Trajectory Analytics: Indexing, Mining, and Applications to Network Optimization

Sayan Ranu, Aditya D. Telang, Vinay Kolar
IBM Research
Manyata Embassy Business Park
Bangalore, India.
{sayanranu, aaditya.telang, vinkolar}@in.ibm.com

1. MOTIVATION

The last decade has witnessed an unprecedented growth in the availability of devices equipped with location-tracking sensors. Examples of such devices include cellphones, in-car navigation systems and weather monitoring gadgets. The widespread usage of these devices has resulted in an abundance of data that are in the form of trajectories. Querying and mining these trajectories is critical to extract the knowledge hidden in the raw datasets and to design intelligent spatio-temporal data management systems. For example, in cellular network analytics, cellphone service providers are interested in identifying "bottleneck" regions that affect a high volume of trajectories due to poor coverage [15]. In zoology, tracking movements of animals is critical to understand their flocking behaviors [8] and migratory patterns [1]. Meteorologists are interested in analyzing trajectories of wildfires and tornadoes to identify "alleys" that are conducive to these environmental disasters [13]. For efficient traffic management, governmental agencies are interested in understanding how congestions spread across a road network so that infrastructure developments can be prioritized in an appropriate manner. In this tutorial, we will survey the rich area of trajectory analytics, which offers solutions to the above problems. Additionally, we will analyze the pros and cons of existing techniques and outline challenges that remain to be solved.

Broadly, in this tutorial, we will focus on three different aspects of trajectory analytics: matching and indexing trajectories, mining patterns from trajectories, and their applications on cellular network optimization.

2. MATCHING AND INDEXING TRAJECTORIES

DTW, originally designed for aligning time series sequences, is among the first distance metrics used for matching trajectories. DTW employs dynamic programming to automatically adapt to *local time shifts* (i.e., detect spatial similarity in trajectories that follow similar paths, but certain sub-paths are shifted in time) and match sequences of different lengths by elastically shifting the time axis. Vlachos et al. [14] analyzed the inadequacies of DTW on matching trajectory data and proposed the LCSS technique. The paper highlighted that in DTW, every point in a trajectory needs to be matched and thus, few strong mismatches can drastically influence the overall distance. Consequently, LCSS ignored mismatches and computed similarity based on the number of points in a trajectory that "match". Two points are considered to "match" if they are spatially within a distance threshold of ϵ. In subsequent works, Chen et al. developed ERP [3] and EDR [4] to further improve LCSS. EDR built on the same threshold-based point-matching intuition proposed by LCSS. However, EDR also incorporates the number of poorly matched regions in the distance computation. More recently DISSIM is proposed by Frentzos et al. [5]. DISSIM targets the scenario where two trajectories are considered similar if they travel through similar regions at similar speeds. In this tutorial, we will discuss each of these trajectory matching techniques, analyze their weaknesses and strengths, and provide insights on how existing techniques need to adapt to deal with the noise inherent in majority of the trajectories.

In addition to identifying the optimal alignment between two trajectories, it is also essential to design index structures so that similarity queries on trajectories can be answered in a scalable manner. Generally, matching trajectories is an expensive task due to its quadratic computation cost. In this tutorial, we will discuss the unique challenges faced while indexing trajectory databases and present an overview of the general strategies employed by existing indexing techniques.

3. MINING PATTERNS FROM TRAJECTORIES

The ability to match trajectories opens up the opportunity to mine patterns from trajectory databases. In the second half of the tutorial, we will survey techniques on mining patterns from a collection of trajectories and summarize their cumulative behaviors. The tutorial will deep dive into two specific pattern mining areas: frequent sub-trajectory mining, and clustering trajectories. Frequent sub-trajectories [6] not only reveal commonalities across a collection of trajectories, but also play an important role in predicting future destinations [11]. Clustering trajectories, on the other hand, allows us to group trajectories based on some common behavior. A number of formulations have been proposed in defining a cluster in the domain of trajectories [7,9,10,12]. In this tutorial, we will first introduce each of these definitions and then present how identifying these clusters can be accomplished.

4. APPLICATION IN CELLULAR NETWORKS

Cellular network operators collect Call Detail Records (CDRs) that record the user's voice or data transaction. CDR includes information such as user's number, base-stations connected during the transaction, and user experience. When CDR data is fused with the

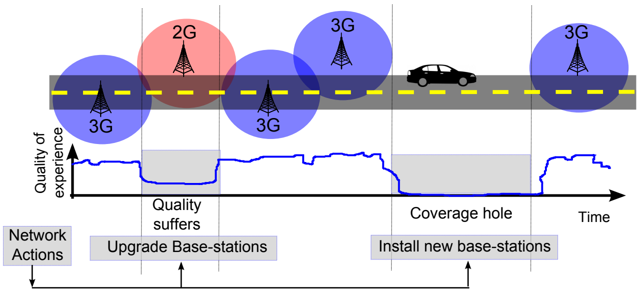

Figure 1: Illustration of cellular network optimization using trajectory data

Geographical Information (latitude, longitude) of the base-stations, cellular operators can derive user locations and trajectories, as well as the density of users in the city [2].

This rich spatio-temporal information provides novel ways for the operator to optimize the network and monetize the user's movement pattern captured (e.g., targetted advertisements to users based on their trajectories and hangouts). The main challenge in such a system are: (1) building a big-data platform that captures billions of CDRs for millions of customers, (2) performing spatio-temporal analytics on the big-data; and (3) Inventing new ways to solve complex problems using the rich CDR data.

In this tutorial, we will introduce the problem of trajectory analytics using the CDR big-data, and discuss novel use-cases that aid network optimization for mobile-users. Figure 1 succinctly illustratives the use-case. A mobile user experiences differentiated quality of experience based on which technology base-station (older 2G or newer 3G base-stations). Clearly, trajectory analytics on CDRs can be utilized in the above example to identify areas of no/low coverage, and provide insights to the operator on 'where' and 'how' to optimize the network. We will discuss such network planning and provisioning problems for mobile-users using trajectory analytics on CDR data.

5. INTENDED AUDIENCE AND PREREQUISITE KNOWLEDGE

This tutorial is targeted towards computer scientists interested in the field of data analytics, which includes graduate students and faculty members from academia as well as industry professionals. The tutorial is organized in a self-contained way and does not assume any particular expertise from the audience. By the end of the tutorial, the goal is to expose the audience to the diverse set of problems arising in trajectory analytics, demonstrate how these problems translate to real life applications, and finally, equip attendees with technical insights on how these problems can be solved.

The tutorial is of interest to the COMAD audience since it delves into the topic of trajectory analytics, which is extremely relevant in today's society due to the increasing pervasiveness of location-tracking technologies. The tutorial will survey techniques from top publication venues while maintaining a striking balance between the theoretical concepts and their practical importance.

6. AUTHOR BIOGRAPHY

Sayan Ranu: Sayan is a researcher at IBM Research India since 2012. Prior to joining IBM, he finished his PhD from University of California, Santa Barbara. His current research interests include spatio-temporal data analytics, graph indexing and mining, and bioinformatics.

Aditya Telang: Aditya is a researcher at IBM Research India since 2011. Prior to joining IBM, he finished his PhD from University of Texas at Arlington. His current research interests include spatio- temporal data analytics, information management, and business analytics.

Vinay Kolar: Vinay works at IBM Research, India. He received his Ph.D. from State University of New York at Binghamton, and was a researcher at Carnegie Mellon University, Qatar and RWTH-Aachen University, Germany. His research interests are in the design of reliable and efficient networking systems.

7. REFERENCES

[1] Movebank, http://www.movebank.org.

[2] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A Tale of One City: Using Cellular Network Data for Urban Planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[3] L. Chen and R. Ng. On the marriage of edit distance and Lp norms. In *VLDB*, 2004.

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

[5] E. Frentzos, K. Gratsias, and Y. Theodoridis. Index-based Most Similar Trajectory Search. In *ICDE*, pages 816–825, 2007.

[6] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.

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

[8] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet. In *ASPLOS X*, pages 96–107, 2002.

[9] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 593–604, New York, NY, USA, 2007. ACM.

[10] Z. Li, B. Ding, J. Han, and R. Kays. Swarm: mining relaxed temporal moving object clusters. *Proc. VLDB Endow.*, 3(1-2):723–734, Sept. 2010.

[11] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 637–646, New York, NY, USA, 2009. ACM.

[12] M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst.*, 27(3):267–289, Nov. 2006.

[13] M. Passe-Smith. Exploring Local Tornado Alleys for Predictive Environmental bibtex. In *ESRI User Conference*, 2006.

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

[15] H. Zang and J. C. Bolot. Mining call and mobility data to improve paging efficiency in cellular networks. In *MobiCom*, 2007.