Go Beyond Raw Trajectory Data: Quality and Semantics

Kai Zheng Han Su
The University of Queensland, Brisbane, Australia
{uqkzheng, h.su1}@uq.edu.au

Abstract

Past decades have witnessed extensive studies from both academia and industries over trajectory data, which are generated from a diverse range of applications. Existing literature mainly focuses on raw trajectories with spatio-temporal features such as location, time, speed, direction and so on. Recently, the pervasive use of smart mobile devices like smart phones, watches and bands have brought about more generation of trajectory by personal users (instead of companies or organizations) and from online space (instead of physical space), where individuals can decide when and where to log on and share their locations with others. The more discentralized and contextualized trajectory sources have brought some unique challenges for database management with respect to the quality and semantics of trajectories data. With more applications and services relying on trajectory data analysis, it is necessary for us to think about how these new issues will affect the traditional way that trajectories are digested and processed. In this paper we will elaborate on these challenges and introduce our recent progress in the respective directions. The message we try to deliver is that raw trajectories themselves no longer satisfy the requirement of today's mainstream applications. To really release the power of trajectory-based applications, we should go beyond the raw trajectory data by enhancing their quality and semantics, which calls for novel computing architectures, paradigms and algorithms with sufficient capabilities to manage and analyse the enhanced trajectory data.

1 Introduction

The increasing availability of location-acquisition technologies including telemetry attached on wildlife, GPS set on cars, WLAN networks, and mobile phones carried by people have enabled tracking of almost any kind of moving objects, which results in huge volumes of spatio-temporal data in the form of trajectories [36]. Trajectory data consists of rich information about when and where a particular moving object is and offers unprecedented opportunity for discovering its mobility patterns. This inspires tremendous amount of research in trajectory data from a variety of aspects in the past decade, ranging from designing effective indexing structures [24] [8] [22] [9] [13] and efficient query processing algorithms [24] [29] [11] [14], to data mining and knowledge discovery [19] [16] [15] [21]. Despite their significant contributions in this area, traditional research on trajectory data has primarily focused on its raw format, i.e., a sequence of spatio-temporal points collected directly from the location-acquisition devices. While there was nothing wrong with this research philosophy especially back in the days when the source and scale of trajectory data are quite limited, recent advances in sensor technologies and

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

location-based social networks (LBSN) have posed new challenges to this community, which can be summarised in the following two aspects.

- **Challenge 1: Data Quality.** Although a trajectory can be theoretically modelled as a continuous function mapping from time to space, in a database it is actually a discrete sequence of spatio-temporal locations sampled from the movement of an object. In other words, when a raw trajectory is reported to the server and stored, it is just a sample of the original travel history. Therefore different sampling rates can result in completely different raw trajectories even for the same travel history. Since the sources of trajectory data are so diversified nowadays, the sampling rates vary significantly from one application to another. As a few examples, a geologist equipped with specialized GPS-devices can report her locations with very high frequency (e.g., every second) while a casual mobile phone user may only provide one location record every couple of hours or even days (via, for example, a check-in service in LBSN). Such variations can also be imposed by external factors (such as availability of on-device battery and wireless signal) and may change at the users discretion. In this big data era, it is not uncommon that we need to integrate trajectories across multiple sources and analyse them altogether. Nevertheless, our previous study [26] has shown that the great variance in sampling rates can render existing trajectory distance functions (e.g., DTW [17], LCSS [29] or EDR [10], ERP [11]) ineffective, which will in turn affect the algorithms, systems and applications relying on those distance functions. From database perspective, this essentially is a data quality problem that can be present in many analytical tasks involving multi-sourced and heterogeneous data. Systematic approaches are desired in order to gain deeper understanding of its root cause and eventually develop a comprehensive solution.
- **Challenge 2: Data Semantics.** Recent years have witnessed the flourish of location-based social networks (LBSN) that enables people to add a location dimension to existing online social networks in a variety of ways. For example, users can upload geo-tagged photos/videos to Flickr [2], Instagram [4] and/or Vimeo [6] to share their great moment with friends, comment on an event in Twitter [5] with geo-tagged tweets, check-in at a restaurant on Foursquare [3], or log bicycle trails for sport analysis and experience sharing on Bikely [1]. The location dimension serves as glue in LBSN that bridges the gap between physical and digital world. In other words, by aggregating all the geo-tagged contents posted by a user in her cyber-space (i.e., LBSN), we can actually know not only where and when she has been, as in the traditional trajectory database, but also what she was doing by extracting the information from the multimedia contents attached to the locations (e.g., text, images, videos). Moreover, we can even transform raw trajectories collected from GPS modules to semantic trajectories by applying semantic annotation techniques [7,30]. With such a large volume of trajectory data enriched with semantic and activity information, we are confronted with challenges in terms of managing, analysing and understanding it. Due to the complex and combinatorial nature of this data, techniques across multiple areas including database, multimedia, data mining and natural language processing should be considered.

In this paper we will categorize and introduce our recent progress that has been made with respect to the above challenges. Generally speaking, we have found that the knowledge derived from raw trajectories is quite limited in most cases and even misleading sometimes. We believe the mainstream location-based services should base themselves on a higher level of abstraction for trajectory data, the one that has been dedicatedly processed to acquire better quality and more semantics. Figure 1 demonstrates our opinion about the relative positions of raw trajectories and enhanced trajectories in modern trajectory data management systems and applications. In the remainder of this paper, we will focus on explaining the research philosophy of our work and their relationships, while referring interested readers to the original papers for technique details.

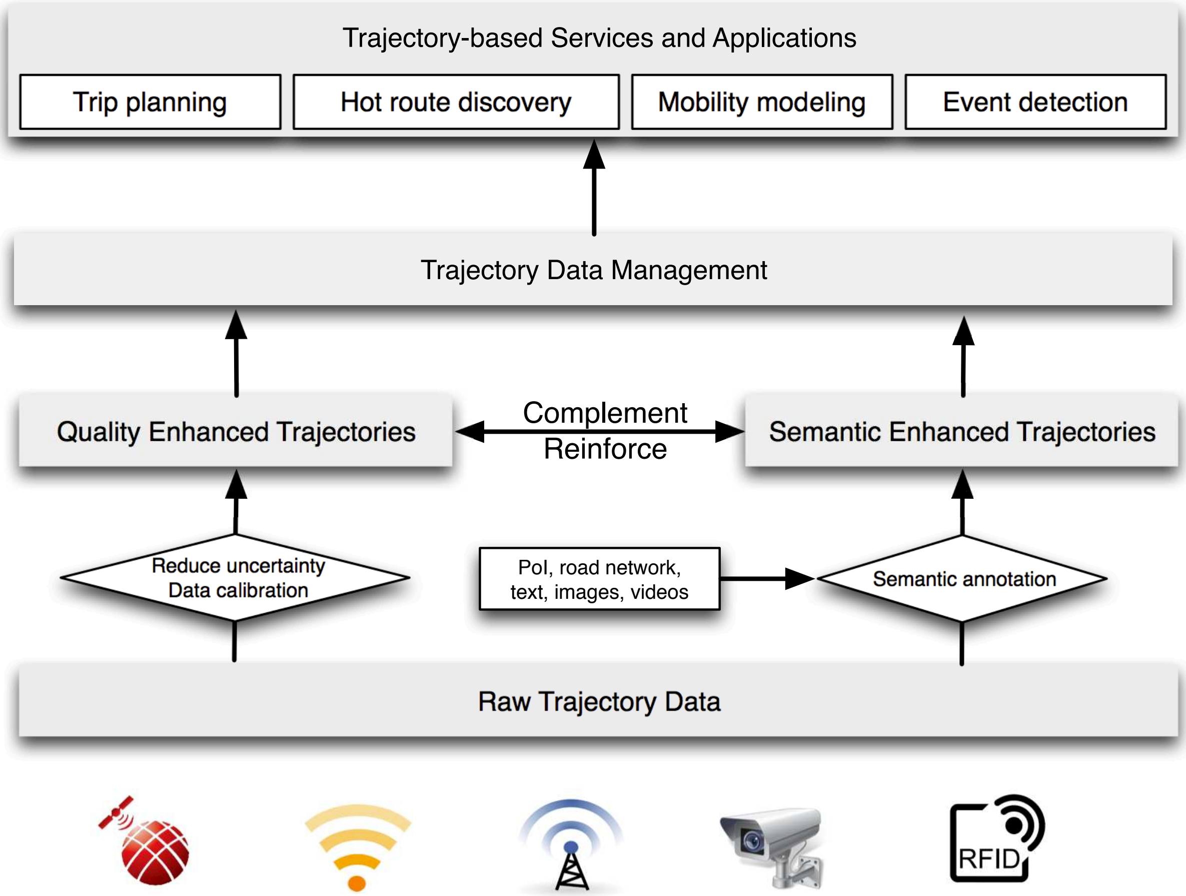

Figure 1: This figure demonstrates the relationship between raw trajectories, quality enhanced trajectories and semantic enhanced trajectories. It also illustrates their relative positions in trajectory data management systems and trajectory-based applications.

2 Quality-Aware Trajectory Management

Nowadays trajectory data can be generated from highly diversified services and applications, resulting in data with different qualities. Generally trajectory data quality issues can arise from two levels: point level and trajectory level. The first one is caused by the inaccuracies of location-acquisition devices and systems, i.e., the reported location is deviated from its actual location. Although this issue seems inherent and inevitable, we normally do not regard it as a major problem due to rapid advances of tracking technologies (e.g., GPS with sub-meter precision). Our focus is then on the second level, which is caused by the sampling rates of trajectories. As mentioned before, a trajectory in database is just a sample of its original travel history. Because nothing is known about the objects' whereabouts in-between two consecutive sampled locations, a trajectory is of low quality or high uncertainty if its sampling rate is low. To deal with an object's location in-between those samples, a typical technique is to apply interpolation [20] by which means the sampled positions become the end points of line segments, and the trajectories are transformed into polylines in 3D (*x*-*y*-*time*) space. However, as pointed out by [12, 18, 23], interpolation cannot reflect the exact movement pattern of an object. In theory, a moving object can be located anywhere within a given (bounded) region, as long as it does not violate physical constraints (e.g., maximum allowed velocity). Some efforts have been made to consider this issue when processing trajectory data by proposing probabilistic queries [28, 31] that, instead of reporting the result only, provides the confidence of the result being true as well. However the quality of the trajectories cannot benefit or be improved from those approaches. In this section, we will introduce our methodologies to tackle this problem – enhancing the trajectory quality, which we believe to be more fundamental and efficient solutions compared to the expensive probabilistic queries.

2.1 Enhancing Trajectory Quality by Reducing Uncertainty

In [32] we aimed at reducing the uncertainty of a trajectory with low sampling rate, which is the main cause of trajectory quality issues. More specifically, given a low-sampled trajectory, our goal is to estimate its original and complete route/path on the underlying road network. At the first glance this seems a mission impossible if we simply act on each low-sampled trajectory separately since no better estimation can be done than linear interpolation for consecutive samples. However we have made two important observations based upon our analysis on real data. First, *travel patterns between certain locations are often highly skewed.* This is due to the fact that, when people travel, they often plan the route based on the experience of their own or others, rather than choosing a path randomly. The skewness of travel pattern distribution makes it feasible to distinguish the possible routes based on their popularity. The second observation, which is more interesting, is that *similar trajectories can often complement each other to make themselves more complete.* This implies that if we consider these trajectories collectively, they may reinforce each other to form a more complete route. These two observations show that the original route of a low-sampled trajectory can be estimated to some extent if a set of historical trajectories within the same spatial domain is available. Now the question is how to leverage this historical data. Intuitively, given a low-sampled trajectory, one can simply search for the historical trajectories that pass by all the sampled locations of the given trajectory and then find the most popular routes. Nevertheless, since the given trajectory can have arbitrary locations, we usually cannot find any historical trajectory that matches the whole part of the query very well. Even if we can, the amount may not be large enough to serve as reliable statistics. Therefore we propose a more practical solution consisting of three steps. Firstly, we divide the whole query into a sequence of sub-queries and search for the *reference trajectories* that can give hints on how each sub-query travels. Then we infer the *local routes* for each sub-query by considering the reference trajectories in a collective manner. At last, we connect consecutive local routes to form the *global routes* and return the ones with the highest scores to the users. As a summary, the essence of the route inference approach in this paper is to extract the travel pattern from history, and infer the possible paths of the query by suggesting a few popular routes. Compared to the original number of possible routes, the uncertainty of the given trajectory is reduced significantly in this way. Please refer to [32] for the detailed algorithms.

2.2 Enhancing Trajectory Quality by Data Calibration

Data quality issues do not just lie in low-sampled trajectories. In [26] we observed that trajectories with inconsistent sampling rate (no matter low or high) are almost incomparable and make the most classical trajectory distance functions less effective. To address this problem, in [26] we take a different philosophy that, instead of manipulating or adjusting the original trajectory data, uses a fixed and data independent set of spatial objects (called reference system) to re-write all the sampled locations of the original trajectories. This process is called trajectory calibration, the aim of which is to reduce the inconsistency in the sampling rates amongst all trajectories and improve the effectiveness of similarity-based trajectory analysis. Nonetheless it is a non-trivial task to perform trajectory calibration. First, building a good reference set is the stepping stone for the entire system. Since our goal is to rewrite the trajectory data using the reference set, we expect a good reference set to be stable, independent of data sources, and have a strong association with the trajectory data. The first and second properties are essential for producing trajectories in a unified form, while the third property ensures that the calibration process will not introduce a large deviation from the original routes. Trajectory calibration may encounter three circumstances when rewriting a trajectory with the reference set: 1) a trajectory point may need to be shifted and aligned onto the reference; 2) some trajectory points may need to be removed or merged (when the sampling rate is higher than necessary); 3) some new trajectory points may need to be inserted (when the sampling rate is too low), all in the context of the chosen reference system. Further, the criteria to judge the goodness of the calibration results need to be established, for the system to enforce efficiently and effectively and for the users to understand to what extent the calibration can improve the data analysis results. The calibration framework we proposed comprises two components: a reference system and a calibration method. For the first component, we present several reference systems by defining different types of anchor points (space-based, data-based, POI-based and feature-based), which are fixed small regions in the underlying space. A series of strategies are designed for the calibration component, including the methods to insert anchor points to trajectories in order to make them more complete without scarifying geometric resemblance to the original routs. Please refer to [26] for more technical details.

3 Semantic Enhanced Trajectory Management

A trajectory in its raw format is just a sequence of spatio-temporal locations (e.g., a GPS point is a triplet (longitude, latitude, timestamp)). Although a lot of research have been done towards mining interesting patterns from a collection of trajectories purely based on their spatio-temporal features [15, 16, 19, 21, 34, 35], the results from those mining algorithms are often hard to explain and interpret for humans.This is because raw trajectory data can only reveal *when* and *where* a person was but cannot tell *what* she was doing (i.e., activity) and *how* she went there (i.e., moving behaviour) without leveraging extra information at semantic level. There have been some preliminary studies that enrich GPS locations with semantic entities such as POIs, roads, regions, resulting in semantic trajectories or annotated trajectories [7, 25]. These work focused on how to determine the correct semantic label for each trajectory point when multiple entities are in its vicinity, i.e., the generation of semantic trajectory data. In this section, we will introduce our recent research in managing and processing trajectories that have been enhanced with semantic information.

3.1 Querying Semantic Trajectories

Even though semantic trajectories contain much more information than raw trajectories, their value cannot be derived and utilised until there is an appropriate way to store, manage and process this data efficiently to a large scale. In this light, we developed a database storage framework to support efficient indexing and query processing over activity trajectory [33], which is a specific kind of semantic trajectory with textual information (e.g., keywords, tags, short phrases) describing user's activity at each location. More precisely, we propose a novel similarity query for activity trajectories by incorporating both geometric distance and activity match into the similarity measure, with the goal of returning more meaningful results to the users. However, answering this new query turns to be a more challenging problem since just making use of either location or activity information for pruning the search space will result in bad query performance. Our approach to this problem starts with a novel grid index called *GAT*, which includes a hierarchy of cells for each activity, an inverted list of trajectories containing each activity within each cell, and a summarized sketch of activities for each trajectory. GAT keeps the advantage of hierarchical spatial index while avoiding the flaws of large "dead zones" when indexing trajectories by minimum bounding boxes. In addition, the index not only uses the local information on trajectory segments within the cells but also preserves some global information for the entire trajectory in the activity sketch, so that its pruning power can be boosted. On top of the index, we develop a best-first search strategy with tighter distance lower bound for all "unseen" trajectories in the database and an efficient algorithm to compute the distance between candidates and the query. The interested readers can find more details about the indexing structure and search algorithms in our paper [33].

3.2 Summarising Trajectories with Short Text

The common way to generate semantic trajectories is to mechanically replace the coordinate of each location with a semantic entity, which often yields excessive information for people to digest and interpret. Therefore another direction we have been working on recently is to find a more compact, expressive and interpretable way to represent semantic trajectories. Inspired by text summarisation in information retrieval, we in [27] proposed to use short text to summarise and represent a trajectory by leveraging a diversified source of auxiliary information (e.g., PoI, road network). We found the textual representation can be superior than raw and semantic trajectories in two aspects. First, as the output is a summarization rather than mechanical transformation of raw trajectories (like semantic trajectories), data volume will be reduced significantly. Second, despite of smaller data size, the information conveyed in the text are strategically focused on the most 'interesting' parts of the trajectories, thus making more sense for humans. A partition-and-summarization framework was proposed in our work. The partition phase tries to find an optimal partition according to the user's granularity requirement, which can minimize the variation of predefined features for the trajectory segments within the same partition. The purpose of this optimization is to use more compact representation to summarize each partition. In the summarization phase, we define a novel measure for the unusualness of each feature by employing the common patterns amongst other trajectories, and generate textual description for the most unusual features with a predefined template. Please refer to [27] for more details about this framework.

4 Concluding Remarks

In this paper we have discussed some new challenges in trajectory data management that were brought about by the emergence of location-based services and explosion of smartphone users. Particular attentions are paid on two aspects – quality and semantics, which are believed as vital dimensions to uncover the true value of trajectory data for government, businesses and personal users. We introduce some of our recent studies in addressing these issues from different angles and highlight the connection between our research and the conventional ones. We hope these discussions can trigger more research interest and efforts in developing modern computing platforms and data management systems for trajectories – one of the most ubiquitous and accessible data today.

5 Acknowledgements

This work has been partially supported by ARC Discovery Early Career Researcher Award (DE140100215).

References

[1] Bikely. `http://www.bikely.com/`.

[2] Flickr. `https://www.flickr.com/`.

[3] Foursquare. `https://foursquare.com/`.

[4] Instagram. `https://instagram.com/`.

[5] Twitter. `https://twitter.com/`.

[6] Vimeo. `https://vimeo.com/`.

[7] L.O. Alvares, V. Bogorny, B. Kuijpers, J.A.F. de Macedo, B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. In *GIS*, pages 1–8, 2007.

[8] Y. Cai and R. Ng. Indexing spatio-temporal trajectories with chebyshev polynomials. In *SIGMOD*, pages 599–610, 2004.

[9] V.P. Chakka, A.C. Everspaugh, and J.M. Patel. Indexing large trajectory data sets with seti. In *CIDR*, 2003.

[10] L. Chen and R. Ng. On the marriage of lp-norms and edit distance. In *VLDB*, pages 792–803, 2004.

[11] L. Chen, M.T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *SIGMOD*, pages 491–502, 2005.

[12] R. Cheng, D.V. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *SIGMOD*, pages 551–562, 2003.

[13] P. Cudre-Mauroux, E. Wu, and S. Madden. Trajstore: An adaptive storage system for very large trajectory data sets. In *ICDE*, pages 109–120, 2010.

[14] E. Frentzos, K. Gratsias, N. Pelekis, and Y. Theodoridis. Nearest neighbor search on moving object trajectories. *SSTD*, pages 328–345, 2005.

[15] H. Jeung, H.T. Shen, and X. Zhou. Convoy queries in spatio-temporal databases. In *ICDE*, pages 1457–1459, 2008.

[16] H. Jeung, M.L. Yiu, X. Zhou, C.S. Jensen, and H.T. Shen. Discovery of convoys in trajectory databases. *Proceedings of the VLDB Endowment*, 1(1):1068–1080, 2008.

[17] J.B. Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, pages 201–237, 1983.

[18] B. Kuijpers and W. Othman. Trajectory databases: data models, uncertainty and complete query languages. *Journal of Computer and System Sciences*, 2009.

[19] J.G. Lee, J. Han, and K.Y. Whang. Trajectory clustering: a partition-and-group framework. In *SIGMOD*, page 604, 2007.

[20] José Antonio Cotelo Lema, Luca Forlizzi, Ralf Hartmut Güting, Enrico Nardelli, and Markus Schneider. Algorithms for moving objects databases. *The Computer Journal*, 46(6):680–712, 2003.

[21] Z. Li, B. Ding, J. Han, and R. Kays. Swarm: Mining relaxed temporal moving object clusters. *Proceedings of the VLDB Endowment*, 3(1-2):723–734, 2010.

[22] J. Ni and C.V. Ravishankar. Indexing spatio-temporal trajectories with efficient polynomial approximations. *TKDE*, 19(5):663–678, 2007.

[23] D. Pfoser and C.S. Jensen. Capturing the uncertainty of moving-object representations. In *SSD*, pages 111–131, 1999.

[24] D. Pfoser, C.S. Jensen, and Y. Theodoridis. Novel approaches to the indexing of moving object trajectories. In *VLDB*, pages 395–406, 2000.

[25] S. Spaccapietra, C. Parent, M.L. Damiani, J.A. De Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146, 2008.

[26] Han Su, Kai Zheng, Haozhou Wang, Jiamin Huang, and Xiaofang Zhou. Calibrating trajectory data for similarity-based analysis. In *SIGMOD*, pages 833–844. ACM, 2013.

[27] Han Su, Kai Zheng, Kai Zeng, Jiamin Huang, Nicholas Jing Yuan, and Xiaofang Zhou. Making sense of trajectory data: A partition-and-summarization approach. In *ICDE*. IEEE, 2015.

[28] G. Trajcevski, R. Tamassia, H. Ding, P. Scheuermann, and I.F. Cruz. Continuous probabilistic nearest-neighbor queries for uncertain trajectories. In *EDBT*, pages 874–885, 2009.

[29] M. Vlachos, D. Gunopoulos, and G. Kollios. Discovering similar multidimensional trajectories. In *ICDE*, page 0673, 2002.

[30] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528. ACM, 2011.

[31] K. Zheng, G. Trajcevski, X. Zhou, and P. Scheuermann. Probabilistic range queries for uncertain trajectories on road networks. In *EDBT*, pages 283–294, 2011.

[32] K. Zheng, Y. Zheng, X. Xie, and X. Zhou. Reducing uncertainty of low-sampling-rate trajectories. In *ICDE*, 2012.

[33] Kai Zheng, Shuo Shang, Nicholas Jing Yuan, and Yi Yang. Towards efficient search for activity trajectories. In *ICDE*, 2013.

[34] Kai Zheng, Yu Zheng, Nicholas Jing Yuan, and Shuo Shang. On discovery of gathering patterns from trajectories. ICDE, 2013.

[35] Kai Zheng, Yu Zheng, Nicholas Jing Yuan, Shuo Shang, and Xiaofang Zhou. Online discovery of gathering patterns over trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1974–1988, 2014.

[36] Y. Zheng and X. Zhou. *Computing with spatial trajectories*. Springer, 2011.