different properties and requirements, such as sensor networks.

Mobile clients cache data in their local memory while moving around. Consequently, it is beneficial to understand how we can fully utilize local cached data. My work addressed two fundamental issues related to mobile client cache: i) the client cache may become invalid due to the change of its location; and ii) new queries may be issued as client moves around. To ensure client cache validity, I developed a method to cache not only the data object but also the corresponding valid region, i.e., a geographical region within which the cached object is guaranteed to be valid for the same query. The novel idea of valid region has been adopted by many researchers. I also investigated different presentations of valid

scopes. To facilitate the processing of newly issued queries, I proposed two caching schemes, proactive caching and complementary space caching, to minimize the cost of a cache miss, i.e., when a client needs help from the server to answer a query.

In addition to air indexing and client side caching, I studied the scheduling algorithm in a multi-channel environment where data objects of various sizes and access frequencies are broadcast in parallel on multiple channels of different bandwidths, with the client listening to one channel at a time. Based on the derived optimal average expected delay for multiple channels, a novel two-level scheduling strategy was proposed to achieve a near-optimal performance. A high-performance privacy preserving scheduling algorithm was also proposed to make available the overall access pattern of data objects without disclosing the individual access pattern.

Selected Publications

Yifan Zhu, Ruiyao Ma, Baihua Zheng, Xiangyu Ke, Lu Chen, Yunjun Gao: GTS: GPU-based Tree Index for Fast Similarity Search. *Proc. the ACM SIGMOD/PODS International Conference of Management of Data (SIGMOD'24)*. Santiago, Chile, June 11-16, 2024

[86] Rui Zhu, Xiaochun Yang, Bin Wang, and Baihua Zheng: Closest Pairs Search Over Data Stream. *Proc. the ACM SIGMOD/PODS International Conference of Management of Data (SIGMOD'24)*. Santiago, Chile, June 11-16, 2024

Yifan Zhu, Lu Chen, Yunjun Gao, Ruiyao Ma, Baihua Zheng, Jingwen Zhao: HJG: An Effective Hierarchical Joint Graph for ANNS in Multi-Metric Spaces. *Proc. the 40th IEEE International Conference on Data Engineering (ICDE'24)*, Utrecht, Netherlands, May 13-16, 2024

[82] Rui Zhu, Yujin Jia, Xiaochun Yang, Baihua Zheng, Bin Wang, Chuanyu Zong: Multiple Continuous Top-K Queries Over Data Stream. *Proc. the 40th IEEE International Conference on Data Engineering (ICDE'24)*, Utrecht, Netherlands, May 13-16, 2024

Congcong Ge, Pengfei Wang, Lu Chen, Xiaoze Liu, Baihua Zheng, Yunjun Gao: CollaborEM: A Self-supervised Entity Matching Framework Using Multi-features Collaboration. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 35(12): 12139-12152, 2023

Yuntao Du, Yujia Hu, Zhikun Zhang, Ziquan Fang, Lu Chen, Baihua Zheng, and Yunjun Gao: LDPTrace: Locally Differentially Private Trajectory Synthesis. *PVLDB* 16(8): 1897-1909, 2023

Janaka Brahmanage, Thivya Kandappu, and Baihua Zheng: A Data-driven Approach for Scheduling Bus Services Subject To Demand Constraints. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 35(7): 6534-6547, 2023

Yuxiang Guo, Lu Chen, Zhengjie Zhou, Baihua Zheng, Ziquan Fang, Zhikun Zhang, Yuren Mao, and Yunjun Gao: CampER: An Effective Framework for Privacy-Aware Deep Entity Resolution. *Proc. the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'23)*. Long Beach, California, USA, August 6-10, 2023: 626-637

Yuqi Chen, Hanyuan Zhang, Weiwei Sun and Baihua Zheng: RNTrajRec: Road Net-work Enhanced Trajectory Recovery with Spatial-Temporal Transformer. *Proc. the 39th IEEE*

International Conference on Data Engineering (ICDE'23), Anaheim, California, USA, April 3 – 7, 2023: 829-842