Bringing Semantics to Spatiotemporal Data Mining: Challenges, Methods, and Applications

Chao Zhang
Department of Computer Science
Univ. of Illinois at Urbana-Champaign
Urbana, Illinois, USA
czhang82@illinois.edu

Quan Yuan
Department of Computer Science
Univ. of Illinois at Urbana-Champaign
Urbana, Illinois, USA
qyuan@illinois.edu

Jiawei Han
Department of Computer Science
Univ. of Illinois at Urbana-Champaign
Urbana, Illinois, USA
hanj@illinois.edu

***Abstract*—The pervasiveness of GPS-equipped mobile devices has been nurturing an unprecedented amount of semantics-rich spatiotemporal data. The confluence of spatiotemporal and semantic information offers new opportunities for extracting valuable knowledge about people's behaviors, but meanwhile also introduces its unique challenges that render conventional spatiotemporal data mining techniques inadequate. Consequently, mining semantics-rich spatiotemporal data has attracted significant research attention from the data mining community in the past few years. In this tutorial, we start with reviewing classic spatiotemporal data mining tasks and identifying the new opportunities introduced by semantics-rich spatiotemporal data. Subsequently, we provide a comprehensive introduction of existing techniques for mining semantics-rich spatiotemporal data, covering topics including spatiotemporal activity mining, spatiotemporal event discovery, and spatiotemporal mobility modeling. Finally, we discuss about the limitations of existing research and identify several important future directions.**

I. Introduction

With the ubiquitous connectivity of wireless networks and the wide proliferation of GPS-enabled mobile devices, recent years have witnessed an explosive growth of spatiotemporal records that are enhanced with rich semantic information (*e.g.*, Point-of-Interest category, text message, tweet). For example, in location-based social networks (*e.g.*, Facebook Place, Twiter, Instagram, Foursquare), a user could create geo-tagged posts from anywhere at any time. Each typical post consists of a location, a timestamp, and a text message, providing a unified *where-when-what* view regarding the user's activity. Around ten million geo-tagged tweets are being created in the Twitterverse every day, and more than eight billion check-ins have been accumulated on Foursquare so far. Meanwhile, the advance of geo-tagging techniques has largely facilitated annotating Web pages, news, and other resources with GPS information, resulting in a massive amount of geo-tagged online documents.

The emergence of such semantics-rich spatiotemporal (abbreviated as SRST onwards) data offers new opportunities to understanding people's behaviors in the physical world. First, due to the confluence of spatiotemporal and semantical information, our knowledge about an individual is no longer limited to her physical location, but also what she is doing on the scene. Second, comparing with traditional spatiotemporal data like GPS trajectories and survey data, it is easier and cheaper to obtain large-scale SRST data because of the massive users, enabling us to obtain reliable knowledge about user behaviors and thus improve various prediction and decision making tasks.

Nevertheless, mining SRST data is by no means trivial because it introduces several unique challenges, which renders many classic spatiotemporal mining techniques inadequate. Those challenges can be summarized as follows: 1) *Mining reliable knowledge from the noisy and sparse data.* Real-life SRST data are often highly noisy and sparse. For instance, prior studies have revealed that a large amount of tweets are just random babbles from the users; and even those meaningful tweets are typically short. Hence, how to extract useful knowledge from such noisy and sparse data is challenging; 2) *Integrating multi-modal data.* There are three different factors involved in SRST data: location, time, and text. Those heterogeneous factors are highly coupled to reflect people's activities in a collective way, yet they have totally different modes, magnitudes, and distributions. How to effectively integrate those different data types for knowledge acquisition remains the second challenge; and 3) *Developing scalable and real-time mining methods.* For most real-life applications, the power of the SRST data can be fully unlocked only when a sheer amount of them is used. Furthermore, many applications such as event detection call for methods that can handle streaming SRST data continuously. Accordingly, the final challenge is to design methods that can handle large-scale and/or streaming SRST data in an efficient way.

Because of the unique opportunities and challenges brought by the emergence of SRST data, the past few years have witnessed extensive research for mining SRST data. In this tutorial, we provide a comprehensive review for existing techniques and systems developed for mining SRST data for useful knowledge. We categorize existing SRST data mining techniques into three classes: 1) spatiotemporal activity mining; 2) spatiotemporal event discovery; and 3) spatiotemporal mobility modeling. We will introduce the major goals and challenges in each class, as well as how existing methods approach them. Finally, we discuss about several interesting future directions for mining SRST data.

II. Indended Audience and Duration

The target audience of this tutorial are typical ICDE attendees who are interested in spatial databases, spatiotemporal data mining, text mining, and data integration. This tutorial is expected to last for 1.5 hours. It will provide the audience

with a comprehensive coverage of the major techniques developed for SRST mining. We anticipate this tutorial to enable researchers to identify the connection between this area and their own interests to inspire important future research, and help practitioners learn about the most recent advances in this area to improve real-life location-based systems.

III. Tutorial Outline

Our 1.5-hour tutorial consists of the following five sections: 1) introduction; 2) spatiotemporal activity mining; 3) spatiotemporal event discovery; 4) spatiotemporal mobility modeling; and 5) summary and future directions. Each section is expected to last for 15-20 minutes. At the end of each section, a two-minute question and answering session is provisioned. Figure 1 provides an overview of the tutorial outline. In the following, we elaborate the content of each section.

1. Introduction (15 min)
   1.1. Motivations
      1.1.1. Classic spatiotemporal data mining
      1.1.2. SRST data: definitions and characteristics
      1.1.3. Opportunities of incorporating semantics
   1.2. SRST data mining overview
      1.2.1. Challenges
      1.2.2. Methods
      1.2.3. Applications

2. Spatiotemporal Acvitiy Mining (20 min)
   2.1. Semantic annotation
   2.2. Geographical topic modeling
   2.3. Personalized activity profiling

3. Spatiotemporal Event Discovery (20 min)
   3.1. Spatiotemporal event detection
      3.1.1. Batch detection methods
      3.1.2. On-line detection methods
   3.2. Spatiotemporal event forecasting

4. Spatiotemporal Mobility Modeling (20 min)
   4.1. Pattern-based approaches
   4.2. Model-based approaches

5. Summary and Future Directions (15 min)
   5.1. Summary of existing SRST data mining methods
      5.1.1. Principles and techniques
      5.1.2. Advantages and limitations
   5.2. Future directions
      5.2.1. Handling data sparsity
      5.2.2. Scalable and on-line mining
      5.2.3. Representation learning for heterogeneity integration

Fig. 1. An outline of the tutorial.

A. Introduction

We begin our tutorial with introducing the unique opportunities and challenges brought by the emergence of semantics-rich spatiotemporal data. Specifically, we first provide a retrospect of traditional spatiotemporal data mining tasks such as trajectory classification and clustering [10], [11], [22], [23], co-location mining [56], [34], [26], [13], [15], periodicity detection [27], [28], and sequential movement pattern discovery [29], [5], [21], [12], [38], [25].

Subsequently, we introduce the characteristics of the SRST data generated from different data sources (*e.g.*, location-based social networks, geo-tagged online documents), and explain why such SRST data sets shed light on improving existing spatiotemporal data mining tasks and even enabling tasks that were almost impossible years ago.

Finally, we provide a brief overview of the existing techniques that are recently developed for mining SRST data. We divide such techniques into three categories: a) spatiotemporal activity mining; b) spatiotemporal event discovery; and c) spatiotemporal mobility modeling.

B. Spatiotemporal Activity Mining

The task of spatiotemporal activity mining has been defined as discovering the people's typical activities in different geographical regions and/or time periods. Accurate activity mining can not only provide an intuitive understanding about people's behaviors in the physical world, but also facilitate various downstream applications such as location prediction and activity recommendation.

There have been a handful of methods that are recently developed for spatiotemporal activity mining based on SRST data. We categorize these methods into three categories: the first is semantic annotation [2], [41], [40], [42], [50], which aims to infer the semantics for geographical objects (*e.g.*, POIs, regions, trajectories); the second is geographical topic modeling methods [30], [36], [39], [33], [32], [18], [44], [46], which are designed for extracting crowd-level topics that are representative for different geographical regions; and the third is personalized activity profiling [8], [58], [14], [47], which unveils personalized spatiotemporal activities for different individuals. Those methods are underpinned by various techniques, including topic modeling, kernel density estimation, Dirichlet process, and Gaussian mixture model. For each of the categories, we summarize the main ideas behind those developed techniques and also demonstrate the suitable applications for them.

C. Spatiotemporal Event Discovery

A spatiotemporal event (*e.g.*, protest, crime, disaster, sport game) is an unusual activity bursted in a local area and within specific duration while engaging a considerable number of participants. While accurately detecting spatiotemporal events was almost impossible years ago due to the lack of reliable data sources, the explosive growth of SRST data sheds light on it because of its sheer size and multi-dimensional nature.

In this section, we summarize existing works that leverage large-scale SRST data for detecting and forecasting spatiotemporal events. For the task of event detection, several pioneering studies [24], [20], [35], [4], [19], [16] have attempted to extract interesting spatiotemporal events on a given static SRST data set in a batch manner, with various techniques such as wavelet transform, spike detection, and comparative analysis. More recently, a number of studies [1], [31], [9], [52] have studied the problem of detecting spatiotemporal events in an on-line manner. We will introduce the key ideas of the studies along both directions and describe their characteristics.

In addition to spatiotemporal event detection, there has also been considerable research [53], [6], [55], [54] on forecasting

spatiotemporal events that may happen in the future, we will explain the insights behind the design of these techniques and discuss about their advantages as well as limitations.

D. Spatiotemporal Mobility Modeling

Spatiotemporal mobility modeling aims at unveiling the regularities underlying human movements. The user information contained in most SRST data sets allows us to extract the records for each user and order them by time. Accordingly, a large number semantic trajectories can be obtained from the raw SRST corpus — each is a sequence of timestamped locations where every location is described by semantic information (*e.g.*, category label, text).

The massive semantic trajectories extracted from SRST data serve as an unprecedentedly valuable source for understanding people's moving behaviors. Compared with conventional GPS trajectory data, we now have access to the rich semantics about the users' activities at different locations and timestamps.

In this section, we introduce state-of-the-art spatiotemporal mobility modeling methods that utilize SRST data. Such methods can be divided into two lines. The first line [3], [45], [28], [42], [57], [49] is *pattern-based*, which extracts pre-defined movement patterns from the semantic trajectory database to reflect the regularities underlying people's movements. Examples of such movement patterns include sequential movement patterns and periodic patterns. The second line [7], [47], [17], [43], [37], [51], [48] is *model-based*, which relies on statistical models (*e.g.*, Hidden Markov Model) to summarize people's moving behaviors. We will give a comprehensive introduction to the major techniques developed in each class, explain how they integrate multiple data types (location, time, text) for mobility regularity mining, and discuss about how to choose different methods for different scenarios.

E. Future Directions

In the final session of the tutorial, we conclude by reviewing the presented research tasks and also introduce several important future research directions for mining semantics-rich spatiotemporal data. In particular, we identify three directions that are currently attracting much research attention and worth future research efforts: 1) developing effective and intelligent algorithms that can alleviate SRST data sparsity; 2) designing scalable and on-line methods that can efficiently handle massive and streaming SRST data; and 3) learning general representations that effectively integrate the multiple dimensions in SRST data.

IV. Related Tutorials

There are no earlier versions of this tutorial presented elsewhere. Two related tutorials that were previously given by the third instructer are listed as follows:

1. Zhenhui (Jessie) Li, Fei Wu, Jiawei Han, "Trajectory Data Mining", (to appear in) the 2016 IEEE International Conference on Big Data, Washington D.C., USA, December 2016.
2. Jiawei Han, Zhenhui Li, and Lu An Tang, "Mining Moving Object, Trajectory and Traffic Data", 2010 International Conference on Database Systems for Advanced Applications, Japan, April 2010.

Those two tutorials are quit different from the current one in that they focus on mining traditional spatiotemporal data without semantic information. In contrast, all the techniques presented in this tutorial aim to discover knowledge from the newly emerged semantics-rich spatiotemporal data.

Biographies

Chao Zhang is a Ph.D. candidate at the Department of Computer Science, University of Illinois at Urbana-Champaign. His research focuses on knowledge discovery from text-rich spatiotemporal data and heterogeneous data mining. He has won the 2015 ECML/PKDD Best Student Paper Runner-up Award, the Microsoft Star of Tomorrow Excellence Award, and the Chiang Chen Overseas Graduate Fellowship.

Quan Yuan is a postdoctoral research associate of the Department of Computer Science at University of Illinois at Urbana-Champaign. He received his Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore in 2015. His research interests include spatio-temporal data mining, recommender systems, and text mining.

Jiawei Han is an Abel Bliss Professor at the Department of Computer Science, UIUC. His research areas encompass data mining, data warehousing, database systems, and information networks, with over 700 publications. He is Fellow of ACM, Fellow of IEEE, Director of IPAN (2009-2016), supported by Network Science Collaborative Technology Alliance program of the U.S. Army Research Lab, and the coDirector of KnowEnG: a Knowledge Engine for Genomics, one of the NIH supported Big Data to Knowledge (BD2K) Centers.

References

[1] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013.

[2] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macêdo, B. Moelans, and A. A. Vaisman. A model for enriching trajectories with semantic geographical information. In *GIS*, page 22, 2007.

[3] L. O. Alvares, V. Bogorny, B. Kuijpers, B. Moelans, J. A. Fern, E. D. Macedo, and A. T. Palma. Towards semantic trajectory knowledge discovery. *Data Mining and Knowledge Discovery*, 2007.

[4] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*, pages 438–441, 2011.

[5] H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. In *ICDM*, pages 82–89, 2005.

[6] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *KDD*, pages 1166–1175. ACM, 2014.

[7] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[8] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1):3, 2011.

[9] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *ICDE*, pages 1561–1572, 2015.

[10] Z. Fu, W. Hu, and T. Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *ICIP*, pages 602–605, 2005.

[11] S. Gaffney and P. Smyth. Trajectory clustering with mixtures of regression models. In *KDD*, pages 63–72, 1999.

[12] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339, 2007.

[13] J. Gudmundsson and M. J. van Kreveld. Computing longest duration flocks in trajectory data. In *GIS*, pages 35–42, 2006.

[14] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[15] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *SSTD*, pages 364–381, 2005.

[16] W. Kang, A. K. H. Tung, W. Chen, X. Li, Q. Song, C. Zhang, F. Zhao, and X. Zhou. Trendspedia: An internet observatory for analyzing and visualizing the evolving web. In *ICDE*, pages 1206–1209, 2014.

[17] Y. Kim, J. Han, and C. Yuan. Toptrac: Topical trajectory pattern mining. In *KDD*, pages 587–596, 2015.

[18] C. C. Kling, J. Kunegis, S. Sizov, and S. Staab. Detecting non-gaussian geographical topics in tagged photo collections. In *WSDM*, pages 603–612, 2014.

[19] J. Krumm and E. Horvitz. Eyewitness: Identifying local events via space-time signals in twitter feeds. In *SIGSPATIAL*, 2015.

[20] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):72, 2012.

[21] A. J. T. Lee, Y.-A. Chen, and W.-C. Ip. Mining frequent trajectory patterns in spatial-temporal databases. *Inf. Sci.*, 179(13):2218–2231, 2009.

[22] J.-G. Lee, J. Han, X. Li, and H. Gonzalez. Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proceedings of the VLDB Endowment*, 1(1):1081–1094, 2008.

[23] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *SIGMOD*, pages 593–604, 2007.

[24] R. Lee, S. Wakamiya, and K. Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349, 2011.

[25] Y. Li, J. Bailey, L. Kulik, and J. Pei. Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases. In *ICDM*, 2013.

[26] Z. Li, B. Ding, J. Han, and R. Kays. Swarm: Mining relaxed temporal moving object clusters. *PVLDB*, 3(1):723–734, 2010.

[27] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *KDD*, pages 1099–1108, 2010.

[28] Z. Li, J. Wang, and J. Han. Mining event periodicity from incomplete observations. In *KDD*, pages 444–452, 2012.

[29] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung. Mining, indexing, and querying historical spatiotemporal data. In *KDD*, pages 236–245, 2004.

[30] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[32] S. Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM*, pages 281–290, 2010.

[33] S. Sizov. Latent geospatial semantics of social media. *ACM Transactions on Intelligent Systems and Technology*, 3(4):64, 2012.

[34] L.-A. Tang, Y. Zheng, J. Yuan, J. Han, A. Leung, C.-C. Hung, and W.-C. Peng. On discovery of traveling companions from streaming trajectories. In *ICDE*, pages 186–197. IEEE, 2012.

[35] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *ECIR*, pages 356–367. 2013.

[36] C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *GIR*, pages 65–70, 2007.

[37] J. Wang, M. Li, J. Han, and X. Wang. Modeling check-in preferences with multidimensional knowledge: A minimax entropy approach. In *WSDM*, pages 297–306. ACM, 2016.

[38] L. Wang, K. Hu, T. Ku, and X. Yan. Mining frequent trajectory pattern based on vague space partition. *Knowl.-Based Syst.*, 50:100–111, 2013.

[39] P. Wang, P. Zhang, C. Zhou, Z. Li, and G. Li. Modeling infinite topics on social behavior data with spatio-temporal dependence. In *CIKM*, pages 1919–1922. ACM, 2015.

[40] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang. Semantic annotation of mobility data using social media. In *WWW*, pages 1253–1263, 2015.

[41] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semitri: a framework for semantic annotation of heterogeneous trajectories. In *EDBT*, pages 259–270, 2011.

[42] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology*, 4(3):49, 2013.

[43] H. Yin, X. Zhou, Y. Shao, H. Wang, and S. Sadiq. Joint modeling of user check-in behaviors for point-of-interest recommendation. In *CIKM*, pages 1631–1640. ACM, 2015.

[44] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.

[45] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng. Semantic trajectory mining for location prediction. In *GIS*, pages 34–43, 2011.

[46] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *KDD*, pages 186–194, 2012.

[47] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *KDD*, pages 605–613, 2013.

[48] Q. Yuan, W. Zhang, C. Zhang, X. Geng, G. Cong, and J. Han. PRED: periodic region detection for mobility modeling of social media users. In *WSDM*, pages 263–272, 2017.

[49] C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta. Splitter: Mining fine-grained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment*, 7(9):769–780, 2014.

[50] C. Zhang, K. Zhang, Q. Yuan, H. Peng, Y. Zheng, T. Hanratty, S. Wang, and J. Han. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *WWW*, 2017.

[51] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han. Gmove: Group-level mobility modeling using geo-tagged social media. In *KDD*, 2016.

[52] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *SIGIR*, 2016.

[53] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In *SDM*, volume 15, pages 963–971. SIAM, 2015.

[54] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *KDD*, pages 1503–1512, 2015.

[55] L. Zhao, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In *KDD*, 2016.

[56] K. Zheng, Y. Zheng, N. J. Yuan, and S. Shang. On discovery of gathering patterns from trajectories. In *ICDE*, pages 242–253, 2013.

[57] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua. Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology*, 3(3):56, 2012.

[58] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You are where you go: Inferring demographic attributes from location check-ins. In *WSDM*, pages 295–304, 2015.