A SURVEY OF SPATIAL, TEMPORAL AND SPATIO-TEMPORAL DATA MINING

R. Geetha
Senior Lecturer
Department Of B.Sc (CT)
K.S.Rangasamy College of Technology
Tiruchengode

N. Sumathi
Senior Lecturer
Department Of B.Sc (CT)
K.S.Rangasamy College of Technology
Tiruchengode

Dr. S. Sathiyabama
Professor
Department Of M.C.A
K.S.Rangasamy College of Technology
Tiruchengode

Abstract

Spatio-temporal data sets are often very large and difficult to analyze and display. Since they are fundamental for decision support in many application contexts, recently a lot of interest has arisen toward data-mining techniques to filter out relevant subsets of very large data repositories as well as visualization tools to effectively display the results. In this paper we propose a data-mining system to deal with very large spatio-temporal data sets. With the growth in the size of datasets, data mining has recently become an important research topic and is receiving substantial interest from both academia and industry.

This short paper provides a few comments on this research and provides a relevant research papers investigating temporal, spatial and spatio-temporal data mining.

Keywords: Data Mining, Temporal Data Mining, Spatial Data Mining, Spatio-Temporal Data Mining

1. Introduction

The development of data mining has naturally led to the exploration of application domains within which data mining may be used. Since many of these domains have an inherently temporal or spatial context, the time and/or space component must be taken into account in the mining process, in order to correctly interpret the collected data.

2. Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. In practice, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables (or) fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data-mining activities into one of two categories:

1. Prediction data mining, which produces the model of the system described by the given data set, or
2. Descriptive data mining, which produces new, nontrivial information based on the available data set

2.1 An Architecture For Data Mining

Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

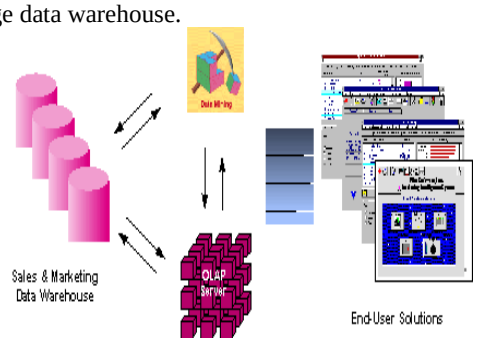

Fig.1 Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure.

2.2 Primary Data Mining Tasks

1. Classification-discovery of a predictive learning function that classifies a data item into one of several predefined classes.
2. Regression-discovery of a predictive learning function, which maps a data item to a real-value prediction variable.
3. Clustering-a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
4. Summarization-an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.
5. Dependency Modeling-finding a local model that describes significant dependencies between variables or between values of a feature in a data set or in a part of a data set.
6. Change and Deviation Detection-discovering the most significant changes in the data set.

3. Temporal Data Mining

Temporal data mining concerns the analysis of events ordered by one or more dimensions of time. We distinguish between two broad directions. One concerns the discovery of causal relationships among temporally-oriented events. The other concerns the discovery of similar patterns within the same time sequence or among different time sequences. This latter area, commonly termed time series analysis (or trend analysis) focuses on the identification of similar pre-specified patterns.

3.1 Mining Temporal Sequences

The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events. The discovery of relations between sequences of events involves mainly three steps: the representation and modeling of the data sequence in a suitable form; the definition of similarity measures between sequences; and the application of models and representations to the actual mining problems. A sequence composed by a series of nominal symbols from a particular alphabet is usually called a temporal sequence and a sequence of continuous, real-valued elements, is known as a time series.

3.2 Representation Of Temporal Sequences

3.2.1. Time-Domain Continuous Representations

A simple approach to represent a sequence of real-valued elements (time series) is using the initial elements, ordered by their instant of occurrence without any preprocessing. An alternative consists in finding a piecewise linear function able to approximately describe the entire initial sequence. The objective is to obtain a representation amenable to the detection of significant changes in the sequence.

3.2.2 Transformation Based Representations

The main idea of Transformation Based Representations is to transform the initial sequences from time to another domain, and then to use a point in this new domain to represent each original sequence. One proposal uses the Discrete Fourier Transform (DFT) to transform a sequence from the time domain to a point in the frequency domain.

A more recent approach uses the Discrete Wavelet Transform (DWT) to translate each sequence from the time domain into the time / frequency domain. The DWT is a linear transformation, which decomposes the original sequence into different frequency components, without loosing the information about the instant of the elements occurrence.

3.3 Temporal Data Mining Tasks

Data mining has been used in a wide range of applications. Temporal data mining tasks may be grouped as follows: (i) prediction, (ii) classification, (iii) clustering, (iv) search & retrieval and (v) pattern discovery. The task of time-series prediction has to do with forecasting (typically) future values of the time series based on its past samples. In order to do this, one needs to build a predictive model for the data.

4. Spatial Data Mining

4.1 Objective

The main difference between data mining in relational DBS and in spatial DBS is that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood which are used by spatial data mining algorithms.

4.2 Database Primitives For Spatial Data Mining

The set of database primitives for mining in spatial databases which are sufficient to express most of the algorithms for spatial data mining and which can be efficiently supported by a DBMS.

4.3 Efficient DBMS Support

Effective filters allow to restrict the search to such neighborhood paths “leading away” from a starting object. Neighborhood indices materialize certain neighborhood graphs to support efficient processing of the database primitives by a DBMS.

4.4 Spatial Data Mining Softwares

(1) CrimeSTAT: A Spatial Statistics Tool for the Analysis of Crime Incidents and their locations.

Description: Crimestat is a spatial statistics tool widely used by crime analysts and practioners. The program is Windows-based and interfaces with most desktop GIS programs. The purpose is to provide supplemental statistical tools to aid law enforcement agencies and criminal justice researchers in their crime mapping efforts. (2) Spatial Statistics Toolbox for Matlab/Fortran by K. Pace : Free software

Description: Matlab and Fortran toolbox for computing simultaneous and conditional spatial autoregressions and mixed regressive spatially autoregressive models, produced by K. Pace of the Dept. of Finance at Lousiana State University.

(3) Spatial Econometrics Library for Matlab: Free software

Description: It's a full set of function expansion for spatial analysis, especially including spatial autoregressive modeling.

(4) ClusterSeer/BoundarySeeer/SpaceStat by TeraSeer : Commercial software

Description: The TeraSeer softwares support spatial clustering, spatial autocorrelation analysis, k-function, and classification.

5. Spatio-Temporal Data Mining

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.

5.1 "Spatialization" And "Temporalization" Of Data Mining Techniques

Spatio-temporal data mining represents the confluence of several fields including spatio-temporal dtabases, machine learning, statistics, geographic visualization, and information theory. First of all, spatial and temporal relationships exist among spatial entities at various levels (scales). 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. 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

5.2. The Spatio-Temporal Data-Mining Process

The data-mining process usually consists of three phases, or steps: (1) pre-processing or data preparation; (2) modeling and validation; and (3) post-processing or deployment. During the first phase, the data may need some cleaning and transformation according to some constraints imposed by some tools, algorithms, or users. The second phase consists of choosing or building a model that better reflects the application behavior. Finally, the third step consists of using the model, evaluated and validated in the second phase, to effectively study the application behavior.

5.3 Spatial-Temporal Data Representation And Infrastructure

Four broad categories of temporality within data are classified in a review of temporal knowledge discovery.

- Static (time has to be traced by external information such as database construction),
- Sequences (ordered list of events, reveals relationships such as before and after, or the richer relationships described as meets, overlaps, contemporary),
- 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).

6. Result

Spatial and temporal databases are an exciting and rapidly advancing field and we have outlined above a few areas we consider worthwhile for doctoral candidates just starting their research career. This paper aims at bringing together researchers and practitioners of spatial, temporal and spatio-temporal data mining.

7. Conclusion

These huge collections of spatio-temporal data often hide possibly interesting information and valuable knowledge. It is obvious that a manual analysis of these data is impossible and data mining might provide useful tools and technology in this setting. Spatio-temporal data mining is an emerging research area that is dedicated to the development of novel algorithms and computational techniques for the successful analysis of large spatio-temporal databases.

References

[1] Abraham, T. and Roddick, J.F. 1997. 'Research issues in spatiotemporal knowledge discovery'. In Proc. SIGMOD'97 Workshop on Data Mining, Arizona, USA. ACM Press. 85.

[2] Nectaria Tryfona and Christian S. Jensen, Conceptual Data Modeling for Spatio-temporal Applications, GeoInformatica 3:3, 245-268 (1999)

[3] Tamas Abraham and John F. Roddick, Survey of Spaio-Temporal Databases, GeoInformatica 3:1, 61-99 (1999).

[4] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2000.

[5] Cristian S. Jensen, Introduction to Temporal Database Research, Temporal Database Management, 2000.

[6] T. Hadzilacos, Y. Manolopoulos, J.F. Roddick, and Y. Theodoridis, eds. Advances in Spatial and Temporal Databases. Proceedings of the Eighth International Symposium on Spatial and Temporal Databases, SSTD 2003, Santorini Island, Greece, volume 2750 of Lecture Notes in Computer Science. Springer, Berlin, 2003.