2019 IEEE 35th International Conference on Data Engineering (ICDE)

A Novel Representation and Compression for Queries on Trajectories in Road Networks (Extended abstract)

Xiaochun Yang Bin Wang Kai Yang
School of Computer Science and Engineering
Northeastern University
Shenyang, China
{yangxc,binwang}mail.neu.edu.cn

Chengfei Liu
Faculty of Science, Engineering and Tech.
Swinburne University of Techenology
Melbourn, Australia
cliu@swin.edu.au

Baihua Zheng
School of Information Systems
Singapore Management University
Singapore
bhzheng@smu.edu.sg

***Abstract*—Recording and querying time-stamped trajectories incurs high cost of data storage and computing. In this paper, we explore characteristics of the trajectories in road networks, which have motivated the idea of coding trajectories by associating timestamps with relative spatial path and locations. Such a representation contains large number of duplicate information to achieve a lower entropy compared with the existing representations, thereby drastically cutting the storage cost. We propose techniques to compress spatial path and locations separately, which can support fast positioning and achieve better compression ratio. For locations, we propose two novel encoding schemes such that the binary code can preserve distance information, which is very helpful for LBS applications. In addition, an unresolved question in this area is whether it is possible to perform search directly on the compressed trajectories, and if the answer is yes, then how. Here we show that directly querying compressed trajectories based on our encoding scheme is possible and can be done efficiently. We design a set of primitive operations for this purpose, and propose index structures to reduce query response time. We demonstrate the advantage of our method and compare it against existing ones through a thorough experimental study on real trajectories in road network.**

***Index Terms*—trajectory, compression, LBS, road network**

I. Introduction

For the purpose of reducing overhead in data storage and processing, trajectory compression is used to compress the size of trajectories while maintaining their utility. In this paper, we consider both storing and querying trajectories in road network. We aim to store trajectories using relatively small space and support queries with high performance.

There are mainly two types of representations for trajectories in road networks. A typical type of expressions combines a timestamp t and a 2D position (x, y) together in the form of (t, x, y) to express a time-stamped position in a trajectory [5]. Such a representation causes big overhead of data storage and computing. The other type of expressions separates spatial locations from timestamps, using consecutive edges $\langle e_i, \dots, e_j \rangle$ to represent a spatial path of a trajectory, and a sequence of distance-time (d_i, t_i) pairs to capture the temporal information. PRESS [3] and the generalized in-network trajectory data model proposed by Sandu-Popa et al in [2] are the latest representative works of the second type of expressions. However, a good compression ratio can only be achieved under a large error bound/error threshold. Literature [2] does not report how to do query processing on their compressed trajectories, and the query processing in [3] heavily relies on decompression.

In this paper, we propose a novel representation, a lossless compression for both spatial path and timestamps, and an error-bounded compression for locations. Such representation and compression can achieve high compression ratio under a small error bound and can support queries efficiently. The first challenge is to design a good representation with small entropy (i.e. the representation contains large amount of duplicate information) to facilitate storing and querying trajectories in road networks, considering both space overhead and efficiency. This means that, from the compression perspective, it prefers entropy of a trajectory representation to be low; and from the querying perspective, it aims at being able to support query processing efficiently. To attain the above goals, we propose a novel representation of trajectories in road networks (called TED-representation) to separately represent spatial entry path, distances, and timestamp. This separation provides ideal properties to support both compression and location-based query processing, with mainly two advantages: (i) it enables us to capture characteristics of trajectories in road networks, and enables our expression to achieve lower entropy than existing representations, and (ii) it allows us to easily associate these three dimensions to build up a close relation among a spatial path, locations, and timestamps, which enables us to effectively cut down the error bound.

The second challenge is to propose compression algorithms to transform TED representation into shorter binary words, such that TED representation can be recoverable from them. We propose several techniques to compress spatial entry paths

The work is partially supported by the National Natural Science Foundation of China (Nos. 61572122, 61532021), Liaoning BaiQianWan Talents Program, ARC Discovery Projects under Grant Nos. DP160102412 and DP170104747, and the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

2375-026X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDE.2019.00253

and locations separately. For spatial entry paths, we propose a fixed-length encoding for a single path and consider the feature of trajectories in a road network to compress multiple trajectories. Such a compression can support fast location and it is also able to achieve a high compression ratio, especially when the total number of trajectories that need compression is large. We will demonstrate that this compression can drastically reduce the storage costs, achieve a high compression ratio, and support all kinds of paths, including non-shortest paths, acyclic single trajectories, periodical trajectories, and multiple trajectories. For locations, we first propose a distance-preserving encoding scheme called *DP-encoding* to encode locations. Then we propose a novel encoding scheme DDP-encoding to make the code decodable, and a pruned DDP-encoding to further save space so that the size of codes for location is close to that of Huffman encoding. We show that these two encoding schemes can preserve distance information as well as process queries efficiently for LBS applications.

The last but not the least challenge is to devise techniques to answer typical queries on compressed trajectories. We list two types of primitive operations, which are fundamental functions to support LBS related applications. We show that query processing can be effectively limited to a small candidate region in compressed trajectories, and only a small part of data needs to be decompressed, which is efficient. We propose novel index structures and algorithms for this purpose, and reduce the primitive operation response time. We then present our algorithms for four types of commonly used LBS queries, demonstrate the advantage of our method and compare it against existing ones through a thorough experimental study on real trajectories in road network.

II. Trajectory Representation and Framework

A trajectory Tr is a series of time-stamped raw positional data p in the form of (t, x, y), where t is a timestamp, and (x, y) refers to a location in a 2D Euclidean space with a latitude x and a longitude y. A road network is generally defined as a directed graph $G = (V, E)$, where V is the vertex set and E is the edge set. Each vertex has different exit entries pointing to different edges. The exit entries of a vertex are unique consecutive numbers starting from 1.

A. Framework

We propose a new framework to compress trajectories for physical storage and to support query processing for LBS applications. Fig. 1 shows our framework. Each GPS trajectory can be converted to an embeded trajectory via map matching process [1], [3]. Like PRESS, we represent each time-stamped data p via a triple $(\underline{t}ime, \underline{e}dge, \underline{d}istance)$, therefore an embedded trajectory can be represented by a *time sequence (T)*, a *spatial entry path (E)*, and a *distance sequence (D)*. Unlike PRESS, spatial information is represented by spatial entry paths and distances with a much lower entropy and hence corresponding compression algorithms are able to achieve much higher compression ratios. The representation of timestamps (T) facilitates the association between entry paths (E) and distances (D). A trajectory represented in this TED format is called a *TED-trajectory*.

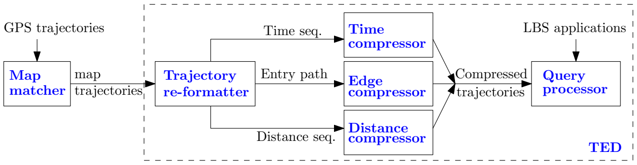

Fig. 1. Our framework for representing and compressing trajectories.

We then compress TED-trajectories by compressing T, E, and D separately. Efficient search algorithms are developed to process LBS queries by only partially decompressing compressed trajectories.

III. TED-Trajectories Compression

Given a trajectory T_r in the form of (t, x, y) sequence, and a compressed trajectory $T_r{}'$ of T_r based on its TED representation, the effectiveness of the TED representation and compression is evaluated by the *compression ratio* which is defined as the rate of T_r's storage cost to that of $T_r{}'$, i.e., $\frac{|T_r|}{|T_r{}'|}$. In this section, we explain how to perform compression on TED-trajectories. The high compression ratio achieved by TED-trajectories well justifies the advantage of TED representation in terms of compression.

IV. Query on Compressed Trajectories

In this section, we propose two types of primitive operations: (i) transformation operations among TED codes, and (ii) mapping operations between the logical format (t, x, y) and their compressed codes. Ideally, we hope to query compressed data directly [4], i.e. without fully decompressing the trajectories, or in the ideal case without decompressing the trajectories at all. Many of LBS queries shown in Table I can be supported using primitive operations.

TABLE I
Major LBS Queries.

Query types	Queries
Basic	*where*(T_r, t), *when*(T_r, x, y)
Range	*distance*(T_r, t_1, t_2), *howlong*$(T_r, x_1, y_1, x_2, y_2)$
Aggregation	*count*(T_{rs}, x, y, r)
General	*kNN*(T_{rs}, x, y, t_1, t_2), *window*$(T_{rs}, x_1, y_1, x_2, y_2, t_1, t_2)$

References

[1] J. Krumm. Trajectory analysis for driving. In *Computing with Spatial Trajectories*, pages 213–241. 2011.

[2] I. S. Popa, K. Zeitouni, V. Oria, and A. Kharrat. Spatio-temporal compression of trajectories in road networks. *GeoInformatica*, 19(1), 2015.

[3] R. Song, W. Sun, B. Zheng, and Y. Zheng. PRESS: A novel framework of trajectory compression in road networks. *PVLDB*, 7(9), 2014.

[4] X. Yang, B. Wang, C. Li, J. Wang, and X. Xie. Efficient direct search on compressed genomic data. In *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, pages 961–972, 2013.

[5] J. Yuan, Y. Zheng, X. Xie, and G. Sun. T-drive: Enhancing driving directions with taxi drivers' intelligence. *TKDE*, 2012.