The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014)

Data Mining Strategies for Real-Time Control in New York City

H. Joon Park*, Patrick H. Kimb, Michael Marsicoa, Naim Rasheeda

*New York City Department of Transportation, 55Water Street, 6Fl, New York, NY 10041, USA
bCrossroads Systems, 42 Brown Dr., Shelton, CT, USA

Abstract

The Data Mining System (DMS) at New York City Department of Transportation (NYCDOT) mainly consists of four database systems for traffic and pedestrian/bicycle volumes, crash data, and signal timing plans as well as the Midtown in Motion (MIM) systems which are used as part of the NYCDOT Intelligent Transportation System (ITS) infrastructure. These database and control systems are operated by different units at NYCDOT as an independent database or operation system. New York City experiences heavy traffic volumes, pedestrians and cyclists in each Central Business District (CBD) area and along key arterial systems. There are consistent and urgent needs in New York City for real-time control to improve mobility and safety for all users of the street networks, and to provide a timely response and management of random incidents. Therefore, it is necessary to develop an integrated DMS for effective real-time control and active transportation management (ATM) in New York City. This paper will present new strategies for New York City suggesting the development of efficient and cost-effective DMS, involving: 1) use of new technology applications such as tablets and smartphone with Global Positioning System (GPS) and wireless communication features for data collection and reduction; 2) interface development among existing database and control systems; and 3) integrated DMS deployment with macroscopic and mesoscopic simulation models in Manhattan. This study paper also suggests a complete data mining process for real-time control with traditional static data, current real timing data from loop detectors, microwave sensors, and video cameras, and new real-time data using the GPS data. GPS data, including using taxi and bus GPS information, and smartphone applications can be obtained in all weather conditions and during anytime of the day. GPS data and smartphone application in NYCDOT DMS is discussed herein as a new concept.

Keywords: Data Mining System (DMS), New York City, real-time control, active transportation management (ATM), GPS data

* Corresponding author. Tel.: +1-212-839-7757; fax: +1-212-839-7000.
E-mail address: hpark@dot.nyc.gov

© 2014 The Authors. Published by Elsevier B.V.
Selection and Peer-review under responsibility of the Program Chairs.
1. Introduction

Private sector industries have utilized data mining technology for several decades. Data mining technology has been used for decision making, forecasting, planning, business strategies, sales forecasting, etc. Not only in business, but also in engineering or science sectors, data mining, which benefits from advancement of technology becomes a core part of advanced system construction.

Traffic and pedestrian data mining is a process by which useful information can be extracted from data sets. Its main purpose is to access useful information derived from a pool of various data for operational control. Generally, important information is not visible with a set of raw data, since raw data are not typically user friendly information, requiring the process of traffic data mining. When the size of raw traffic data are small, data mining is not necessary or it is not called data mining. Instead it is usually called analysis, traffic data processing or data manipulation. However, the huge size of traffic data or complicated data reduction process requires data mining. Real-time controls and transportation analyses rely heavily on validated data. With data mining process and simulation model applications, the useful traffic and pedestrian/bike information can be applied for efficient real-time control and management. New York City Department of Transportation (NYCDOT) has expanded its database systems in various areas such as the signal system, accident data, traffic and pedestrian/bicycle data, and developed AIMsUN macroscopic/mesoscopic/microscopic simulation models for Manhattan. Also, the Midtown in Motion (MIM) system has been introduced as a part of the active transportation management (ATM) system through adaptive controls in Midtown Manhattan since 2011. Nevertheless, these systems are operated independently as a separate database and operation system.

2. Related Works

The existing data mining system in transportation consists of a two-level traffic information process: 1) traffic information extraction from massive data sets; and 2) forecasting for future traffic conditions. A number of previous studies introduced a new hybrid intelligent data mining model for traffic information monitoring and forecasting purposes using a conceptual approach (1, 2). Over the past decade, many transportation agencies in the United States have implemented performance measurement systems to monitor real-time traffic conditions along corridors and provide data to transportation planning processes. These deployed systems collect data from various Intelligent Transportation System (ITS) devices, including Bluetooth readers, radar detectors including Remote Traffic Microwave Sensor (RTMS), GPS-based probe vehicles, loop detectors, and traffic cameras, etc. Collected traffic data is integrated with other data sources which significantly impacts traffic conditions, such as incident and weather information for in-depth traffic operational analysis. The University of California at Berkeley in cooperation with Caltrans (California DOT) developed a freeway performance measurement system (PeMS) which extracts information from real-time and historical traffic data. PeMS receives the 30-second traffic flow and occupancy data, and calculates the average vehicle speed from the collected data. The Vehicle Probe Project Suite (VPP Suite) developed by the Center for Advanced Transportation Technology Laboratory (CATT) at the University of Maryland allows transportation agencies to support operations, planning, analysis, and performance measures estimation using INRIX probe data combined with other agency transportation data.

The Moving Ahead for Progress in the 21st Century Act (MAP-21) requires measuring and improving safety, congestion, system reliability, and freight movement. To comply with this, many State Departments of Transportation and Metropolitan Planning Organizations (MPOs) adopted private probe vehicle traffic data (3). However, it still has not provided real-time traffic information dynamically during the short-interval periods for non-recurring congestion (4). In addition, there are limitations in the number of vehicle-probe drivers for metropolitan urban streets, traffic-oriented detection technologies, and a lack of crash, pedestrian and transit operation information.

3. Traffic Database Systems in New York City

NYCDOT has equipped several database systems such as signal timing repository database system, DOT Safety application including corridor and intersection crash data, the Traffic Information Management System (TIMS), and
NYCDOT has equipped several database systems such as signal timing repository database system, DOT Safety application including corridor and intersection crash data, the Traffic Information Management System (TIMS), and real-time data warehousing of the MIM project. The MIM warehouse processes real-time data while other database systems take a role as static data repository. The NYCDOT ITS infrastructure for MIM includes 103 advanced solid-state traffic controllers, 32 video cameras, 100 microwave sensors, 23 electronic toll collection readers at intersections within a 110 square block area or “zone” from 2nd to 6th Avenues, and 42nd to 57th Streets. In 2013, MIM deployment was expanded to 1st, 8th and 9th Avenues, and 42nd to 57th Streets. Fig. 1 illustrates data mining work flow of MIM. Taxi Global Positioning System (GPS) data are currently used for supplemental data of MIM for validating travel time within the aforementioned study area.

TIMS is used by NYCDOT to query, display and download traffic data within the data warehouse. The TIMS data includes 9,200 Automatic Traffic Recorder (ATR) counts, bicycle counts of 1,530 segments/intersections, pedestrian counts of 1,800 segments/intersections, 2,800 turning movement and vehicle classification counts, and 3,230 spot speeds. Phase 2 of TIMS will explore an integration of the number of injuries from the NYCDOT Safety application. Therefore, it is necessary to develop an integrated Data Mining System (DMS) for effective real-time control and ATM system in New York City.

4. Component of Traffic Data Mining

4.1. Database

Database is a storage for data containing not just data. It can contain logic and process for data manipulation, process and analysis. The current TIMS and signal timing repository database system belongs within this category. When designing a database schema as an integrated system, a core concept for the entire design scope is normalization, which distributes elements to fit to the needs of the analysis required in real-time traffic control.
Fig. 2. Example of data schema

One example schema is shown in Fig. 2. Different colors represent different sub-schema. Online Analytical Processing (OLAP) may need to run daily for further analysis or reporting service to provide various data mining capabilities. Various parts of analysis functions can be embedded in the database layer using built-in database functions.

4.2. Data Input Layer

In many cases, data comes from many manual input systems or application processes. However in traffic engineering, when the term, data is referenced, it stands for data collected from the field. Field data collected may involve vehicle, cyclist and pedestrian volumes, and travel times. Data of this type should be entered into a database using an automated system. The first step toward best practice management of data mining in traffic engineering is the development of proper methodologies in data inputs. Today’s technology provides many useful devices such as smartphone or tablet. These devices are not just devices for personal use but can be used as a data input device since it is a computer not just a gadget or a personal device. Furthermore, it is not expensive and is quite versatile in its capabilities and highly mobile compared to laptops or desktops. Development of applications for these devices is not more difficult than that for regular personal computer (PC)’s. Another benefit realized from this type of device is internet access, WIFI connection, camera and GPS. These functions allow data collection required for traffic control to be available using this type of device.

Through the use of these devices, one can collect traffic data such as vehicle volumes, signal phasing, pedestrian movements, record the video while driving in the field and easily upload the data directly to the database server from the field right after the measurement as well as entering data after transferring to PC’s from the office. In the case of smartphone, many people carry smartphone daily, which indicates that data collection is virtually always possible, easy and most importantly, inexpensive. There is no need to purchase equipment designed only for data collection.
In addition to those benefits mentioned, the data can contain GPS information, increase accuracy of data and be easily manipulated.

4.3. OLAP and Data Access Layer

OLAP tools help traffic engineers/analysts perform different analysis of multidimensional data such as time series and trend analysis of extensive periods. Database stores data in flat 2-dimensional format, OLAP can treat records in database in a format of multiple dimensions. When the OLAP service layer is in place, data access layer can provide a gateway for higher layer of services to access records in various formats and dimensions. When large data set requires analyzing with various and customizable dimensions, OLAP is a useful tool. However, the current trend in technology is to provide less heavy tools like the Tabular service of Microsoft, which can provide similar functions as OLAP but with less load on system memory.

4.4. Service Layer

The service layer resides on top of access layer or OLAP layer through which the layers which are higher in hierarchical structure can benefit dynamic and various analysis services. This layer can have multiple layers as needed and its degree of multiplicity depends on needs and complexity of the business.

4.5. Analysis Layer

Sometimes an analysis layer can be mixed with the service layer or located under the service layer. The functionality of the service layer is to provide an analysis process performed based on the data set requested to and provided from the service layer. The analysis layer includes a statistical analysis showing time period trends and summary of travel times or speeds along corridors and segments.

Fig. 3. Example of traffic data mining using hand-held device and smartphone
4.6. Presentation Layer

The actual and tangible format of the data is provided to the user via this layer. The user will view, manipulate and process the data finally provided via this layer. In advanced platforms, this layer provides highly sophisticated forms of information visualization extracted from and processed through the data. The presentation layer for traffic and pedestrian/bicycle data mining can be a dynamic simulation model such as the AIMSUN simulation model of Manhattan traffic model and MIM. An example of traffic data mining using hand-held device and smartphone is shown in Fig. 3.

5. Challenge and solutions of static data for traffic and pedestrian/bicycle data mining

The structure of real-time control system is comprised of data collection, data storage, data process, and operation control. The details of communication and data exchange among these components vary depending on the design. Data storage is mostly facilitated by the database. Data process is designed to process raw data to generate information. In the final stage, operation control is a component wherein a decision is made to manage the traffic conditions. Streamlining data collection though various sources, and operation control with data processes are the keys for a successful real time control in heavy traffic and pedestrian conditions.

5.1. Challenge

Four types of data are required for integrating data mining: vehicle and pedestrian/bicycle volumes; signal timing plans; travel time; and accident data. In most cases, signal data is obtained via two-way communication, and control of signal timing is also done via two-way communication. Besides signal timing plans, data collection or interface with other types of data should be considered carefully. Many suggestions were made regarding collecting data using conventional methodologies or somewhat advanced technologies; however, from a practical perspective, some important facts should be also examined. The communication method for traffic data transfer can be done wirelessly or through copper wire. When wireless communication is considered, the first thing to confirm is that the media for communication is available to transmit the data. No matter what methodology is used, it is necessary to transmit the data to the data center where the database is located. Currently, two ways of communication between the point of data collection and the data center are possible – internet or WIFI connection. However, this communication method is available only in limited regions.

Video data at NYCDOT are basically collected through closed circuit television (CCTV) cameras. For image processing, it should be remembered that the technology for image processing of object detection is still under development. Even though it is frequently heard that commercial products using this technology are available, there are limitations. It is mostly for static image objects like a human face and/or for the system where there are not many objects to detect. The current technology to detect a single object with a few frames of scene may take a time of range of 50~100 milliseconds. This processing time is fast enough to detect vehicle motion. For an urban intersection, the total number of vehicles in motion are sometimes over 3,000/hour or 1/sec, which indicates image processing is possible within an acceptable time range. However the above value (50~100 milliseconds) is when there are not many objects in motion in a frame. If a scene has many vehicles in motion and some of them are overlapped in the image or some are similar-looking vehicles moving in parallel, the current technology of image processing may not produce accurate data. Another limitation of image processing is it requires a sufficient amount of light for the camera to take a scene and process it for object detection. There are techniques using thermal lens or infrared (IR) lens but this technique requires special equipment and can be expensive.

5.2. Suggested Solution

As described earlier, real-time data collection of traffic volumes does not have many options except loop detector. The alternative method is to apply semi-real-time method. Smart devices can be utilized to transmit the data of vehicle volumes. The data collection using a smart device is done via human work. The data collection can be done periodically and the data is stored in data storage. The data undergoes a process to prepare a proper set of
information. Then the available real-time data is combined with a proper set of semi-real-time data of traffic, cyclist and pedestrian volume by matching the pattern intelligently. This is possible using high level techniques such as categorical analysis or statistical clustering analysis or pattern recognition in computer science.

Finally, this set of information is transferred to operation control for final decision. In order to achieve this, two facts are critical – periodic data collection and congestion index of traffic conditions. As the size of data increases the latter may require more computing power. This requirement can be satisfied using proper technology in software engineering. A conventional technique for horizontal scalability may be a fitting and easy solution but it would be costly since additional purchasing of hardware is required. On the other hand, vertical scalability is another solution with less cost but it would require more robust system design of during software development. One candidate for the latter is distributed computing. This type of technique is already available and used in commercial products or open source community.

With regard to traffic and travel time data, a simple and smart application for electronic hand-held devices can be developed and distributed to the drivers with minimal cost. Then the driver would act not just as a consumer of services provided by DMS, they can be a data provider. This application could collect travel information like time, incident location and congestion levels. It can detect connection state for WIFI. When it is connected to WIFI, it transmits the data to macroscopic and mesoscopic transportation simulation models or the real-time control system.

Several years ago, NYCDOT developed a Manhattan Traffic Modal (MTM) including Lower and Midtown Manhattan areas based on PARAMICS and AIMSUN models. This MTM development process helped to identify several issues in building a large scale model including: 1) simulation model compatibility needs; 2) tremendous manual data entry efforts; 3) inability to collect all-day traffic and pedestrian data as well as regular data updates; and 4) limited hardware platform. These lessons in the early modeling stages have directed NYCDOT to explore new traffic data sources and advanced traffic detection technologies which are capable of interfacing macroscopic and mesoscopic transportation model with effective data mining process. These models not only provide a relationship between volume or density and speed across the entire corridor and regions, but also illustrate dynamic time-series features on essential MOEs by managing many data sets quickly. NYCDOT has continuously expanded the transportation simulation model as a regional model including Bronx, Queens, Brooklyn, Staten Island and West New York in New Jersey. The streamline interfaces of data mining and macroscopic and mesoscopic models will provide dynamic and fast-responsive controls for recurring and non-recurring congestion in New York City and innovative systematic controls on region-wide transportation effects.

6. GPS application in traffic data mining

The application used for data collection requires simple analysis functions for data reduction process. This gives engineers/analysts a preview of the data, which allows an engineer/analyst to have improved planning in his/her task for the project. Some analytic layers or similar layers can be started its function as soon as new data is uploaded. For example, when the last piece of data is uploaded and stored in the database, the internal module of database can call for proper analysis performed, such as levels of service (LOS), queuing analyses and demand projections.

There are many GPS data (i.e., taxi, bus, and freight movement) available in New York City. These data require a long processing time for analysis process. Therefore the data mining with automated algorithms and step process can be scheduled and run for non-peak hour periods depending on the schedule. Engineers/analysts will save time because all the processes described up to now can be done in a very short time when data collection is complete.

Once all the necessary data is stored in the database and relevant analyses are completed, the entire system is ready for providing high level services to the users. In other words, this high level service is ready for the consumer (engineer or analyst). When the service is requested via either web or proprietary application, the service agent (the software module delegating service request from the user to the service layer) will ask for the proper service. The service layer will provide the proper data set to a service provider and/or a user. Web browser or application interface will be provided through simulation presentation layer which enables data to be presented in a more meaningful format. Other issues that application of data mining can resolve are data exchange. Sometimes it is required to exchange data between two different institutions or agencies like city DOT and state DOT. The issue in this case is the data format in two different parties may be different. The simple answer to resolve this issue is standardizing the format. Unfortunately, the cost is huge and it will take quite a long time to standardize. But if DM
is put in-place with a proper interface engine, the issue will be easily resolved. The data exchange engine is a software module which sits in the system like other service layers. It is designed to understand different formats and data structure from different agencies (city, state, etc.).

The exchange engine will request data from one agent and convert to a proper format and then deliver the data. When the data is delivered to the requester, it will be saved in a permanent or temporary place in the database and then a proper process will run the analysis, as necessary. There are many additional benefits which are not mentioned here. Generic statistical analysis, pattern and trend analysis, data visualization, optimization analysis and many other works can be integrated in the system together with data. With the help of HTML5 and WebGL technology, a simple simulation is possible. Also, since the data may contain GPS information if the device used is equipped with GPS function, geographic mapping and integration with GIS data is available within the same web platform. With DMS, engineers can access various databases not just for data but for information and analysis results.

7. Conclusions and suggestions for further research

There are many smartphone and tablets applications which are available for providing useful information such as travel time, bottleneck locations and incidents using wireless networks. It also connects drivers with real-time traffic data. Drivers using these applications take a role as real-time data sensors and detectors. They, as a data feeding community for active online data, help people to improve the quality of everyone's daily driving and commuting. Public agencies should consider utilizing cost-effective real-time data resources while providing traffic information. New strategies in NYCDOT are suggested which will aid the development of efficient and cost-effective DMS, including: 1) new technology application such as smart phone for data collection and reduction; 2) interface development among existing traditional database and control systems; and 3) integrated DMS deployment with a macroscopic or mesoscopic model to monitor traffic conditions, project future traffic patterns, visualize congestion and safety data and provide effective real-time control policy along key corridors and across regional networks.

References

4. Kim K, Motiani D; Spasovic L.N.; Dimitrijevic B; Chien SI, Assessment of Speed Information Based on Probe Vehicle Data: Case Study in New Jersey, Presented at 93rd annual meeting of the Transportation Research Board, Washington, DC, 2014
5. Midtown in motion technical memorandum #4, KLD Engineering, PC. August 2013
8. TIMS weekly report, New York City Department of Transportation, January 2014