Data-Mining of State Transportation Agencies Projects Databases

Khaled Nassar, PhD
Department of Construction and Architectural Engineering,
American University in Cairo

Data mining is widely used in business applications including market segmentation, fraud detection, credit risk analysis as well as many other applications. In the construction domain however, the use of data mining has been extremely limited. Data mining usually requires the availability of a large database of previous cases to be analyzed. Therefore applications in the construction industry must be geared to those situations where such databases are readily available. This paper describes a research effort to explore a potential use of data mining in the construction industry. Real data about asphalt paving projects was collected from various IDOT (Illinois Department of Transportation) sources and analyzed using data mining techniques. The results indicate that data mining can provide information beyond the use of general statistical analysis. Various rules and patterns were derived from the original database, which could be applied to support decision-making. The limitations of data mining are also noted including the need to verify and test the discovered patterns.

Key Words: Data mining, Construction Databases, State Transportation Agencies Databases, knowledge discovery

Introduction

It has been estimated that the quantity of data in the world roughly doubles every year, while the amount of meaningful information decreases rapidly [1]. Properly analyzing data and detecting these patterns is therefore of great importance to businesses. The construction industry is no exception. Construction companies collect data on a daily basis for activities and operations. Similarly, state transportation agencies (STAs) maintain their own project databases. Public or semi-public access is sometimes provided on the Internet, such as those of OSHA, FHWA and various DOTs. Data mining can provide a great tool for discovering the wealth of information contained in this data [2]. The term "KDD" is generally employed to describe the whole process of extraction of knowledge from data and the term "data mining" is often used exclusively for the discovery stage of the KDD process [1,3,4].

This paper describes a research effort undertaken to explore the applicability of data mining to a potential application in the construction industry. Data mining techniques were applied to a state transportation agency's (STA) database containing information about asphalt paving projects such as cost and schedule data. The goal was to discover any hidden rules of patterns stored within the data. In the next section a brief introduction to data mining techniques is presented. In the following section, data collected from the state of Illinois DOT's projects-database containing typical asphalt paving projects is presented and analyzed. Data mining was used to reveal unknown patterns and trends in the database of paving projects. Examples of the extracted patterns and rules are presented. Finally, the limitations and the conclusions are presented.

Data-Mining Applications in the Construction Industry

Data mining is widely applied in business applications including market segmentation, customer profiling, fraud detection, evaluation of retail promotions, credit risk analysis insurance policy, and in some military operations [5]. In the construction domain, the use of data mining techniques has been limited. Nii and Okine presented a data mining approach to pavement rehabilitation and maintenance decision support using rough set theory [6]. The data used in the study was collected from the Florida Department of Transportation (FDOT) district 6 in 1995 for flexible pavement. The conclusion was that the preliminary results indicate that the rough set theory application may well work for a PMS system. Leu et al investigated the applicability of data mining in the prediction of tunnel support stability using an artificial neural networks (ANN) algorithm [7]. Data from a railway tunnel construction in western Taiwan were used to establish the model. The main objective was to develop a neural network model for the prediction of tunnel support performance. After a thorough data cleaning, 470 records remained for the ANN analysis. The number of rock mechanical and construction related attributes used as input variable totaled 14. The data types were both numeric (directly read from rock mechanical logging and daily reports) and logical. In addition, Lucio et al. discussed the data preparation process for construction knowledge generation through knowledge discovery in databases [8], as well as construction knowledge generation and dissemination [9].

Application of Data Mining To an STA's Paving Project Database

In this paper, the application of data mining to a database containing data on construction asphalt paving operations projects was explored. The main purpose was to explore any relationship between relevant variables that might reveal hidden knowledge about the paving projects. Ideally, the most interesting relationships to be identified are those between project cost and other variables, traffic control and traffic control cost and other variables, contractors and any cost variables, as well as any other general relationship between variables undetected. In the following sections the different steps carried out will be discussed.

Data available on the paving project in various IDOT sources was explored. In addition, contractors (superintendents, estimators and management personnel) and various IDOT personnel were contacted and interviewed for further knowledge and support. A suitable amount of project instances was needed for the analysis. Therefore, all nine IDOT-districts were contacted and their input requested. A flat file - dataset consisting of 414 instances (individual projects) and 21 attributes was created. Each instance has a number of attributes. The attributes were divided into four main categories: General issues, Project characteristics, Traffic control issues and Contractor's issues. The collected attributes are listed in table 1 along with the type of each variable.

Table 1

Attributes and their state

Attribute	Type
General issues	
Contract no.	Numerical
District	Numerical
County	Logical
Location	Logical
Project characteristics	
Type of project	Logical
Distance	Numerical
No. of lanes	Numerical
Planned working days	Numerical
Actual working days	Numerical

DBE	Numerical
Volume of asphalt concrete	Numerical
Surface mix	Logical
Superpave	Numerical/Logical
Time of day	Logical
Traffic control issues	
Traffic control	Numerical/Logical
Total traffic control cost	Numerical
Contractor issues	
Contractors no.	Numerical
Name of contractor	Logical
Contractor's bid	Numerical
Percent change in bid	Numerical
No. of unsuccessful bidders	Numerical

General issues include attributes such as Contract number, District, County and Location. These have no other meaning other than distinguishing between the instances. The "Type of project" attribute describes the type of project being constructed, which can be: Surfacing, Resurfacing, Patching, Widening, or a combination of the two. The Planned Working days, Actual Working days, Length/Distance, Number of lanes, Volume of asphalt concrete, Mixture and Superpave attributes all represent typical aspects of operations. The Volume is represented as QC/QA tons (quality control and quality assurance), which is used as a rough approximation of the total asphalt concrete for each project. The Mixture attribute is used to identify the asphalt concrete mixture used for every project, which in this study includes mixtures, C, D, E and, F (or a combination in case of multiple overlays). The Superpave attribute (Superior Performing Asphalt Pavements) indicates whether the project was a Superpave project or not. This is a factor that can affect the performance and productivity of the contractor [8]. DBE (Disadvantage Business Enterprises) is the participation in % of total estimated contractor cost for minority or women businesses in every project. This attribute can be used to investigate whether DBE influences the contractor cost. "Traffic control" which is a Boolean attribute, and associated "Total cost of traffic control" are the only two attributes included in the Traffic Control category. The associated total cost usually consists of several pay-items, which had to be found and added up to make the total cost for traffic control. These attributes can help in identifying the impact of traffic control costs. This group consists of the Contractor (its name and bidding number), Contractor's bid (bid price), Number of unsuccessful bidders (for each bid) and Percent change from contractor's bid.

Descriptive statistical analysis was performed on the complete dataset to identify all essential information about the data. Scatter graphs were generated to identify sub-sets that may identify potential correlation between any two variables, check for outliers (table 2, figure 1) and the potential need for normalizing certain attributes. The analysis did not indicate any correlation between the variables nor the need for normalizing any particular attributes. Table 2 lists only a few of the attributes used for the data-mining process as well as their descriptive statistics. The statistical analysis revealed the fact that some variables in this study are not suitable for the data mining. For instance, after the cleaning process, only 2 nighttime projects out of overall 338 were left for data mining, making it impossible to create rules and draw any conclusions about nighttime paving operations. Therefore, the time of day attribute was left out when the dataset was mined.

Table 2

Descriptive statistics for Some of the attributes Selected

Attribute	Average	Stdev.	Range	
			Max	Min
Distance [miles]	3.33	3.40	24.00	0.07
No. of lanes	*	*	*	*

Planned working days	48.0	39.8	310	15
Actual working days	37.8	33.6	191.5	2
DBE	0.06	0.04	0.16	0
Volume of asph. concrete [tons]	9936	17545	152151	0
Total traffic control cost [$]	$ 19,647	$ 38,126	$ 465,250	$ -
Bid price [$]	$ 878,913	$ 1,387,366	$ 15,003,639	$ 33,576
Percent change from bid	4.1	10.9	74.0	-28.2
No. of unsuccessful bidders	1.7	1.4	7	0
Actual wd/Planned wd	0.61	0.42	2.69	0

During the analysis, it became obvious that additional attributes were needed. Location as recorded, for example, had no meaning to the problem. Some projects took place in the same area (or on the same highway) but were distributed statewide in different locations, which could result in unreliable rules being extracted from the dataset. To include the location of each instance in some form, an attribute was added to the dataset. Each project or instance was coded as either 0 or 1, roadway or highway respectively.

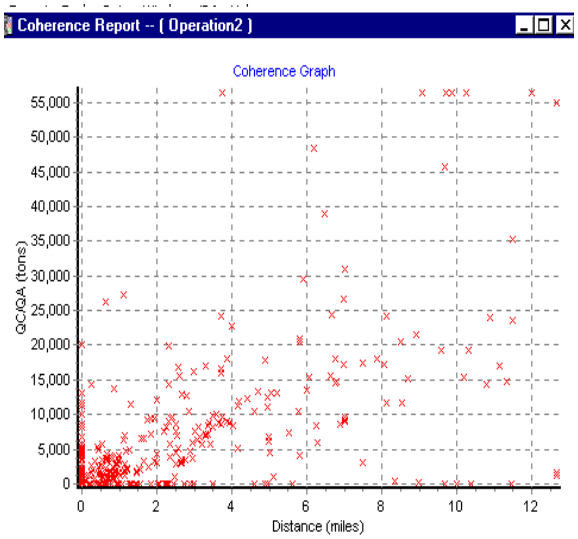

Figure 1. A coherence graph between tonnage and Distance

A combination of classification-neural networks and statistical analysis was used to generate the rules described in the next section. In particular, since there are no domain-specific solutions for the construction industry that exist to date, general-purpose data mining software were utilized. Details about these tools can be found in [11].

Results

The generated rules are presented in the following format:

IF Location = 0 AND Type of project = Surfacing (and) Bid <= 508391 THEN Total Traffic Cost <= 9125 (92.9% confidence/13 cases)

That is for Roadways where new surface is being laid and contractor's bid price was less than or equal to $508,391 the total traffic cost was less than or equal to $9125. This can be stated with approximately 93% confidence and is supported with 13 cases. The goal is not just to find rules similar to the one above, but instead to extract a number of similar rules that would resemble a trend. Certain patterns were identified from the hundreds of rules extracted. For example,

- There is a general trend for new surface type of projects for highways to be more often complete within scheduled time. Any combination of rules in this regard (and there are several) has usually a high confidence level and is largely supported.
- In district 3, projects tend to be within scheduled time when planned working days are less than 35. This is in particular the trend when the project is of the resurfacing type (overlay). Moreover, the bid price tends to be less than $283,000 if working days are less than 35 working days.
- If projects in district 2 are scheduled longer than 35 days, it can be implied with more than 50% confidence that projects overrun the time-schedule. Also if the volume of asphalt concrete is more than 5130 (QC/QA) tons the bid tends to be more than $760,000.
- For resurfacing types of projects in general, when projects are scheduled for less than 35 working days, tend to be either low in volume (tons of asphalt concrete) or the bid price is lower than $ 283,000.
- For district 7, if volume of asphalt concrete is less than 5130 tons and either traffic control cost lower than $5250 then bids tend to be lower than $283,000.
- Moreover, the bid price tends to be less than $283,000 if working days are less than 35 working days. Also if the volume of asphalt concrete is more than 5130 (QC/QA) tons the bid tends to be more than $760,000.

After these patterns were generated, IDOT was contacted to verify some of these patterns. IDOT personnel could confirm some of the patterns and did provide explanation for them. For example, one of the reasons suggested for the extracted rules about exceeding schedules in some rural IDOT districts relates to generous mobilization times needed in those districts, which has to be a minimum of 15 days.

Conclusions and Recommendations

The study presented here describes an application of data-mining analysis to a typical construction database containing information about asphalt projects in Illinois. A case study was presented to test the applicability of data mining as an analysis method. A database was constructed with collected data from IDOT sources. Data-mining technique was utilized to analyze the created dataset and rules generated. Based on the generated results and interpretation, certain previously unknown patterns were discovered. The study shows that data mining can provide information on a dataset/database beyond statistical methods only and provide a source of valuable information (that could not have been detected otherwise) to support decision-making. If the time-consuming data collection process can be reduced, the method can extract information faster than other analysis methods. Suggestions for future research include increasing the size of the dataset used, as well as trying other software and techniques to verify the extracted rules and trends. One of the main characteristics of data mining is the large amount of data needed to generate rules. The major rules generated in this study usually have high confidence but only limited amount of cases supporting the rule. Therefore, increasing the data set will significantly enhance the quality and reliability of the generated rules and trends.

Acknowledgment

The author would like to acknowledge the support provided by IDOTpersonnel in providing the data and the follow up analysis.

References

[1] Adrians, Pieter, Zantinge, Dolf. "Data mining." Addison-Wesley Longman, England, 1996.
[2] Cabena, Peter, et al. Discovering data mining: From concept to implementation. Prentice Hall, NJ, 1997.
[3] Han, Jiawei. Data mining: Concepts and techniques. Morgan Kaufmann Publishers, San Francisco, 2001.
[4] Hand, D.J, Mannila, H., Smyth, P. Principles of data mining. MIT press, Massachusetts, 2001.
[5] Witten, I. H., Frank, E. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufman, California, 2000.
[6] Nii O. Attoh-Okine, Rough set application to data-mining principles in pavement management database, J. Comput. Civ. Eng., Am. Soc. Civ. Eng. 11 (4) (1997) 231-237.

[7] Sou-Sen Leu, Chee-Nan, Shiu-Lin Chang. Data mining for tunnel support: neural network approach. www.elsevier.com, automation in construction, 2000.
[8] Soibelman, Lucio, Kim, Hyunjoo. Data preparation process for construction knowledge generation through knowledge discovery in databases. Journal of Computing in Civil Engineering, ASCE, 16 (1), (2002) 39-47.
[9] Soibelman, Lucio. Construction knowledge generation and dissemination. Berkeley-Stanford CE&M workshop: Defining a research agenda for AEC process/product development in 2000 and beyond.
[10] Two Crows Corporation. Introduction to data mining and knowledge discovery. Third edition. Two Crows Corporation, 1999. http://www.kdnuggets.com/software/index.html, 2002