Data Mining and Visualization of Twin-Cities Traffic Data

S. Shekhar, C. T. Lu, S. Chawla, P. Zhang
Computer Science Department, University of Minnesota
200 Union Street SE, Minneapolis, MN-55455
[*shekhar, ctlu, chawla, pusheng*]@cs.umn.edu TEL:(612) 6248307 FAX:(612)6250572
http://www.cs.umn.edu/Research/shashi-group

August 2, 2000

Abstract

Data Mining(DM) is the process of extracting implicit, valuable, and interesting information from large sets of data. As huge amounts of data have been stored in traffic and transportation databases, data warehouses, geographic information systems, and other information repositories, data mining is receiving substantial interest from both academia and industry. The Twin-Cities traffic archival stores sensor network measurements collected from the freeway system in the Twin-Cities metropolitan area. In this paper, we construct a traffic data warehousing model which facilitates on-line analytical processing(OLAP), employ current data mining techniques to analyze the Twin-Cities traffic data set, and visualize the discoveries on the highway map. We also discuss some research issues in mining traffic and transportation data.

Keywords: Data mining, spatial-temporal data mining, traffic data, visualization

1 Introduction

Data Mining(DM) is the process of extracting implicit, valuable, and interesting information from large sets of data. Data mining is a technology developed from the confluence of research in machine learning, pattern recognition, statistics, database systems, and data visualization. As transportation agencies and companies collect huge amount of data concerning their operations and systems, data mining is expected to play an important role in analysis. Data mining will allow agencies and companies to extract useful information from these databases, thus making more effective decisions. For Intelligent Transportation System(ITS), which deals with the sensors, instrumentation, communication and control of the transportation system, data mining techniques are particularly important. Current ITS data is primarily used for real-time decision making. It has not been extensively and systemically applied for long-term data analysis and decision-making.

In this paper, we formulate a general framework for mining transportation data; provide some practical mining techniques and examples using Twin-Cites traffic data archival; and list some data mining opportunities in ITS for further exploration.

The rest of the paper is organized as follows. Section 2 introduces our application domain and some basic concepts of data warehousing and data mining. In section 3, the design model and OLAP operations for data warehousing are discussed. Section 4 demonstrates our mining examples using visual display. The further research opportunities related to traffic data mining applications are discussed in Section 5. Finally, we summarize our work in Section 6.

2 Basic Concepts

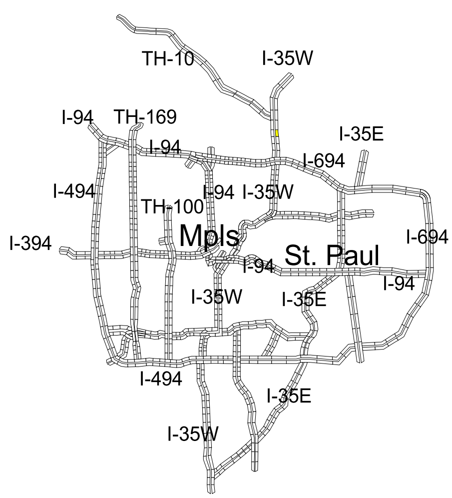
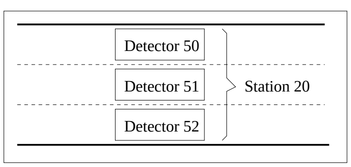


Figure 1: Detector map in station level

2.1 Application Domain: Twin-Cities Traffic Data

In 1995, the University of Minnesota, the Traffic Management Center(TMC) Freeway Operations group started the development of a database to archive sensor network measurements from the freeway system in the Twin Cities. The sensor network includes about nine hundred stations, each of which contains one to four loop detectors, depending on the number of lanes, as shown in Figure 2(a) Sensors embedded in the freeways and interstate monitor the occupancy and volume of traffic on the road. At regular interval, this information is sent to the Traffic Management Center for operational purposes, e.g., ramp meter control, as well as research on traffic modeling and experiments. Figure 1 shows a map of the stations on highways within the Twin-Cities metropolitan area, where each polygon represents one station. The interstate freeways include I-35W, I35E, I-94, I-394, I-494, and I-694. The state trunk highways include TH-100, TH-169, TH-212, TH-252, TH-5, TH-55, TH-62, TH-65, and TH-77. I-494 and I-694 together form a ring around the Twin-Cities. I-94 passes from East to North-West, I-35W and I-35E are in South-North direction. Minneapolis downtown is located on the intersection of I-94, I-394, and I-35W, and Saint Paul downtown is located on the intersection of I-35E and I-94.

Figure 2(b) shows the three basic data-tables for the traffic data. The *station* table stores the geographical location and some related attributes for each station. The relationship between each detector and its corresponding station is captured in the *detector* table. The *volume* table records all the volume and occupancy information within each 5-minute time slot at each particular station.



(a) Relationship between detectors and stations

Station Table

Station	Polygon_id	Polygon Boundary	Location	Freeway	Direction	Zone	.....
1	P1	(3,5),(4,10),..	26th St.	I-35W	N	Q4	
2	P3	(5,7),(6,4),....	28th St.	I-35W	N	Q4	
....							

Detector Table

Detector	Station
1	1
2	1
....	

Value Table

Time	Detector	Volume	Occupancy
1997 1 12 12:30	1	50	3
1997 1 12 12:50	2	60	12
....			

(b) Three basic tables

Figure 2: Detector-station Relationship and Basic Tables

We use Figure 3 to illustrate data flows and major modules of our system. The basic map and raw data are cleaned, transformed, and loaded into the data warehouse, which provides the multidimensional views and the OLAP operations for a variety of front end data mining analysis tools, e.g, classification, clustering, association rules. The discovered patterns or rules are then visually displayed as maps, tables, or charts for further interpretation. We describe the data warehouse component and data mining/knowledge discovery component in forthcoming subsection.

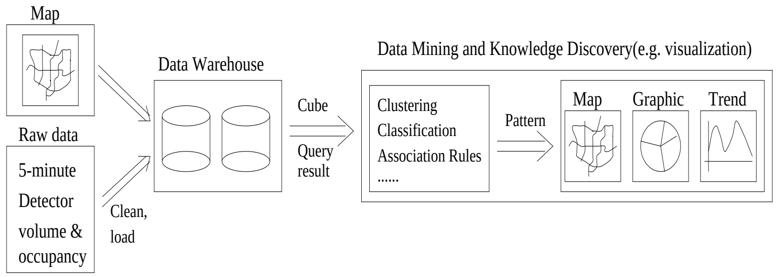


Figure 3: Data-flow and main modules in our system

2.2 Data warehouse

A data warehouse(DW) [5, 6, 22, 23, 24, 26, 25] is a repository of subject-oriented, integrated, and non-volatile information, aimed at supporting knowledge workers(executives, managers, analysts) to make better and faster decisions. Data warehouses contain large amounts of information, which is collected from a variety of independent sources and is often maintained separately from the operational databases. Traditionally, operational databases are optimized for on-line transaction processing (OLTP), where consistency and recoverability are critical. Transactions are typically small and access a small number of individual records based on the primary key. Operational databases maintain current state information. In contrast, data warehouses maintain historical, summarized, and consolidated information, and are designed for on-line analytical processing (OLAP) [12, 13]. The data in the warehouse are often modeled as a multidimensional space to facilitate the query engines for OLAP, where queries typically aggregate data across many dimensions in order to detect trends and anomalies [29]. There is a set of numeric measures that are the subjects of analysis in a multidimensional data model. Each of the numeric measures is determined by a set of dimensions. In a census data warehouse, for example, the measure is population, and the dimensions of interest include age group, ethnicity, income type, time (year), and location(census tract). Given N dimensions, the measures can be aggregated in 2^N different ways, The SQL aggregate functions and the group-by operators only produce one out of 2^N aggregates at a time. A data cube [20] operator computes all 2^N aggregates in one shot.

Spatial data warehouses, such as transportation data warehouses, contain geographic data, e.g., satellite images, aerial photography [14, 21, 28, 36], in addition to non-spatial data. Examples of spatial data-warehouses include the US Census data-set [1, 19], Earth Observation System archives of satellite imagery [41], Sequoia 2000 [39], and highway traffic measurement archives. The research in spatial data warehouses has focused on case-studies [14, 28] and on the per-dimension concept hierarchy [21]. A major difference between conventional and spatial data warehouses lies in the visualization of the results. Conventional data warehouse OLAP results are often shown as summary tables or spread sheets of text and numbers, whereas in the case of spatial data warehouse the results may be albums of maps. It is not trivial to convert the alpha-numeric output of a data cube on spatial data warehouses into an organized collection of maps. Another issue concerns the aggregate operators on geometry data types(e.g., point, line, polygon). Existing databases and the emerging standard for geographic data, OGIS [30] need to address these issues.

2.3 Data mining and Knowledge Discovery

Data mining is a process to extract implicit, nontrivial, previously unknown and potentially useful information(such as knowledge rules, constraints, regularities) from data in databases [32]. The explosive growth in data and databases used in business management, government administration, and scientific data analysis has created a need for tools that can automatically transform the processed data into useful information and knowledge. Consequently, data mining has become a research area with increasing importance [17]. The major tasks of data mining can be divided into description method and prediction method [18]. The description methods are used to find human-interpretable patterns and rules which better explaining the data; the prediction methods use some variables to predict unknown or future values of other variables.

Spatial data mining is a process of discovering interesting and useful but implicit spatial patterns. With the huge amount of spatial data obtained from satellite images, medical images, GIS, etc., it is a non-trivial task for humans to explore spatial data in detail. Spatial datasets and patterns are abundant in many application domains related to NASA, EPA, NIST, and USDOT. A key goal of spatial data mining is to partially 'automate' knowledge discovery, i.e. search for "nuggets" of information embedded in very large quantities of spatial data.

Efficient tools for extracting information from spatial data sets can be of importance to organizations which own, generate and manage large geo-spatial data sets. The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships and assuming independence between different data points. However, classical data mining methods often perform poorly on spatial data sets which have high spatial auto-correlation. Spatial auto-correlation is the property that spatially referenced objects influence other objects in the neighborhood [7]. This property is often referred to as the first law of geography: everything is related to everything else but nearby things are more related than distant things. Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Spatial statistics techniques on the other hand do take spatial autocorrelation directly into account but the resulting models are computationally expensive and are solved via complex numerical solvers or sampling based Markov Chain Monte Carlo(MCMC) methods.

Most data mining tools can process data stored in conventional databases. The construction of data warehousing is not required. However, with its support for convenient OLAP multidimensional grouping, aggregation, and queries, data warehousing is often used as a backbone for data mining modules, such as clustering, classification, and association rule discovery.

Knowledge discovery is the process of identifying valid, novel, potentially useful, and understandable patterns in data [16]. In this process, a semi-automated environment is built to discover patterns by utilizing the automated data analysis tools, such as data mining techniques. The search of new models and theories involves both the domain specific information and the guidance of domain experts. The basic components of knowledge discovery include automated interpretation, model construction, and hypothesis validation. Some related work such as advanced data visualization and discovery demonstration for a specific domain are also covered in knowledge discovery.

3 Traffic Data Warehouse

The data in the warehouse are often modeled as a multidimensional space to facilitate the query engines for OLAP, where queries typically aggregate data across many dimensions in order to detect trends and anomalies [29]. There is a set of numeric measures that are the subjects of analysis in a multidimensional data model. Each of the numeric measures is determined by a set of dimensions. In a traffic data warehouse, for example, the measures are volume and occupancy, and the dimensions are *time* and *space*. Dimensions are hierarchy by nature. In figure 4, for example, the *time* dimension can be grouped into "Hour","Date","Month","Week","Season", or "Year", which form a lattice structure indicating a partial order for the dimension. Similarly, the *Space* dimension can be grouped into "Station","County","Freeway", or "Region". Given the dimensions and hierarchy, the measures can be aggregated in different ways, The SQL aggregate functions and the group-by operator only produce one out of all possible aggregates at a time. A data cube [20] is an aggregate operator which computes all possible aggregates in one shot.

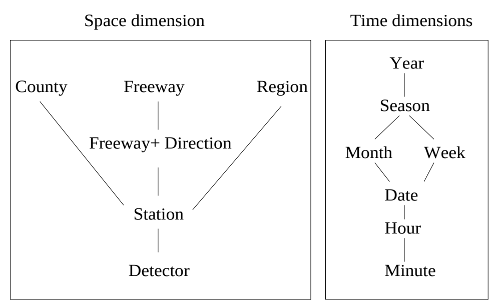
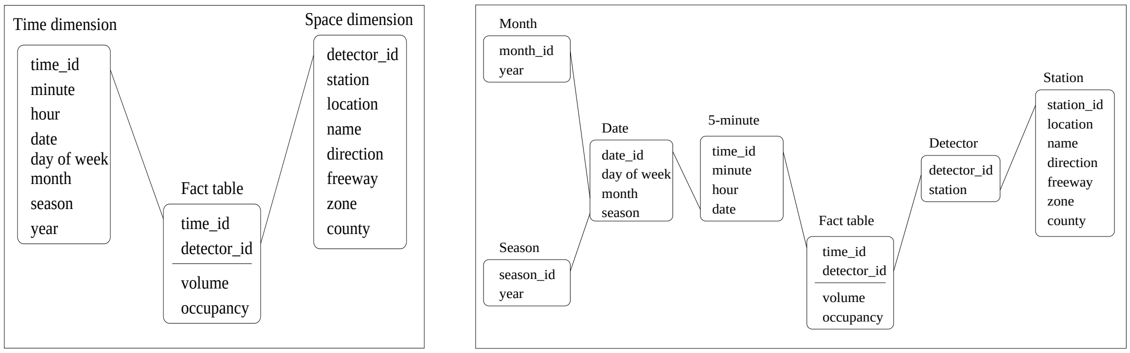


Figure 4: Concept Hierarchies for Dimensions

Most data warehouses use a star or snowflake schemas to present the multidimensional data model [5, 22, 6]. Figure 5(a) shows a star schema representation for the traffic data. The table in the center is called *fact* table, which connects to other dimension tables. In traffic data warehouse 2(b), the *volume* table can be thought of the *Fact* table. *Detector* and *Station* table together form the space dimension. The *time* dimension data can be derived from calendar models. A snowflake schema, as shown in Figure 5(b), provide a refinement of a star schema by decomposition of the dimension tables. Notice that the *Detector* and *Station* tables are separate in snowflake schema of traffic data warehouse. Aggregate structure of the dimensional tables in star schemas may be more appropriate for many analysis [6]. Accordingly, we use the star schema to construct the traffic data set. We describe aggregate functions and the data cube operator in the rest of this section.

3.1 Aggregate Function

Aggregate functions compute statistics for a given set of values within each cuboid. Examples of aggregate functions include sum, average, and centroid. Aggregate functions can be grouped



(a) A Star Schema

(b) A Snowflake Schema

Figure 5: Design Schema

into three categories, namely, distributive, algebraic, and holistic as suggested by Gray et al [20]. We define these functions in this section and provide some examples from the GIS domain. Table 1 shows all of these aggregation functions for different data types.

	Aggregation Function		
Data Type	Distributive Function	Algebraic Function	Holistic Function
Set of numbers	Count, Min, Max, Sum	Average, Standard Deviation, MaxN, MinN()	Median, MostFrequent, Rank
Set of points, lines, polygons	Minimal Orthogonal Bounding Box, Geometric Union, Geometric Intersection	Centroid, Center of mass, Center of gravity	Nearest neighbor index, Equi-partition

Table 1: Aggregation Operations

- **Distributive**: An aggregate function F is called distributive if there exists a function G such that the value of F for an N-dimensional cuboid can be computed by applying a G function to the value of F in $(N+1)$-dimensional cuboid.

- **Algebraic**: An aggregate function F is algebraic if F of an N-dimensional cuboid can be computed using a fixed number of aggregates of the $(N+1)$-dimensional cuboid.

- **Holistic**: An aggregate function F is called holistic if the value of F for an N-dimensional cuboid cannot be computed from a constant number of aggregates of the (N+1)-dimensional cuboid.

The computation of aggregate functions has graduated difficulty. The distributive function can be computed from the next lower level dimension values. The algebraic function can be computed from a set of aggregates of the next lower level data. The holistic function needs the base data to compute the result in all levels of dimension.

3.2 Cube Operator

The cube operator [20, 37] generalizes the histogram, cross-tabulation, roll-up, drill-down, and sub-total constructs. It is the N-dimensional generalization of simple aggregate functions. Table 2 is the *base* table for the traffic data cube. This is produced by joining the dimension tables with *Fact* table in a star schema. Recall traffic data warehouse star schema(Figure 5(a)). The *base* table has columns for all attributes of *fact* table and various dimension tables. The key column/attribute for *base* table are the same as those for the *fact* table. Due to the lack of space, Table 2 omits a few columns, e.g., season, day of week.

There are two dimensions: *time* and *space*, and two measure: *volume* and *occupancy*. Each dimension has its corresponding attributes. The lower attributes can be aggregates to higher or derived attributes according to its concept hierarchies as shown in Figure 4.

Space Dimension							Time Dimension							Measures	
Det_id	Stat_id	Freeway	Direction	Zone	County	...	Time_id	Yr	Mn	Day	Hr	Min	...	Vol.	Occu.
1	1	I-35W	N	6M	Hennepin	..	1	1997	1	15	6	20	..	80	15
2	1	I-35W	N	6M	Hennepin	..	1	1997	1	15	6	20	..	75	12
3	10	I-35W	S	8P	Hennepin	..	1	1997	1	15	6	20	..	65	18
4	10	I-35W	S	8P	Hennepin	..	1	1997	1	15	6	20	..	45	17
5	100	I-94	E	7L	Scott	..	1	1997	1	15	6	20	..	120	30
6	100	I-94	E	7L	Scott	..	1	1997	1	15	6	20	..	115	35
7	120	I-94	W	8Q	Ramsey	..	1	1997	1	15	6	20	..	134	25
8	120	I-94	W	8Q	Ramsey	..	1	1997	1	15	6	20	..	125	15
....	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

Table 2: The base table of traffic data cube

To support OLAP, the data cube provides the following operators : roll-up, drill-down, slice and dice, and pivot. We now define these operators.

- Roll-up: aggregate. This operator generalizes one or more dimensions and aggregates the corresponding measures. For example, Table 3 is the roll-up of the *base* Table 2 in both *space* and *time* dimension. From the dimension hierarchy in Figure 4, while *space* dimension is aggregated to *freeway* level, the *time* dimension is aggregated to *year* level. The aggregate function used in this example is the Average on a daily basis.

Space Dimension	Time Dimension	Measures	
Freeway	Year	Volume(Avg. per day)	Occupancy(Avg. per day)
I-35W	1997	60,345	20.3
I-35E	1997	69,730	14.5
I-94	1997	86,782	19.5
....	..	..	..

Table 3: Example of roll-up

- Drill-down: disaggregate. It specializes in one or a few dimensions and presents low-level aggregations. For example, we drill down Table 3 in *time* dimension, adding the *month* attribute.

- Slice and Dice: selection and projection. Slicing into one dimension is very much like drilling one level down into that dimension, but the number of entries displayed is limited to that specified in the slice command. A dice operation is like a slice on more than one dimension.

Space Dimension	Time Dimension		Measures	
Freeway	Year	Month	Volume(Ave. per day)	Occupancy(Avg. per day)
I-35W	1997	1	55,340	23.3
I-35W	1997	2	65,645	10.1
I-35W	1997	3	68,395	24.3
...				
I-35E	1997	12	85,345	20.9
I-35E	1997	1	55,375	24.3
I-35E	1997	2	62,335	12.1
I-35E	1997	3	70,945	23.0
...	..	..	..	..

Table 4: Example of drill-down

- Pivoting: re-orienting the multidimensional view of data. It presents the measures in different cross-tabular layouts. It is more typical of spreadsheets. For example, rows may represent months of a year. Column may represent freeways. Cells may show average volume per day. This spreadsheet would result from pivoting part of Table 4.

4 Traffic Data Mining and Knowledge Discovery

Traffic data mining is concerned with identification of non-trivial, useful, and interesting patters in traffic data. Knowledge discovery, e.g., visualization, helps traffic professionals study the patterns identified by data mining. We will examine two kinds of patterns, namely, classification and clustering, in this section.

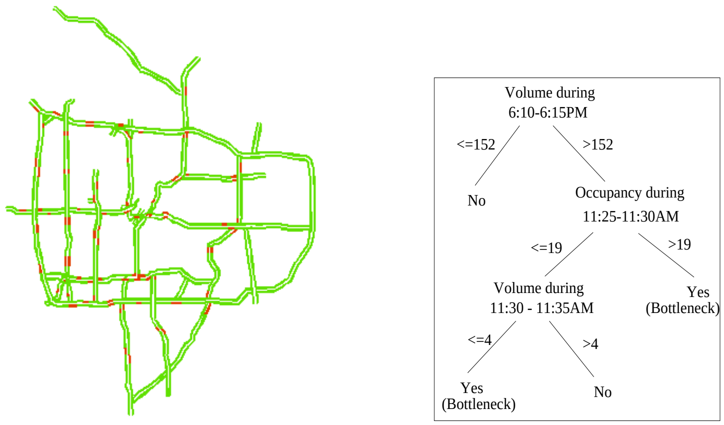
4.1 Classification

Given a set of example objects, called the training set, each of which contains n attributes(features) and one class label, the objective of classification is to analyze the training set and build a model for each class using the n attributes in the data set. The class models are then used to classify test set, which the class labels are not provided.

A decision-tree-based classification [33] is a supervised learning method that constructs decision trees from a set of training examples. In our traffic dataset, we apply the C.5 decision tree clustering algorithm [34] provided by a data mining software Clementine [9] to classify the bottleneck station, as shown in Figure 6(a). The bottleneck stations are pre-determined by MNDOT TMC. The training set is the average traffic volume and occupancy per-lane on 5-minute interval for January 15 1997, while the testing set is the traffic flow on January 17 1997. The training accuracy is 89% and testing accuracy is 87%. The decision tree for determining a bottleneck station is shown in Figure 6(b).

4.2 Clustering

Clustering, or unsupervised classification, is the process of grouping a set of abstract objects into different clusters, such that objects within a cluster are more similar to one another and objects in separate clusters are less similar to one another. Each object in the cluster contains a set of attributes. Euclidean distance is a commonly used similarity measure between data points if the attributes are continuous.



(a) Bottleneck Stations from MNDOT

(b) Decision Tree

Figure 6: A Classification Example

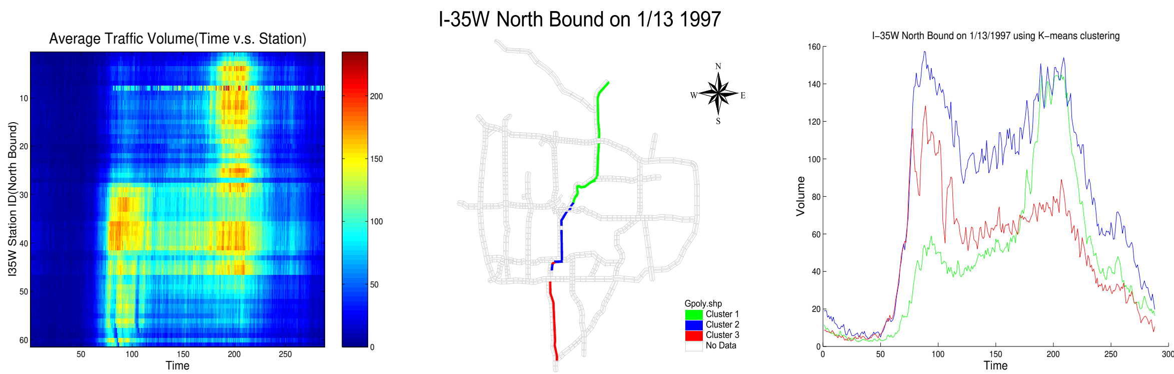
It is of interest to group stations which exhibit similar traffic flow patterns. Here, each station S_i is modeled as a point in a n-dimensional space with the form $S_i =< t_1, t_2, \ldots t_j, \ldots, t_n >$, where t_j denotes the traffic volume for station i at time j and n is the number of the time slots.

Spatial zone formation consists of segmenting traffic stations into smaller pieces that are relatively homogeneous in some sense. While these zones can be specified directly by researchers, particularly when certain areas are of interest, hierarchical clustering provides a general data mining approach for automatically creating a "zone hierarchy." The leaf nodes are the individual station, while intermediate nodes represent larger groups of contiguous stations. The regions at the lowest level represent smaller, more homogeneous regions, while regions at higher levels represent larger, but less homogeneous regions. There has been some research into clustering contiguous spatial data [7, 38], but it is still a relatively new problem. Figure 7 shows an example of clustering I-35W north bound into three homogeneous zones, where stations within each zone exhibit similar traffic pattern. Figure 7(a) is the data(average traffic volume) map of each 5-minute time slot v.s. station located in the I-35W north bound. The X-axis is the 5-minute time slot for the whole day and the Y-axis is the label of the stations installed on the highway, starting from 1 in the north end to 61 in the sound end. We will call it *Space-Time* volume map. Each station is modeled in the form $S_i =< t_1, t_2, \ldots t_j, \ldots, t_n >$, where n is 288, denoting the 5-minute time slots within one day. The clusters after applying K-means clustering algorithm are shown in Figure 7(b). The distance between two station vectors is the Euclidean distance, i.e., sum of the square of differences in volume at each time-point. We looked for three cluster of stations, primarily because the data map(Figure 7(a)) shows at least three distinct groups of stations. The results of the clustering algorithms can be visualized in attribute space (volume over time) for further knowledge discovery. Figure 7(c) shows the average volume within each cluster. The differences between each clusters can be easily observed. The stations in cluster 1 have high traffic volume during the afternoon rush hour; the stations in cluster 3 have peak traffic volume during the morning rush hour; the stations in cluster 2 exhibit high traffic volume during both morning and afternoon rush hour. Table 5 summarizes all the time periods with peak traffic volume (greater than a pre-defined threshold) within each cluster.

Comments from Sanjay

In Figure 7, we have shown three different ways of visualizing the station-time-volume relationship. This is analogous to pivoting, which is a standard way of data exploration in OLAP and data warehousing. The fact that traffic data has a spatio-temporal component allows for intuitive visualization of the different pivot operations.

Assuming that the three plots in Figure 7 are generated from traffic data of a "typical" day, an unusual but repeatable deviation from the normal pattern can help traffic managers plan for unexpected events which may disrupt the flow of traffic. For example, how will a major traffic accident, a sport game, or an entertainment event affect the composition of the three identified clusters? Such questions can be answered by plotting historical traffic data of days when such events were known to have occured. "



(a) Data(average volume) map

(b) K-means clustering using volume data

(c) Average volume within each cluster

Figure 7: I35W North Bound

	I-35W North Bound		
	Cluster 1	Cluster 2	Cluster 3
No of Station	27	18	16
Duration (minutes) greater than threshold (90)	200	740	135
Time period greater than threshold (90)	2:45PM-2:50PM, 3:10PM-6:20PM	6:20AM-10:00AM, 10:20PM-6:55PM	6:15AM-6:45AM, 7:05AM-8:35AM, 9:15AM-9:20AM

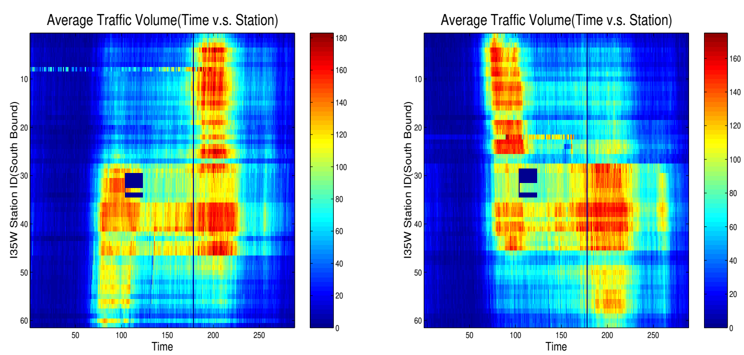
Table 5: Description of each cluster

5 Other Traffic Data Mining Opportunities

We list a few opportunities for mining interesting patterns in traffic data in this section.

5.1 Outliers/Exceptions detection

Knowledge discovery tasks can be classified into four general categories: (a) dependency detection (e.g. association rules) (b) class identification (e.g. classification, clustering) (c) class description (e.g. concept generalization), and (d) exception /outlier detection [27]. Most research has concentrated on the first three categories, which correspond to patterns that apply to a large percentage of objects in the dataset. In contrast, outlier detection focuses on a very small percentage of data objects, which are often ignored as noise. An outlier in a set of data is an observation or a point that is considerably dissimilar to or inconsistent with the remainder of the data [35]. Figure 8 shows an example of traffic flow outliers. Figure 8(a) and (b) are the volume maps for I-35W North Bound and South Bound, respectively, on 1/21 1997. The X-axis is the 5-minute time slot for the whole day and the Y-axis is the label of the stations installed on the highway, starting from 1 in the north end to 61 in the sound end. The abnormal dark blue line at time slot 177 and the dark blue rectangle during time slot 100 to 120 on X-axis and between station 29 to 34 on Y-axis can be easily observed from both (a) and (b). Moreover, station 9 in Figure 8(a) exhibits inconsistent traffic flow compared with its neighbor stations.



(a) I-35W North Bound (b) I-35W South Bound

Figure 8: An example of outlier

Most of the existing work on outlier detection lies in the field of statistics [4]. Numerous outlier tests have been developed for different circumstances, depending on: 1)the data distribution; 2) whether or not the distribution parameters are known; 3)the number of expected outliers; 4)the types of expected outliers [4]. For applying those tests to our traffic data set, there are two serious problems. First, almost all of these tests consider only one attribute, which makes them unsuitable for traffic data cube with time and space dimensions. Secondly, they do not take into account the effect of auto-correlation. Spatial auto-correlation is the property that spatially referenced objects influence other objects in the neighborhood. Time auto-correlation is the property that temporally referenced values influence other values during a time period for the same object. Spatial-temporal auto-correlation is the influence of both spatial and temporal objects. With the dynamic traffic flow, incorporating the concept of auto-correlation is expected to improve the performance of outlier detection test. We are planning to investigate the auto-correlation based algorithm for outlier detection.

Comments from Sanjay

For example, we can begin by assuming that each station is independent of each other and analyze the time-series of each station. We can then aggregate the result over space and calculate the difference between the expected result and the actual result. Part of the difference may be due to the noise present in the data. However, with the existence of spatial autocorrelation, there is an underlying spatial trend that needs to be incorporated in the analysis. There are several tests, including the Moran's I [40], Local Indicators of Spatial Autocorrelation(LISA) test [3, 11, 10], which can be used to quantify spatial autocorrelation.

Similarly, we can first calculate the spatial autocorrelation at each time-step and then plot the spatial autocorrelation as a function of time.

5.2 Association Rules Discovery

Given a set of records, with each record contains some number of items, it is desirable to discover the dependency rules such that the occurrence of an item can be predicted based on occurrences of other items. Agrawal et al. [2] proposed a formal model to define the association rules. Let $\imath = I_1, I_2, \ldots, I_m$ be a set items and T be a database of transactions. An association rule is an implication of the form $X \Longrightarrow I_j$, where X is a set of some items in $\imath$, and I_j is a single item in $\imath$ that is not present in X. The rule $X \Longrightarrow I_j$ holds in the set of transactions T with confidence c if at least $c\%$ of transactions in T that contain X also contain I_j. The support for the rule is defined as a fraction of transaction in T that contains the union of items in X and I_j. Confidence is an indication of the rule's strength and support is a measure of statistical significance.

Traditional application domain of data mining is in the analysis of transactional data. In this domain, the database system stores information about user transactions, which is a collection of items. The association rules captures the interrelationships between various items. Association rules are one possible approach for capturing long range dependencies (in space and time) hidden in traffic data. For example the following rule may be captured by the judicious use of association rule discovery: *A major traffic accident on Station A of Highway X during time $T1$ and $T2$ results in unusual high traffic volume on Station B of Highway Y during $T2+2$ and $T2+3$.* Unlike correlation analysis, the promise of association rule analysis is that such dependencies do not have to be hypothesized but are discovered automatically.

Spatial-temporal association rule extensions to the general form of rule requires the use of spatial and temporal predicates [15]. For temporal association rules, the emphasis movies from the data itself to changes in the data [8]. The discovery of temporal association rules is similar to trend analysis, which is to look for similar trends across the time axis. Given a set of spatial objects, each of which contains a set of time varying numerical or categorical attributes; and a sequence of snapshots are taken at a fixed interval. The problem of spatial-temporal association rule discovery is to find the spatial correlation among object evolutions.

For traffic data, it is common to observe the cyclic pattern, e.g., peak traffic flow during rush hour on weekdays. Ozden et al. [31] have proposed two algorithms for finding association rules which display cyclic variation over time, where users specify period(s) and segment size(s) of interest. However, their techniques focus on item set of transactions, and may not be applicable for continuous traffic data.

5.3 Sequential Pattern Discovery

Given a set of objects, with each object associated with its own timeline of events, the problem of mining sequential pattern is to find the rules that predict strong sequential dependencies among different events.

Discovering interesting patterns, e.g., correlations, from the traffic datasets is challenging due to their spatio-temporal, multiscale nature and their large size (10's of gigabytes). Sequential pattern discovery consists of two major components, namely, modeling of events and algorithms for finding spatio-temporal patterns formed by these events. Forming events is challenging due to the influence of neighboring areas on the properties of a spatial zone and the multi-scale nature of the data. Algorithms for finding patterns need to address the impact of redundancies due to overlap across neighborhoods, (which require redefinition of traditional measures of correlation and association), incorporate spatio-temporal properties and traffic domain knowledge to prune and filter uninteresting patterns, and scale to large data sets.

6 Conclusion

Data mining and knowledge discovery are fast expanding fields with many new research results and new prototypes for various application developed recently. To support intelligent transportation system using current available techniques, data mining faces many challenges and unsolved problems which pose new research opportunities for further study. In this paper, we formalize the integration of data warehousing and data mining techniques to support traffic data analysis, provide real examples by utilizing clustering and classification tools, and propose some research directions. In the future, we are planning to develop new outliers detection algorithms based on the concept of auto-correlation, and design suitable techniques for traffic data to discover cyclic association rules.

7 Acknowledgment

We are particularly grateful to Professor Vipin Kumar, and our Spatial Database Group members, Dr. Xuan Liu, Weili Wu, Hui Xiong, and Xiaobin Ma, for their helpful comments and valuable discussions. We would also like to express our thanks to Professor Stephen Strother, Department of Radiology, University of Minnesota for his insightful suggestion of the *Space-Time* volume map.

This work is supported in part by USDOT grant on high performance spatial visualization of traffic data, and is sponsored in part by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAH04-95-2-0003/contract number DAAH04-95-C-0008, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. This work was also supported in part by NSF grant #9631539.

References

[1] In *URL: http://www.census.gov.*

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining Associations between Sets of Items in Massive Databases. In *Proc. of the ACM-SIGMOD Int'l Conference on Management of Data*, pages 207–216, May 1993.

[3] L Anselin. *Spatial Econometrics: methods and models.* Kluwer, Dordrecht, Netherlands, 1988.

[4] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley, New York, 3rd edition, 1994.

[5] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. In *Proc. VLDB Conference*, page 205, 1996.

[6] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. (1):65–74, March 1997.

[7] S. Chawla, S. Shekhar, W-L Wu, and U. Ozesmi. Modeling spatial dependencies for mining geospatial data: An introduction. In *Harvey Miller and Jiawei Han, editors, Geographic data mining and Knowledge Discovery (GKD)*, 1999.

[8] X. Chen, I. Petrounias, and H. Heathfield. Discovering temporal association rules in temporal databases. In *Proceedings of the International Workshop on Issues and Applications of Database Technology (IADT98), Berlin, Germany*, 1998.

[9] Clementine. Spss inc. http://www.spss.com/clementine/.

[10] A. Cliff and J. Ord. *Spatial Autocorrelation*. London Pion Ltd., 1973.

[11] A.D. Cliff and J.K. Ord. *Spatial processes: Models and Applications*. London Pion Ltd., 1981.

[12] E.F. Codd. Twelve rules for on-line analytic processing. In *Computerworld, April 13 1995*.

[13] E.F. Codd, S.B. Codd, and C.T. Salley. Providing olap(on-line analytical processing) to user-analysts: An it mandate. In *Arbor Software Corporation. Avaliable at http://www.arborsoft.com/essbase/wht_ppr/coddToc.html*, 1993.

[14] ESRI. Spatial data warehousing. http://www.esriau.com.au/wp.htm.

[15] V. Estivill-Castro and A. Murray. Discovering associations in spatial data - an efficient medoid based approach. In *The 2nd Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD-98), Springer-Verlag*, pages 110–121, 1998.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining toward a unifying framework. In *Proc. of The Second Int. Conference on Knowledge Discovery and Data Mining*, pages 82–88, 1996.

[17] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, Cambridge, MA, 1996.

[18] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press*, pages 1–34, 1996.

[19] P. Ferguson. Census 2000 behinds the scenes. In *Intelligent Enterprise*, October 1999.

[20] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In *Proceedings of the Twelfth IEEE International Conference on Data Engineering*, pages 152–159, 1995.

[21] J. Han, N. Stefanovic, and K. Koperski. Selective materialization: An efficient method for spatial data cube construction. In *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'98)*, pages 144–158, 1998.

[22] W.H. Inmon. *Building the Data Warehouse*. New York, NY: John Wiley & Sons, 1993.

[23] W.H. Inmon and R.D. Hackathorn. *Using the Data Warehouse*. New York, NY: John Wiley & Sons, 1994.

[24] W.H. Inmon, J.D. Welch, and K.L. Glassey. *Managing the Data Warehouse*. New York, NY: John Wiley & Sons, 1997.

[25] J. Widom. Research Problems in Data Warehousing. In *Proc. of 4th Int'l Conf. on Information and Knowledge Management(CIKM), Nov. 1995.*

[26] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite. *The Data warehouse Lifecycle Toolkit*. Wiley, 1998.

[27] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. 24th VLDB Conference*, 1998.

[28] MICROSOFT. Terraserver: A spatial data warehouse. http://www.microsoft.com.

[29] I.S. Mumick, D. Quass, and B.S. Mumick. Maintenance of data cubes and summary tables in a warehouse. In *SIGMOD*, pages 100–111, 1997.

[30] OPEN GIS Consortium. OpenGIS Simple Features Specification for SQL. In *URL: http://www.opengis.org/public/abstract.html*.

[31] B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In *Proc. 1998 Int. Conf. Data Engineering (ICDE'98), Orlando, FL*, pages 412–421, Feburary 1998.

[32] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[33] J. Quinlan. Induction of decision trees. In *Machine Learning*, number 1, pages 81–106, 1986.

[34] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[35] S. Ramaswamy, R. Rastongi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Bell Laboratories, Murray Hill, NJ*.

[36] S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C.T. Lu. Spatial databases: Accomplishments and research needs. *IEEE Transactions on Knowledge and Data Engineering(TKDE)*, 11(1):45–55, 1999.

[37] S. Shekhar, C.T. Lu, X. Tan, S. Chawla, and R. R. Vatsavai. Mapcubes: A visualization tool for spatial data warehouses. In *Harvey Miller and Jiawei Han, editors, Geographic data mining and Knowledge Discovery (GKD)*, 1999.

[38] P. Stolortz, H. Nakamura, E. Mesrobian, and et al. Fast Spatio-Temporal Data Mining of Large Geophysical Datasets. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press*, pages 300–305, 1995.

[39] M. Stonebraker, J. Frew, and J. Dozier. The sequoia 2000 project. In *Proceedings of the Third International Symposium on Large Spatial Databases*, 1993.

[40] M. Tiefelsdorf and B. Boots. The exact distribution of Moran's I. *Environment and Planning(Publisher: London Pion Ltd.)*, 27:985–999, 1995.

[41] USGS. National satellite land remote sensing data archive. In *http://edc.usgs.gov/programs/nslrsda/overview.html*.