Mining and Analysis of Traffic Safety and Roadway Condition Data

Hong Lin (linh@itsc.uah.edu)
Sudheer Cheedella (cheedes@email.uah.edu)
John Rushing (jrushing@itsc.uah.edu)
Michael Anderson (mikea@cee.uah.edu)
Ken Keiser (kkeiser@itsc.uah.edu)
Sara J. Graves (sgraves@itsc.uah.edu)

The University of Alabama in Huntsville

ABSTRACT

Decision makers working to improve transportation systems must constantly balance the need to improve roadway safety through infrastructure investment with constraints of available resources. We have explored the additional information available to transportation decision makers by integrating two independent datasets, traffic accident and roadway pavement condition data. The roadway condition data are collected by the Alabama Department of Transportation. The traffic accident data are available in the University of Alabama's Critical Analysis Reporting Environment (CARE) system. The Algorithm Development and Mining (ADaM) [Rushing] system, a data mining toolkit developed by the University of Alabama in Huntsville (UAH) Information Technology and Systems Center (ITSC) was used to examine the possible relationships between roadway data and traffic safety data. The preliminary study showed that discernable and predictable associations can be identified through analysis of the roadway condition and reported traffic accident events. The analysis results indicate that data mining is a successful method to perform advanced analysis to improve infrastructure investment decisions.

INTRODUCTION

Accidents on our nation's roadway are a serious threat to the traveling public. Although operator or operational factors are the major cause of most of automobile accidents, the road-operating environment can often itself be a potential cause-factor in an accident. Poor pavement and shoulder conditions can play a role in the occurrence of accidents. Bumps, potholes, pavement roughness and pavement edge drop-off are just a few pavement conditions that could cause difficulty for the drivers. Decreasing accidents and improving roadway condition are two potentially connected goals of any transportation administration.

Often, transportation decision makers must balance the need to improve roadway conditions with the constraints of available resources. To assist decision makers in making crucial infrastructure investment decisions, large data collection efforts have been undertaken by Alabama Department of Transportation, collecting data on accidents, pavements, bridges, construction, maintenance activities and more. As the size of these databases increases rapidly both spatially and temporally, it is quite a challenge to analysis and extract useful information from them without using advanced data analysis tools.

Data mining as an emerging new data analysis technique has received a great deal of attention in recent years due to the fast growing ability of collecting and storing data. It has been used widely to support decisions in business management, production control, market analysis, engineering design and science exploration. Up-to-date, data mining techniques have been applied to safety [Hardin, Chong] or roadway data [Amado], but not across the combination of both. [El-Seoud] applied clustering techniques in merged datasets of traffic safety and pavement conditions in analyzing the traffic safety in the state of Florida's transportation system. GIS (Geographic Information Systems) was utilized in their study to identify relevant freeway features at each accident location and to integrate them with the accident database. But they emphasize clustering accident data and consider only six pavement feature attributes: number of lanes, speed limit, local name, median width, median type and shoulder type. Many other key pavement conditions mentioned earlier were not addressed.

To fully utilize the database resources, we have developed an automatic dataset integrating process to merge the pavement condition data and traffic safety data. Advanced multivariate data analysis techniques and data mining algorithms will be used to determine whether these techniques can identify inherent roadway safety problems that may have been previously unidentified. This study can address the relationships using combinations of variables such as roadway conditions, weather conditions, and traffic patterns using classification and association data mining techniques. The pavement conditions and traffic safety data are merged spatially based upon the common key attribute of geo-location, known as the "milepost" in both datasets. Considering the pavement conditions are quite different along the different directions of the route, travel direction is also taken into consideration. In the following sections, we will first brief the data mining techniques used in the work, and then explain the data preprocessing necessary for this analytical approach. Finally we will show some case studies results with conclusions.

DATA MINING

The ever increasing tremendous amount of data, collected and stored in large and numerous data bases, has far exceeded human ability for comprehension without the use of powerful tools [Han]. Consequently, important decisions are often made based not on the information-rich data stored in databases but rather on a decision maker's intuitions due to the lack of tools to extract the valuable knowledge embedded in the vast amounts of data. This is why data mining has received great attention in recent years. Data mining refers to extracting or "mining" knowledge from large amounts of data [Han]. It can be viewed as an essential step in the process of knowledge discovery in databases. This is different from traditional statistical analysis, which typically deals with a relatively small dataset and has questions in mind when collecting the data and developing models seeking answers [Hand]. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis [Han]. General data mining principles, including Associations, Sequential Patterns, Classifications, Predictions, and Clustering, can be applied to many areas. Association rule mining finds interesting association or correlation relationships among a large set of data attributes.

ADaM (Algorithm Development and Mining) is a data mining toolkit designed and developed in the Information Technology and Systems Center at the University of Alabama in Huntsville (UAH). It provides classification, clustering and association rule mining methods that are common to many data mining systems. The toolkit consists of over 75 interoperable mining and image processing components. Each lightweight and autonomous component is provided with a C++ application programming interface (API), an executable in support of scripting tools (e.g. Perl, Python, Tcl, Shell). ADaM is extensible and scalable, and has been successfully used in several diverse data mining applications [Rushing]. ADaM is the primary data mining tool used in this pavement condition and traffic safety association study. In particular, the association rule mining algorithm in ADaM was used to discover the pavement conditions that occur frequently together in the given accident and pavement condition datasets and mine the association rules among those interesting attributes.

DATA PREPARATION

A data mining system normally consists of a sequence of steps, starting from data cleaning, data integration, data selection and transformation to data mining and knowledge presentation. In this section, these steps will be addressed in detail regarding the traffic safety data and pavement condition data. All input data is ultimately transformed into ARFF (Attribute-Relation File Format), for compatibility with ADaM and other data mining tools.

Data Sources and Description

The traffic safety data used in this study were obtained through the *CARE* system. *CARE* (Critical Analysis Reporting Environment) is a data analysis software package designed for problem identification and countermeasure development purposes for traffic safety information by the staff of the *CARE* Research & Development Laboratory (CRDL) at the University of Alabama. By querying the accident reports on certain counties, detailed accident records on each route can be extracted. This safety data is a comprehensive record with up to 231 separate attributes. The *CARE* system has accident location information recorded as milepost on the roadway with a resolution of about 0.1 miles. This is understandable considering the confusing scenario at accident sites. The pavement condition datasets were provided by the Alabama Department of Transportation. It includes the roadway pavement condition surveys of years 2000 and 2002 with up to 72 attributes. The pavement conditions include such attributes as rutting, roughness, elevation, shoulder type, etc. Since pavement data collection is automated using equipment with advance scanning technology, it has a resolution of 0.01 miles or higher.

Data Cleaning and Integration

Noise and inconsistent data are inevitable in large data collections. In the accident report, all information regarding the accident is recorded manually on site. Some important information might be missing, such as milepost information and travel direction of the involved vehicles. As mentioned earlier, milepost and travel direction are two important keys used in merging the traffic safety data and pavement condition data. Hence, the records without the geo-location information are useless in this study, and were discarded during the data cleaning process.

As mentioned in the previous section, the pavement condition data and traffic safety data are two completely independent data sets and are collected and maintained by different transportation administration groups. The same attributes may have different names. The milepost attribute in the safety data is M__POST, while in pavement conditions data it is "COND_MILEPOST". The travel direction attribute in the safety data is "DIR OF TRAVEL_VEH C", with number of 1, 2, 3 and 4 representing four directions, while in pavement condition data, it is "COND_DIRECTION" with "N, S, W, E" representing four directions respectively. All these disparities introduce difficulties in using simple queries to merge these datasets. So, we developed automated procedures to resolve these discrepancies during merging. These two datasets were merged based on the spatial location information, together with the data cleaning process. Also, due to the limited amount of usable traffic safety data, we did not consider the time-matching issue between pavement data and accident data within that year. We also realized that it is possible that the accidents could happen before the date of pavement condition surveyed for that year and the condition data might be questionable for that accident record. But since the pavement conditions are not recorded by the date of improvement, merging these two datasets based on the year is the best we can do now.

Another issue we encountered during the data preparation is the inconsistent unit of measure in the pavement datasets we obtained. The pavement condition data of year 2000 are in metric, while the 2002 data are in English. Hence, we have to analyze them separately.

Considering the use of coarse-resolution milepost information for accident location in the accident report, pavement data aggregation is also conducted during the data preprocessing. We have computed the maximum, minimum and median values of all the pavement attributes within 0.1 miles of the each accident location. The worst case among these values is used in data mining analysis, so the nearby conditions of the accident location can be considered.

Data Categorization

The attribute values of pavement conditions are in both categorical and numerical values. For example, there are 5 categories of shoulder conditions, Good, Fair and Poor. But for attributes such as Rutting values can range from 0 to 24 mm. Since our goal is to correlate pavement conditions with accident data, we will categorize numerical-valued attributes, such as IRI (International Roughness Index), patching size, elevation, etc. We analyzed all the pavement datasets available for that year, and found the minimum and maximum values of all the numerical attributes. Ten categories are used and are equally distributed from the minimum to the maximum. The following table gives an example of how the rutting values are categorized.

Value [mm]	(0, 2.4]	(2.4, 4.8]	(4.8, 7.2]	(7.2, 9.6]	(9.6, 12]	(12, 14.4]	(14.4, 16.8]	(16.8, 19.2]	(19.2, 21.6]	(21.6, 2.4]
Category	1	2	3	4	5	6	7	8	9	10

CASE STUDY ANALYSIS AND RESULTS

In our preliminary study, we used collected data for complete roadways spanning Alabama, while at the same time striving to include roadway diversity. To honor privacy concerns, specific roadway identification information are not revealed in this paper since this effort is to suggest the development of possible future decision making tools rather than focus on specific concerns at this point.

The Alabama Department of Transportation (ALDOT) provided roadway data concerning distress characteristics of pavement sections for all federal and state routes in Alabama. The data elements contained in the database included location, pavement type, pavement condition, number of lanes, lane width, and shoulder type. Since it would be difficult to study every pavement section in this limited study, we selected two US highways (indexed as route A and C), one state highway (indexed as route B) and one interstate highway (indexed as route D). Each of these routes spans more than 250 miles across the state. The traffic safety datasets for each of these routes were extracted from the CARE system. The traffic safety and pavement datasets are merged for each route using the procedures explained in previous section. The final datasets for ADaM are saved in ARFF formatted text files for use with the data mining tools.

Our preliminary analysis will compare the occurrence rate (OR for simplicity) of each value of each condition attribute among the accident locations and among the entire pavement condition dataset of each route. If the OR of an attribute's value among the accident locations is higher than its corresponding rate in the pavement condition dataset, we argued that this pavement condition might be a good candidate for pavement rehabilitation. Otherwise, that condition plays no role regarding the cause of accidents. Here we made the assumption that the pavement data are sampled evenly across each of the route.

We first applied the association rule mining to the pavement condition datasets of selected routes with cardinality of one and computed the ORs of each value (category) of each attribute in the pavement data. Then we applied the same algorithm to the merged datasets, i.e., the dataset of pavement conditions at the location of accidents. This gives the ORs of each category value of each pavement condition among the accident locations. Further analysis among these "flagged" attributes with higher ORs will help decision makers identify the location(s) with those pavement conditions for a certain route.

Rutting

Rutting, identified in the pavement database as RRUT and LRUT, is an indication of surface depressions. Rutting is a potential safety concern as ruts pool water, leading to vehicle hydroplaning, and tend to pull a vehicle towards the rut path as it is steered across the rut. Pavement with deeper ruts should be leveled and overlaid.

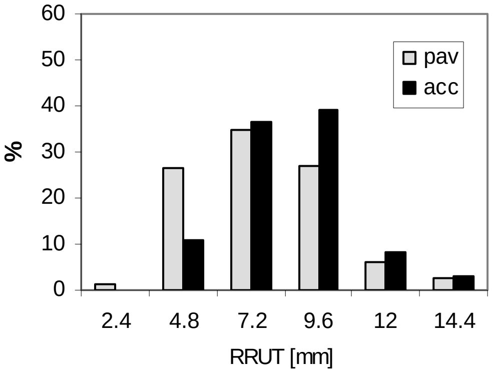

Figure 1(a) Left Side Wheel Path Rutting

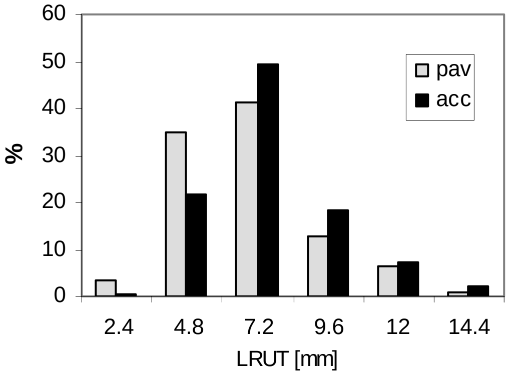

Figure 1(b) Right-side Wheel Path Rutting

Figure 1 shows the frequencies of rut values for Route A of 2000 data sets, Figure 1(a) is for right side wheel path and Figure 1(b) is for left-side wheel path. The black bars are the frequency of occurrences of each rut category among the accident locations; the gray bars are the corresponding frequency along the entire route of pavement condition records. As one can see, the frequencies of higher rut values among the accidents are much higher than the corresponding occurrence along the entire route.

Shoulder Type

Shoulder type can have big effects on automobile accidents [Web]. No shoulder or bad shoulder conditions can cause difficulty in operating vehicles in certain circumstances, and potentially lead to increase in accident severity as avoidance, or escape routes is minimized. Figure 2 shows the analysis result of 2002 data of difference shoulder type among the accident locations. Here the shoulder type of zero is considered as not recorded. Among all the shoulder types, the grass type has much higher occurrence rate (~ 37%) among the accident locations than its corresponding distribution rate (~ 20%) along the route. Grass shoulders are often found on low traffic volume, older roads that have steep hills, sharp curves, narrow pavements, etc. They are generally more susceptible to crashes.

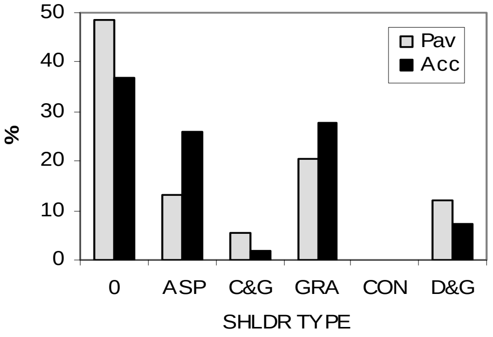

Figure 2(a) Shoulder type of route A

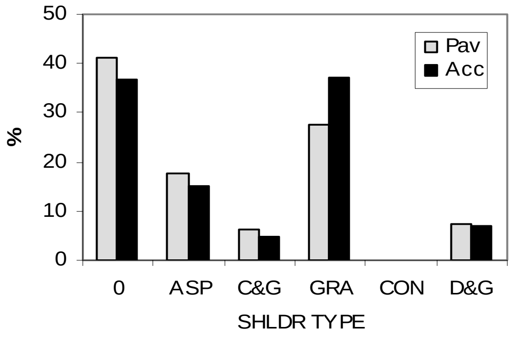

Figure 2(b) Shoulder type of route C

Cracking

Cracking is another index of the pavement condition. Figure 3 shows the analysis result of route B and C of 2002 datasets for transverse cracking on asphalt and concrete, severity 2. Although the majority of the route has a cracking count less than or equal to one, at location with more cracks, the occurrence rate of accidents (in black) are higher than the corresponding ones along the route (in gray).

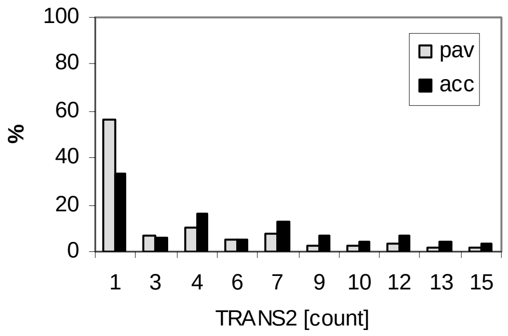

Figure 3(a). Transverse Cracking of B

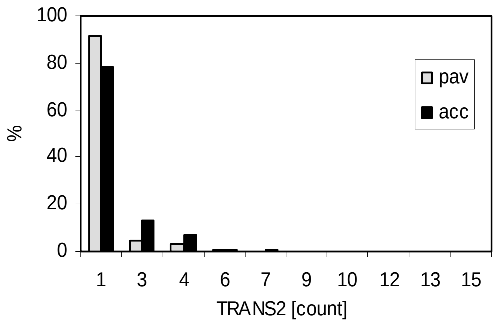

Figure 3(b) Transverse Cracking of C

Roughness

Pavement roughness is an expression of irregularities in the pavement surface that adversely affect a vehicle's ride quality. Every DOT in the US uses the International Roughness Index (IRI), to quantify roughness, and generally IRI ranges from 0 to 1267 inches/mile. Higher values indicate increased roughness, an unfavorable vehicle operating condition. Figure 4(a) and 4(b) represent the analysis results of IRI for route D of 2002 datasets. Figure 4(a) is for IRI1, the average of left IRI and Figure 4(b) is for IRI2, the average of right IRI. Route D is an interstate highway and most of IRI are in the good value range of less than 100 inches/mile. However, it is still apparent that higher roughness values have a higher occurrence among the accident locations. Higher roughness conditions could be one of the factors causing an accident or making the avoidance of an accident difficult. This is no surprise considering the higher travel speed limit on interstate highways.

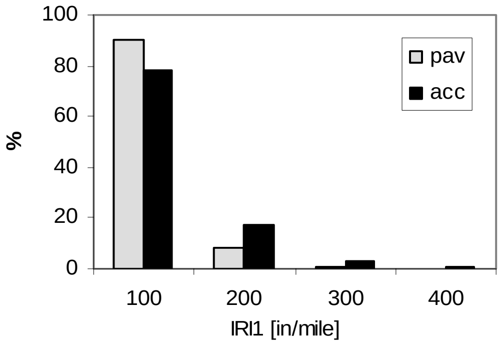

Figure 4(a) Left side of route D

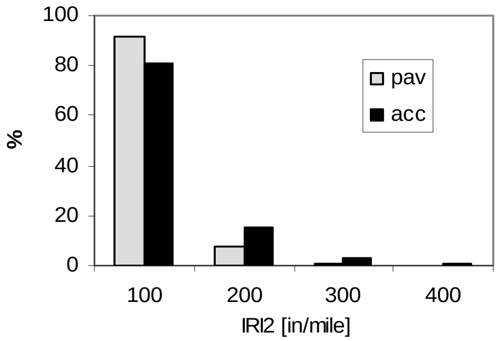

Figure 4(b) Right side of route D

CONCLUSIONS AND FUTURE WORK

This research effort involved merging and associating two different data sources, the traffic safety data and the pavement condition data, for individual routes based on geo-location information inside these databases. By applying advanced data mining techniques, we are able to identify certain pavement conditions in which there is a disproportionate number of accidents. Preliminary analysis has shown the pavement conditions whose values, or value ranges, have much higher occurrence rates among the accident locations than the occurrence rate along each of the selected routes. Here we did not consider the accident rates on each selected route due to the data problem mentioned earlier, i.e., portion of the accident records has to be discarded due to the lack of geo-location information.

Though the data analysis we have done is very preliminary at this time, the results are very encouraging, and have strongly indicated that data mining is a successful method to perform advanced analysis to traffic safety and condition data. By querying these attributes and their "risk" values among the pavement datasets, problematic locations can be identified and targeted for further analysis and evaluation by transportation experts. This approach seems capable of providing valuable information for transportation decision makers in roadway administration and maintenance.

In future and further analysis, more attributes from accident data will be considered, such as traffic volume, driver's information, weather conditions, etc. Currently, traffic volume is the strongest predictor used for high number of accidents by the ALDOT. Traffic volume, especially high volume of heavy trucks also causes accelerated deterioration of pavement conditions, such as roughness, rutting, cracking, etc. It would also be very interesting to explore if there are any pavement conditions causing more driving difficulties to one age group than others.

SUGGESTIONS

The quality of the collected data is essential to the data mining and analysis results. As mentioned in the early sections, we have encountered some problems in traffic safety data since some records have missing attribute values. This significantly reduces the amount of data for the analysis. Any record without geo-locations will be useless in roadway safety analysis. Improved methodologies of collecting accurate geo-location information at accident sites are needed.

Consistence in name and value interpretation of the attributes across different datasets can ease the data preparation process greatly. Unfortunately, inconsistence does exist due to either historic reason or others. For example, in the datasets used for this study, the attribute name for the geo-location is "M__POST" is named as "DIR OF TRAVEL_VEH C" in the safety data, and as "COND_DIRECTION" in the pavement data. Another example is when units of measure have varied between different years of pavement datasets, such as in

ACKNOWLEDGEMENTS

This project is supported and funded by the University Transportation Center for Alabama (UTCA) through Project #04310. We thank ALDOT for providing the roadway pavement condition data for this research. We also thank the University of Alabama for providing the traffic safety data through the CARE system.

REFERENCES

Amado, V. "*Expanding the Use of Pavement management Data*," 2000 MTC Transportation Scholars Conference, Ames, Iowa.

Chong, M. "*Traffic Accident Analysis Using Decision Trees and neural Networks*," IADIS International Conference on Applied Computing, Portugal, IADIS Press, Pedro, 2004.

El-Seoud, M, Elbadrawi, H. "*Data Mining and GIS Technologies to Support Highway Safety Management Systems*," IAMOT 2004.

Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, Academic Press, ISBN 1-55860-489-8.

Hand, D. "*Statistics and Data Mining: Intersecting Disciplines*," ACM SIGKDD, June 1999.

Hardin J. M. and Conerly, M. "*Traffic Safety Analyses*: *A Data Mining Approach*," UTCA Project # 02115.

Hill, S. and Jay K Lindly, J. "*Red Light Running Prediction and Analysis*," UTCA Project # 02112.

Randy K. Smith "*Data Mining to Improve Traffic Safety*," UTCA Project # 04107.

Rushing, J., et al "*ADaM: A Data Mining Toolkit for Scientists and Engineers*," Computers and Geosciences, accepted in Nov. 2004.

Web, *New York Court Finds That Defective Shoulder Caused Death in Automobile Accidents*, URL: *http://www.usroads.com/journals/rilj/9702/ri970202.htm*