

Research Issues in Spatio-temporal Data Mining

A white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, Nov. 18-20, 2003.

Xiaobai Yao
Department of Geography, University of Georgia
Room 204 GG Bldg. Athens, GA 30602
Phone: 706-583-0326, Email: xyao@uga.edu

Introduction

The availability of gigantic volume of geospatial data, often continually updated (e.g. remote sensing data), has greatly challenged our ability to digest the data and to gain useful knowledge that would otherwise be lost. Currently, most research efforts on data mining in geospatial data take the static view of geospatial phenomena, which captures “spatiality” only. However, as all geographic phenomena evolve over time, both “spatiality” and “temporality” is central to our understanding of geographic process and events. In addition, knowledge extracted from spatio-temporal data will help us to have better prediction of spatial processes or events. It is therefore very important to conduct research on data mining in spatio-temporal datasets.

Parallel to the Koperski et al (1998)’s definition of spatial data mining, spatio-temporal data mining here refers to the extraction of implicit knowledge, spatial and temporal relationships, or other patterns not explicitly stored in spatio-temporal databases. It is a subfield of data mining and knowledge discovery in databases (DM and KDD), a research area that started in computer science and information technology in the last decade and now penetrates into almost all data-rich environments. In geographic information science, the study of spatio-temporal data mining has started off recently (Roddick and Lees 2001). Both the “spatial” and the “temporal” prefixes have added substantial complexity to data mining tasks. As the discussion of “spatial data mining” have been appeared earlier in literature (Koperski et al. 1996, 1998, Shekhar et al. 2001), I will focus more on the temporality part of spatio-temporal data mining. Among other issues, this white paper identifies the following two research needs for efficient and effective knowledge discovery in spatio-temporal data.

1 “Spatialization” and “Temporalization” of Data Mining Techniques

Spatio-temporal data mining represents the confluence of several fields including spatio-temporal databases, machine learning, statistics, geographic visualization, and

information theory. Exploration of *spatial* data mining (Koperski et al. 1998, Miller and Han 2001) and *temporal* data mining (Roddick and Spiliopoulou 2002) has received much attention independently in KDD and DM research community. Nevertheless, the need to investigate both “spatial” and “temporal” relations at the same time complicates the data mining tasks even further. A crucial challenge in spatio-temporal data mining is the exploration of efficient methods due to the large amount of spatio-temporal data and the complexity of spatio-temporal data types, data representation, and spatial data structure.

First of all, spatial and temporal relationships exist among spatial entities at various levels (scales). One of the purposes of spatio-temporal data mining is to reveal such relationships. The spatial relations, both metric (such as distance) and non-metric (such as topology, directions, shape, etc.), and temporal relations (such as before or after) may be explicit or implicit in the geographic databases. In both cases, such relationships are information bearing and therefore need to be considered in the mining techniques. Secondly, spatial and temporal dependency and heterogeneity are intrinsic characteristic of spatio-temporal databases. Thirdly, scale effect in space and time is a challenging research issue in geographic analysis. Scale, in terms of spatial resolution or temporal granularity, can have direct impacts on the kinds and/or strength of relationships that can be identified in the datasets. All in all, the unique characteristics of spatio-temporal databases require significant modification of many data mining techniques that are otherwise only good for non-geographic databases.

Spatiality and temporality are two unique dimensions in geography. Recently proposed space-time data models (e.g. Peuquet and Duan 1995, Yuan 1997) integrate time and space as the primary dimension(s) of data for spatio-temporal relations to exist and for spatio-temporal processes to take place, while other attributes are subordinate to the integrated spatio and temporal dimensions. However, conventional data mining methods in artificial intelligence do not recognize the uniqueness of spatial and temporal dimensions. Most current data mining techniques applied to geographic datasets generally use very simple representations of geographic objects and spatial relationship (Buttenfield et al. 2000). The data mining techniques should be modified so that they can, to the largest extent, exploit the rich spatial temporal relationships/patterns embedded in the datasets.

Spatial data mining tasks and techniques can be roughly classified (Fayyad et al. 1996, Miller and Han 2001) into five categories including segmentation, dependency analysis, deviation and outlier analysis, trend discovery, and generalization and characterization. Table 1 applies this classification to spatio-temporal data mining tasks. The last column on the right indicates the need of temporal extensions of most current spatial data mining techniques. Some extensions are being investigated outside of the GIScience community (e.g. temporal association rules in the KDD research communities) without the consideration of “spatiality,” others remain to be studied by GIScientists. This table is open for further discussions.

Table 1. A possible classification of spatio-temporal data mining tasks and techniques (modified from Fayyad et al 1996 and Miller and Han 2001)

Spatio-temporal Data mining task	Descriptions	Techniques	
		Static spatial data	Spatio-temporal data
Segmentation	Clustering Classification	<ul style="list-style-type: none"> ❖ Cluster analysis ❖ Bayesian classification ❖ Decision tree ❖ Artificial neural networks 	<ul style="list-style-type: none"> ❖ temporal extensions to clustering ❖ Temporal extensions to classification
Dependency analysis	Finding rules to predict the value of some attribute based on the value of other attributes over time	<ul style="list-style-type: none"> ❖ Association rules ❖ Bayesian networks 	<ul style="list-style-type: none"> ❖ Temporal Association rules ❖ Temporal extension to Bayesian networks
Deviation and outlier analysis	Finding data items that exhibit unusual deviations from expectations Clustering and other data mining methods	<ul style="list-style-type: none"> ❖ Clustering and other data mining methods ❖ Outlier detection 	Temporal extension to techniques in the left column
Trend Discovery	Prediction of lines and curves Summarizing the database, often over time Discover correlations among the events in sequences	<ul style="list-style-type: none"> Discovery of common trends ❖ Regression 	<ul style="list-style-type: none"> ❖ Sequence mining
Generalization and characterization	Compact descriptions of the data	<ul style="list-style-type: none"> ❖ Bayesian networks ❖ Attribute-oriented induction 	Temporal extension to techniques in the left column

2. Spatial-temporal data representation and infrastructure

Four broad categories of temporality within data are classified in a review of temporal knowledge discovery (Roddick and Spiliopoulou 2002). A similar taxonomy can be applied to the temporality in spatio-temporal data as follows.

- Static (time has to be traced by external information such as database construction, etc.),
- Sequences (ordered list of events, reveals relationships such as before and after, or the richer relationships described as meets, overlaps, contemporary of , etc.),
- Timestamped (a timed sequence of static data taken at more or less regular intervals),
- Fully temporal (integrated spatio-temporal data, e.g. via events, processes, etc.).

The modeling of temporality of geographic data has received significant attention especially in the past decade. Earliest model takes snapshots approach which is still predominant in most commercial GIS packages. Later models include those that use space-time cube (Hagerstrand 1970), space-time composite, and base state with amendments (Langran 1992). Most recent studies in this line advocate for modeling spatio-temporal phenomena via events (Claramunt and Theriault 1995, Peuquet and Duan 1995, Chen and Jiang 1998), activities (Wang and Cheng 2001), processes (Yuan 1997), or the evolution of geographic features/objects (Usery 1996, Wachowicz 1999, Hornsby and Egenhofer 2000).

Developing spatio-temporal data mining methods and developing the spatial data infrastructure should come hand in hand for efficient and effective spatio-temporal data mining. Each spatio-temporal representation approach and the corresponding data structure(s) may impose some unique challenges to the data mining algorithms/methods. In addition, different approaches are able to embed different levels of temporality, ranging from static to fully temporal. Given appropriate data mining methods, the knowledge extracted from a spatio-temporal dataset is highly dependent on the spatio-temporal data types, representation, data structure that are used in the database.

Errors and uncertainty are facts of life in all information systems. It could be worse when the underlying dataset is obtained from heterogeneous data sources. The problem may present itself as unexpected errors, data conflicts, and so on. Although data conflicts occurs in non-spatial data mining as well (Fan et al. 2001), the conflicts reconciling strategies proposed in the non-spatial data mining systems are not readily applicable for a spatio-temporal datasets. This is because that some spatial relations and temporal relations can be misrepresented as a result of data. Part of the solution rest in cleaner data infrastructure and data integration checking tools. Another critical research need in response to the problem is to develop mechanisms to test and validate spatio-temporal data mining results, particularly that test the validity of spatial and temporal relations, and to reconcile discrepancies in data.

References

- Buttenfield, B., Gahegan, M., Miller, H., Yuan, M. (2001), *Geospatial data mining and Knowledge Discovery*. UCGIS white paper on Emergent Research Themes.
- Chen, J. and Jiang, J. (1998), An event-based approach to spatio-temporal data modeling in land subdivision systems. *Geoinformatica*. 2:387-402.
- Claramunt, C. and Theriault, M. (1995), Managing time in GIS: An event-oriented approach. In J.Clifford and A. Tuzhilin (eds), *Recent advances in temporal databases*. Berlin: Springer-Verlag, 23-42.
- Fan, W., Lu, H. Madnick, S.E., Cheung, D. (2001), Discovering and reconciling value conflicts for numerical data integration. *Information Systems*. 26:635-656.
- Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P. (1996), From data mining to knowledge discovery: An Overview. In Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P. Ulthurusamy, R. (eds) *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA:MIT Press, 1-34.
- Hagerstrand, T., (1970), What about people in regional science? *Paper of the Regional Science Association*. 14:7-21.
- Hornsby, K. and Egenhofer, M.J. (2000), Identity-based change: A foundation for spatio-temporal knowledge representation.
- Koperski, K., Adhikary, J., and Han, J. (1996), Spatial Data Mining: Progress and challenges. *Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*. Montreal, Canada: 55-70.
- Koperski, K, Han, J., and Adhikary, J.(1998) Mining Knowledge in Geographical Data, *Communications of ACM* , 1998
- Langran, G., (1992), *Time in geographic information systems*. London:Taylor & Francis.
- Miller, H.J. and Han, J. (2001), Geographic data mining and knowledge discovery: an overview. In Miller, H.J. and Han, J. (eds) *Geographic data mining and knowledge discovery*. London, New York : Taylor & Francis, 3-32.
- Peuquet, D.J. and Duan, N. (1995), An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International journal of Geographic Information systems*. 9:7-24.
- Roddick, J.F. and Lees, B.G. (2001), Paradigms for spatial and spatio-temporal data mining. In H.G. Miller and J. Han (eds), *Geographic Data Mining and Knowledge Discovery*. London: Taylor & Francis.

Roddick, J.F. and Spiliopoulou, M. (2002), A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and data engineering*. 14(4):750-767.

Shekhar, S., Huang, Y. Wu, W., Lu, C.T., and Chawla, S., (2001), What's spatial about spatial data mining: three case studies. In R.Grossman, C. Kamath, V. Kumar, and R. Namburu (eds), *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers.

Usery, E.L. (1996), A feature-based geographic information system model. *Photogrammetric Engineering & Remote Sensing*. 62: 833-838.

Wachowicz, M. (1999), *Object-oriented design for temporal GIS*. London: Taylor & Francis.

Wang, D. and Cheng, T. (2001), A spatio-temporal data model for activity-based transport demand modeling. *International Journal of Geographic Information Science*. 15:561-585.

Yuan, M. (1997), Use of knowledge acquisition to build wildfire representation in geographical information systems. *International Journal of Geographic Information Science*. 11:723-745.