

Mining Spatiotemporal Knowledge: Methodologies and Research Issues *

(A position paper)

Jiawei Han
Department of Computer Science
University of Illinois
Urbana, IL 61801
hanj@cs.uiuc.edu

ABSTRACT

Geographic information systems have proceeded to the stage that vast amount of spatiotemporal data has been captured in the form of data files, images, and data streams by various kinds of data collection tools, including cameras, sensors and mobile devices. It poses great challenges to data analysts to mine such data in order to extract spatiotemporal knowledge for traffic control, weather forecasting, wildfire control, disease spread watching, homeland security, and many other important applications.

In this position paper, we will discuss general methodologies for mining spatiotemporal data and pose some interesting research issues, with the emphasis on multidimensional analysis of spatiotemporal data, finding dynamics in data streams, mining for classification, clustering, outlier analysis, frequent pattern or correlation analysis, and the discovery of sequential patterns for moving objects and changing environments.

1. INTRODUCTION

Owing to the great progress of computer and sensor technologies, huge amount of geographic and spatiotemporal data has been collected by various kinds of data gathering tools, including video cameras, satellites, sensors, mobile devices, and so on, and the speed for doubling the volume of data has been increasing rapidly. Such kinds of data are often in the form of data files, geo-referenced digital imagery, and data streams. Because of the huge volume of data and its fast transition speed, it poses great challenges to data analysts for mining such data to extract spatiotem-

*The work was supported in part by U.S. National Science Foundation NSF-IIS-02-09199 and NSF-IIS-0308215, the University of Illinois, and an IBM Faculty Award. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2003 KDVis Workshop X-XXXXX-XX-X/XX/XX ...\$5.00.

poral knowledge in many important applications, including traffic control, weather forecasting, wildfire control, disease spread watching, homeland security, etc. Thus research into efficient and real-time spatiotemporal data mining is one of the most important themes in research into geographic information systems and spatial databases [14, 22, 8, 16, 15].

Here we present our view on the methodologies of spatiotemporal data mining and a few important research problems, which may promote more in-depth discussions in this workshop as well as in other forums.

2. MAJOR SPATIOTEMPORAL DATA MINING TASKS

Spatiotemporal data mining is to find patterns, outliers or interesting knowledge related to both space and time, i.e., changes related to geo-spatial locations. Typical examples including mining for regularities and irregularities of moving objects and devices (including cell phones, in-vehicle navigation systems and wireless Internet clients and so on), changing thematic maps due to construction, forest fire, etc., changing temperature and other measures within certain geographic locations, the congestion of highway networks due to traffic jams, etc. Based on our observation, spatiotemporal data mining may cover a wide spectrum of data mining activities, include the following major tasks.

1. **Multidimensional analysis of spatiotemporal data.** Since spatiotemporal data often carries multidimensional information such as time, location, properties of moving objects or thematic maps, and so on, it is crucial to summarize the data in multi-dimensional space and perform multidimensional analysis [5]. Such analysis could be performed in the context of both spatial data warehousing and spatial data mining. Since powerful analysis may often involve high-dimensional space, classical data cube model may encounter difficulties in both time and space to compute and store such high-dimensional aggregations, new technology needs to be developed for high-dimensional fast analysis of spatiotemporal data.
2. **Mining dynamics in spatiotemporal data streams.** With remote sensing and other tools for dynamic collection of spatiotemporal data, a lot of data is in the form of data streams, which are in huge volume, or potentially infinite, ordered based on their incoming timestamps,

flowing in and out in fast speed, changing dynamically, and may only be able to be scanned once. This calls for the development of fast and online stream data mining methods. Although there have been a lot of studies recently on stream data processing and mining, such as [3, 6, 7, 9, 4, 1], there are not many studies on mining the dynamics in spatiotemporal data streams. We believe this is an essential new task in knowledge discovery from spatiotemporal data.

3. **Spatiotemporal data classification.** The spatiotemporal data classification process takes classified spatiotemporal data as the training set, construct model using various kinds of classification methods, which can then be used to predict new examples [11]. For example, one may use the volumes of data related to forest fire to build up models to predict at what kinds of conditions and in which areas are likely to caught forest fire, and sufficient precaution and preparation should be taken to avoid serious damages caused by such fires.
4. **Trend prediction for spatiotemporal data.** Similarly, one can build up predication models based on trend analysis using spatiotemporal data [8]. For example, one may predict the spread of a disease to different regions based on the geographic locations, highway networks, temperature, wind velocity, time, and many other factors using regression and other predictive modeling methods.
5. **Spatiotemporal data clustering.** It is a popular data mining practice to cluster spatial data [21, 2]. However, it is often more important to cluster data with time taken as an important factor. For example, clustering automobiles or other moving objects, not only based on their location but also based on their moving directions and speed may help disclose important common behavior of moving objects.
6. **Spatiotemporal outlier analysis.** Outlier analysis has been an essential theme in data mining [12]. This is to disclose some strange moving objects or those changing rather differently in space, direction, speed, and time. For fighting against terrorism, it could be useful to find and examine those outlier objects which move rather differently from others.
7. **Frequent pattern or correlation analysis for spatiotemporal data.** There are many applications that may like to find correlated spatiotemporal objects, i.e., find a group of objects that change directions, speed, and geographic locations in a similar way. One may discover the inherent linkages among those objects and disclose some common useful properties.
8. **Discovery of sequential patterns for moving objects and changing environments.** Among many objects, one may find some regularity of location changes with time, such as if A moves eastward, likely B will be moving eastward within 30 minutes. Such patterns can be considered as sequential patterns related to both time and space. Discovery of sequential patterns [17] for moving objects or changing environments may help prediction of the trends and discovery of outliers, and so on. Thus sequential pattern mining for spatiotemporal data is another important mining task.

3. GENERAL METHODOLOGIES FOR SPATIOTEMPORAL DATA MINING

There have been a lot of data mining and data analysis methods developed for mining relational, text and multimedia data. However, spatiotemporal data has many unique features, and some new and distinct methods may need to be developed for effective mining.

The mining needs to consider the following factors,

1. Efficiency and scalability, because of handling huge volumes of spatiotemporal data,
2. fast response time in many cases since many discovery need to take timely response, such as earth quake, forest fire, tornado, terrorist attack, and other emergency cases,
3. enforcing of spatial and time constraints, since the explanations without considering such constraints may not make sense, and
4. geographic and time proximity of the answers, since a lot of answers, even approximate in time and location, could be quite useful if obtained within a confined time frame.

Based on the above considerations, we propose some general methodologies of data analysis which could be useful for effective and efficient spatiotemporal data mining

1. **Progressive deepening in spatiotemporal data mining.** Owing to massive data, sophisticated spatiotemporal analysis programs, and the requirement of fast response time, one needs to consider mining should proceed in a progressive way, first mining at a rough scale to estimate potential patterns, find interesting hot-spots, and facilitate the detailed examination of interesting portions of data. Such a mining methodology can be considered as “multi-resolution” model but such model may not be simply based on the data, but be based on knowledge mined or data/model interaction. Our previous studies at mining spatial association rules [13], spatial classification [11], spatial clustering [1], adopt this methodology and achieve high efficiency and high mining quality.
2. **Tilted time window modeling in spatiotemporal stream data mining.** Since it is difficult to store the vast amount of the data at different time frames for stream data analysis but it is important to take the historical aspects of data into consideration in stream data mining, it is important to use various kinds of tilted time window models in stream data analysis. The general philosophy of tilted time window is to register the general history of data across the full spectrum of time span but with different resolution scales. In most cases, one may put more emphasis in most recent events (thus at the finest resolution) and less emphasis on more remote data (thus at coarser resolution). The remote events will be compressed in the statistical summary form to give room for more detailed analysis of the recent data. However, comparative analysis can still be performed effectively with user adjusted weights on historical data. Our recent studies on stream data analysis, such as [6, 10, 1] have adopted such a modeling method with satisfactory results.

3. Micro-clustering could be a scalable and effective pre-processing technique in spatiotemporal stream data mining. Micro-clustering [21, 2] which groups closely related data points as one micro-cluster and thus effectively reduces the number of points to be analyzed in many other data mining tasks can be considered as a powerful preprocessing technique. Our study in constraint-based clustering [19, 18], stream clustering [1] and scalable support vector machine algorithm [20] have demonstrated the effectiveness of this method.

4. CONCLUSIONS

Spatiotemporal data mining is a promising research frontier in both geographical information system and data mining. It poses many challenging research problems and promising applications. In this position paper, we have outlined a few important tasks for spatiotemporal data mining and discussed some interesting mining methodology. As a data mining researcher, I would like to work together with the experts on geographic information systems, spatial databases, and visual data analysis to make good contributions to this exciting research frontier.

5. REFERENCES

- [1] C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03)*, Berlin, Germany, Sept. 2003.
- [2] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, pages 49–60, Philadelphia, PA, June 1999.
- [3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.
- [4] S. Babu and J. Widom. Continuous queries over data streams. *SIGMOD Record*, 30:109–120, 2001.
- [5] Y. Bédard, T. Merrett, and J. Han. Fundamentals of geospatial data warehousing for geographic knowledge discovery. In H. J. Miller and J. Han, editors, *Geographic Data Mining and Discovery*, pages 53–73. Taylor and Francis, 2001.
- [6] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 323–334, Hong Kong, China, Aug. 2002.
- [7] P. Domingos and G. Hulthen. Mining high-speed data streams. In *Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00)*, pages 71–80, Boston, MA, Aug. 2000.
- [8] M. Ester, H.-P. Kriegel, and J. Sander. Spatial data mining: A database approach. In *Proc. 1997 Int. Symp. Large Spatial Databases (SSD'97)*, pages 47–66, Berlin, Germany, July 1997.
- [9] M. Garofalakis, J. Gehrke, and R. Rastogi. Querying and mining data streams: You only get one look (a tutorial). In *Proc. 2002 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'02)*, page 635, Madison, WI, June 2002.
- [10] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu. Mining frequent patterns in data streams at multiple time granularities. In H. Kargupta, A. Joshi, K. Sivakumar, , and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, 2003.
- [11] J. Han, A. K. H. Tung, and J. He. SPARC: Spatial association rule-based classification. In R. L. Grossman, et al. (eds.), *Data Mining for Scientific and Engineering Applications*, pages 461–486, Kluwer Academic Publishers, 2001.
- [12] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 392–403, New York, NY, Aug. 1998.
- [13] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 1995 Int. Symp. Large Spatial Databases (SSD'95)*, pages 47–66, Portland, Maine, Aug. 1995.
- [14] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis: London, UK, 2001.
- [15] P. Rigaux, M. O. Scholl, and A. Voisard. *Spatial Databases: With Application to GIS*. Morgan Kaufman, 2001.
- [16] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.
- [17] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proc. 5th Int. Conf. Extending Database Technology (EDBT'96)*, pages 3–17, Avignon, France, Mar. 1996.
- [18] A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-based clustering in large databases. In *Proc. 2001 Int. Conf. Database Theory (ICDT'01)*, pages 405–419, London, U.K., Jan. 2001.
- [19] A. K. H. Tung, J. Hou, and J. Han. Spatial clustering in the presence of obstacles. In *Proc. 2001 Int. Conf. Data Engineering (ICDE'01)*, pages 359–367, Heidelberg, Germany, April 2001.
- [20] H. Yu, J. Yang, and J. Han. Classifying large data sets using svm with hierarchical clusters. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, Washington, D.C., Aug. 2003.
- [21] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'96)*, pages 103–114, Montreal, Canada, June 1996.
- [22] S. Zhou and C. B. Jones. Design and implementation of multi-scale spatial databases. In *Proc. 2001 Int. Symp. Spatial and Temporal Databases (SSTD'01)*, Redondo Beach, CA, July 2001.