University of Alberta

Leveraging Operational Data for Intelligent Decision Support in Construction Equipment Management

by

Hongqin Fan ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Construction Engineering and Management

Department of Civil and Environmental Engineering

Edmonton, Alberta
Fall, 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-32955-9
Our file Notre référence
ISBN: 978-0-494-32955-9

Dedicated to Jingwen Xu, Jadon Fan and Lily Fan for your love and support

H. F.

ABSTRACT

Construction equipment management is aimed at managing equipment resources in order to maximize return on capital investments and satisfy the needs of project management in a timely and cost-effective manner. A rapid development of computer software and hardware, along with various automation methods for data acquisition, has catalyzed the computerization of construction equipment management in recent years. The primary objective of this research is to investigate the application of cutting edge data warehousing and data mining technologies for intelligent decision-making in construction equipment management.

In cooperation with Standard General Inc., a large road building and maintenance contractor in Alberta, Canada, and based on the M-Track equipment information management system developed by NSERC/Alberta Construction Industry Research Chair as well as nine years of equipment operational data, this research proposes to improve the M-Track system in its data analysis and decision support capabilities using advanced computer tools. As a response to the problems existent in Standard General Inc., and common to the construction industry in general, the research addresses issues of (i) How to design and implement an information infrastructure which advocates data sharing, information retrieval, and knowledge discovery for fact-based intelligent construction equipment management; (ii) how to apply the data warehousing technique in construction equipment management for decision support; (iii) how to leverage the large amounts of

operational data for automated knowledge generation and decision support using data mining techniques.

A number of tests and demonstrations on the prototype applications have proven that data warehousing and data mining are suitable technologies for improving the current practice of decision support through an integrated data repository, need-based information retrieval and decision analysis, and automatic identification of trends, patterns, or rules from data. The novel, non-parametric outlier mining algorithm developed in this research has proved to be effective and efficient in detecting the top-N most interesting outliers from equipment database. Although developed in the field of construction equipment management, the documented findings, proposed methods, and recommended best practice are equally applicable to other areas of construction management.

PREFACE

This thesis has been prepared in paper-based format, consisting of seven chapters. Each chapter is an independent paper and can be read separately. However, all the chapters are logically coherent and pertinent to the broad theme of the thesis. The thesis begins with an introductory chapter elucidating the general research motivations, research methodologies, research objectives, and expected contributions with regard to both academia and the construction industry. Chapter 2 overviews the current state of practice in construction equipment management, including the opportunities associated with the wide application of information and communication technology, as well as the challenges in data, information, and knowledge management and decision analysis. Chapter 3 proposes a general framework for intelligent construction equipment management, with an emphasis on incorporating data mining techniques into the current equipment information management system for automated knowledge generation, dissemination, and utilization. Chapter 4 investigates the application of data warehousing techniques in construction equipment management, the general methodology for application, as well as the corresponding benefits and challenges. Chapter 5 presents the application of data mining for assessing equipment residual values in the market using a predictive data mining algorithm. Chapter 6 proposes a novel nonparametric outlier mining algorithm in support of anomaly detection from the construction equipment database, as well as other decision support tasks involved in equipment management. Finally, chapter 7 concludes the thesis and makes recommendations for future research in mining equipment data for decision support.

Chapters 3, 4, 5, 6 are written based on published papers in referred journals and conference proceedings during the course of this research. The writer of this thesis is the primary and major contributor of each published paper among the coauthors, the other coauthors provided supervisory works, offered critiques, and made suggestions on the manuscripts.

ACKNOWLEDGEMENTS

This thesis, like most others, was written with a great deal of time and effort. The dissertation would not have been possible without the generous help and support of the following people:

First of all, I would like to acknowledge the contributions of my supervisors, Dr. Simaan AbouRizk, and Dr. Hyoungkwan Kim, who introduced me to the exciting field of data mining and knowledge discovery in construction equipment management. Their vision, supervision, guidance, financial and spiritual supports helped me through this important period in my life time, aided me in realizing my academic goals and preparing me for future undertakings;

Secondly, I feel indebted to Dr. Osmar Zaïane in the Department of Computing Science. His excellent course on data mining armed me with the knowledge and tools needed for my research; his comments, suggestions, and guidance on my research work led to a number of quality publications. I am also very impressed by the enthusiasm and energy he instills in his students, including myself.

Thirdly, I must express my sincere appreciation to Dale Tillapaugh, the equipment manager of Standard General Inc. for sharing his equipment management knowledge and experience, guiding my research and providing help when I was in Standard General and throughout my research.

Fourthly, I wish to thank Steve Hague, Ryan Gurnett, and Jeff King in NSERC/Alberta Construction industry Research Chair for their help in computer programming and technical writing.

Finally, I wish to extend special thanks to my wife, Jingwen Xu, and my two children, Jadon and Lily, for their continuous support, understanding, and love.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1

OVERVIEW 1

RESEARCH MOTIVATION 5

Decision Support in Construction Equipment Management: A Contractor's Perspective 5

Emerging Data Warehousing Technology 10

Emerging Data Mining Technology 14

RESEARCH OBJECTIVES 19

RESEARCH METHODOLOGIES 20

Literature Review for Problem Identification 20

Site Visit for Fact Finding 21

Interview and Meetings with Domain Experts for Problem Formulation 21

Computer Modeling and Programming for Design, Experimentation, and Validation 21

ACADEMIC CONTRIBUTIONS 27

INDUSTRIAL CONTRIBUTIONS 29

CONCLUSIONS 30

REFERENCES 31

CHAPTER 2. CONSTRUCTION EQUIPMENT MANAGEMENT: STATE-OF-THE-PRACTICE 33

INTRODUCTION 33

CONSTRUCTION EQUIPMENT DATA MANAGEMENT 34

Data Acquisition 36

Data Delivery and Distribution 40

Data Storage 43

Data Exploitation 43

COMPUTERIZED CONSTRUCTION EQUIPMENT MANAGEMENT 46
DECISION SUPPORT IN CONSTRUCTION EQUIPMENT MANAGEMENT 51
Pushed Information and Knowledge 53
Pulled Information and Knowledge 53
CONCLUSIONS 54
REFERENCES 56

CHAPTER 3. BUILDING INTELLIGENT APPLICATIONS FOR CONSTRUCTION EQUIPMENT MANAGEMENT 59
INTRODUCTION 59
LITERATURE REVIEW 62
CONSTRUCTION EQUIPMENT MANAGEMENT VERSUS DATA MINING 63
SYSTEM ARCHITECTURE FOR INTELLIGENT CONSTRUCTION EQUIPMENT MANAGEMENT 66
INTELLIGENT FEATURES OF THE PROPOSED FRAMEWORK 69
CASE STUDY: WORK ORDER EVALUATION 73
CONCLUSIONS 78
REFERENCES 80

CHAPTER 4. DATA WAREHOUSING FOR CONSTRUCTION EQUIPMENT MANAGEMENT 83
INTRODUCTION 83
LITERATURE REVIEW 88
ARCHITECTURAL DESIGN OF EQUIPMENT DATA WAREHOUSE 89
MULTIDIMENSIONAL MODELING 93
EQUIPMENT DATA WAREHOUSING 98
Data Sources 99
Data Staging 100
Data Presentation 101
EQUIPMENT DECISION SUPPORT SYSTEM DESIGN 102

BENEFITS OF EQUIPMENT DATA WAREHOUSE 108
SUMMARY AND CONCLUSIONS 111
REFERENCES: 113

CHAPTER 5. ASSESSING RESIDUAL VALUE OF HEAVY CONSTRUCTION EQUIPMENT USING PREDICTIVE DATA MINING MODEL 116
INTRODUCTION 116
LITERATURE REVIEW 120
DATA MINING FOR PREDICTION OF EQUIPMENT RESIDUAL VALUE 123
AUTOREGRESSIVE TREE (ART) ALGORITHM 126
ART MODEL FOR PREDICTION OF EQUIPMENT RESIDUAL VALUE 130
Data Sources 131
Feature Selection 131
Data Quality Control 134
Model Generation and Validation 135
COMPARISON OF ART WITH ANN AND MLR MODELS 141
DEPLOYMENT OF PREDICTIVE DATA MINING MODELS FOR EQUIPMENT RESIDUAL VALUE 143
DISCUSSIONS 145
SUMMARY AND CONCLUSIONS 147
REFERENCES 150

CHAPTER 6. DECISION SUPPORT IN CONSTRUCTION EQUIPMENT MANAGEMENT USING A NONPARAMETRIC OUTLIER MINING ALGORITHM 154
INTRODUCTION 154
MULTIVARIATE OUTLIER DETECTION 159
RESOLUTION-BASED OUTLIER 162
RB-OUTLIER MINING ALGORITHM 166

COMPARISON OF RB-OUTLIER, DB-OUTLIER AND LOF-OUTLIER MINING ALGORITHMS 169
TEST RESULTS ON A THREE-DIMENSIONAL EQUIPMENT DATASET........ 173
APPLICATION ON CONSTRUCTION EQUIPMENT MANAGEMENT............. 178
CONCLUSIONS AND FUTURE RESEARCH 180
REFERENCES 181

CHAPTER 7. CONCLUSIONS AND FUTURE RESEARCH 183
SUMMARY OF THE THESIS 183
A Framework for Intelligent Construction Equipment Management 184
Data Warehousing for Construction Equipment Management 185
Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model 186
A Nonparametric Outlier Mining Algorithm for Detecting Anomalies from Equipment Data 188
CONCLUSIONS 189
RECOMMENDATIONS FOR FUTURE RESEARCH........ 190
Time Series Analysis for Financial Planning and Inventory Control 190
Characterization Analysis for Decision Support in Equipment Management 193
Mining Simulation Data for Optimization of Equipment Resources 194
FINAL REMARKS 197
REFERENCES 198

LIST OF TABLES

Table 1- 1. Data Mining Tasks versus Exemplar Decision Support Tasks in Equipment Management 16

Table 3- 1. Differences between Prior Knowledge and Derived Knowledge 71

Table 4- 1. Results of Sample MDX Query on Maintenance and Repair Cost Data Cube 108

Table 5- 1. Test Results of 10-fold Cross-Validation 141

Table 5- 2. Comparison of ART with ANN and MLE in prediction accuracy 143

Table 6- 1. Comparison of DB-outlier, LOF-outlier and RB-outlier 171

Table 6- 2. Top-20 Outliers Identified by DB-outlier, LOF-outlier and RB-outlier Algorithms 176

LIST OF FIGURES

Figure 1- 1. Work Order Processing Module in M-Track 7

Figure 1- 2. Multidimensional Data Model for Equipment Revenue 13

Figure 1- 3. Comparison between Data Mining Model and Statistical, Mathematical or Simulation Model 17

Figure 1- 4. Equipment Data Warehousing for Decision Support 24

Figure 1- 5. Procedures for Building Data Mining Modules 25

Figure 2- 1. KOMTRAX® Data Communication System for Equipment Management ...42

Figure 2- 2. Typical Functional Design of a CCEM System 47

Figure 3- 1. System Architecture for Intelligent Equipment Management 67

Figure 3- 2. A Portion of the Induced Work Order Evaluation ART Model 76

Figure 3- 3. Scatter Plot of Actual Deviations vs. Predicted Deviations Using the Test Dataset 77

Figure 3- 4. Recommendation on Labour Hour Estimate from Embedded ART Model ..78

Figure 4- 1. Comparison between Process-oriented View of Equipment IMS and Subject-oriented View of Equipment DSS 87

Figure 4- 2. Bus Matrix for Equipment Data Warehouse 92

Figure 4- 3. Dimensional Model for Maintenance and Repair Cost Data Cube 94

Figure 4- 4. Examples of OLAP Operations on the Maintenance and Repair Cost Data Cube (drill down and slice, slice, roll up) 97

Figure 4- 5. Equipment Data Warehousing and Data Access 99

Figure 4-6. Equipment Data Warehouse System Architecture 103

Figure 4- 7. Visual Browsing of Data Cubes in Equipment Decision Support System...105

Figure 4- 8. Customized MDX Queries in the Equipment Decision Support System.....105

Figure 5- 1. Histogram of Wheel Loader Auction Price...135

Figure 5- 2. Partial ART Data Mining Model for Prediction of Equipment Residual Value...137

Figure 5- 3. Boxplots for Prediction Error Rates of Cross-Validation Tests...................140

Figure 5- 4. Screenshot of the Equipment Information Management System.................144

Figure 6- 1. Identification of Resolution-Based Outliers...166

Figure 6- 2. Comparison of Results from DB-outlier, LOF-outlier, and RB-outlier on a 200-tuple Synthetic Dataset..172

Figure 6- 3. Visualization of the Three-dimensional Equipment Dataset.......................174

Figure 7- 1. Forecasting Monthly Fleet Maintenance Costs and Fleet Repair Costs Using Time Series Data Mining Model...194

CHAPTER 1. INTRODUCTION

OVERVIEW

Construction equipment management refers to the management of equipment resources to maximize the return of capital investments and satisfy the needs of project management in a timely and cost-effective manner (Vorster and Livermore 1994). Heavy construction, including most transportation projects, and industrial installations, highly relies on construction equipment. With the increasing scale, complexity, and safety requirements in modern construction, construction equipment management plays an increasingly critical role in the entire process of construction management, contractors need to continuously upgrade and expand their fleet of construction equipment in order to increase their competitive edge in the construction market. According to a recent survey and report on the U.S. construction equipment market (Reed 2007):

- Following a 14.6% increase in 2006, the total spending on heavy construction in 2007 is expected to reach $224.6 billion US, an increase of 11.6% over 2006;
- The heavy construction market is very competitive, with eight out of 10 contractors specializing in this type of work describing competition as “intensely” or “very” competitive;

- 33% of the contractors reported net fleet expansion, measured by the number of machines added to the fleet in the year of 2006. This percentage is expected to be 27% in 2007;

- The replacement rate of contractors' fleets, measured by the number of machines taken out of the fleet and replaced, was 9.4% in 2005, 10.4% in 2006 and is forecast to be 10.3% in 2007;

- Purchase remains the primary strategy for fleet managers who are acquiring major machines. The machines with a higher sticker price are more likely to be purchased outright than through financing among contractors. Short-term rental remains a viable acquisition strategy among contractors; and steady annual increases in both short-term and long-term rentals have been reported.

This report partially reflects the following trends in equipment management in heavy construction in recent years:

Increased construction volume, but increased competition – The economic boom in recent years has led to continuous increase in heavy construction investment every year; however, the contractors are faced with fiercer competitions in keeping and increasing their market share;

Need for expansion and upgrading of current fleets, but faced with stringent cost control – To satisfy the needs for modern construction and increase their competitive edge, the contractors must dispose the deteriorated, obsolete machines and constantly purchase new powerful productive machines. The contractors must keep the equipment owning and operating costs to a minimum while providing the needed equipment for project construction;

More options in equipment acquisition – The construction equipment market is diversified, which means construction equipment may be acquired through various means, including purchase outright, purchase through financing, rent and buy, and short-term rent. The contractors need to make long-term and short-term decisions on equipment acquisition in order to maximize the return of investment in fixed assets.

Innovations in construction equipment have brought about many changes and opportunities in construction equipment management. For examples, computer control of equipment systems is used to regulate and control fuel delivery and efficiency, exhaust emissions, hydraulic systems, power transfer, load sensing and operation tracking, recording, and regulating (Gransberg et al. 2006). Later models of construction machines are designed with increased power, safety and fuel economy, and automated monitoring and diagnostic systems; some construction equipment operations have been revolutionized through automated control systems and sensing technology although complete automation of construction equipment will not be possible for the near future due to economic and technical considerations (Russell and Kim 2003). The computer information management system has been widely used in large equipment-owning organizations for daily operational control and tracking. Research and development of all these innovative technologies in the areas of construction equipment and equipment management are aimed at reducing equipment operational costs, and improve efficiency, quality and safety. The contractors that fail to take the opportunity to arm themselves with powerful tools will risk losing their competitive edge and market share.

While no fundamental changes have been reported up to this time and neither will they be expected in the near future in heavy construction equipment in terms of performing basic mechanical operations such as digging, pushing, loading, and lifting, the application of advanced computer technologies (hardware and software) is bringing significant changes to equipment operations and management (Gransberg and Ryan 2006). These changes, from the standpoint of equipment management, have enhanced decision support capabilities of contractors in the following aspects: (i) the entire suite of data needed for equipment tracking, monitoring and operational support is collected and distributed through various automation approaches for various purposes; (ii) the equipment information systems are capable of storing these data and generating various reports and displays to support decision making; and (iii) equipment management is greatly facilitated by the increased availability of just-in-time data and information through wireless transmissions, mobile electronic devices, and the internet. Although these changes have benefited the daily operations of construction equipment, decision makers at the corporate level found that these data do not support the equipment management decisions to a great extent. Large amounts of data cannot be turned into actionable information and knowledge, and human interpretation of these data is difficult without powerful computer tools available for analyzing these data.

Using the case of Standard General Inc., a large road building contractor in Alberta, Canada, this research addresses an industry-wide problem in construction equipment management: decision makers cannot turn the large amounts of equipment data into meaningful information and knowledge for fact-based decision support due to lack of support from the current information architecture. To address this problem, the research takes advantages of the recent developments in data warehousing and data mining techniques in computing science to support interactive information generation and automated knowledge discovery for construction equipment management.

RESEARCH MOTIVATION

This research is motivated by the common decision support problems faced by heavy construction contractors, enabling technologies of data warehousing and data mining in computing science.

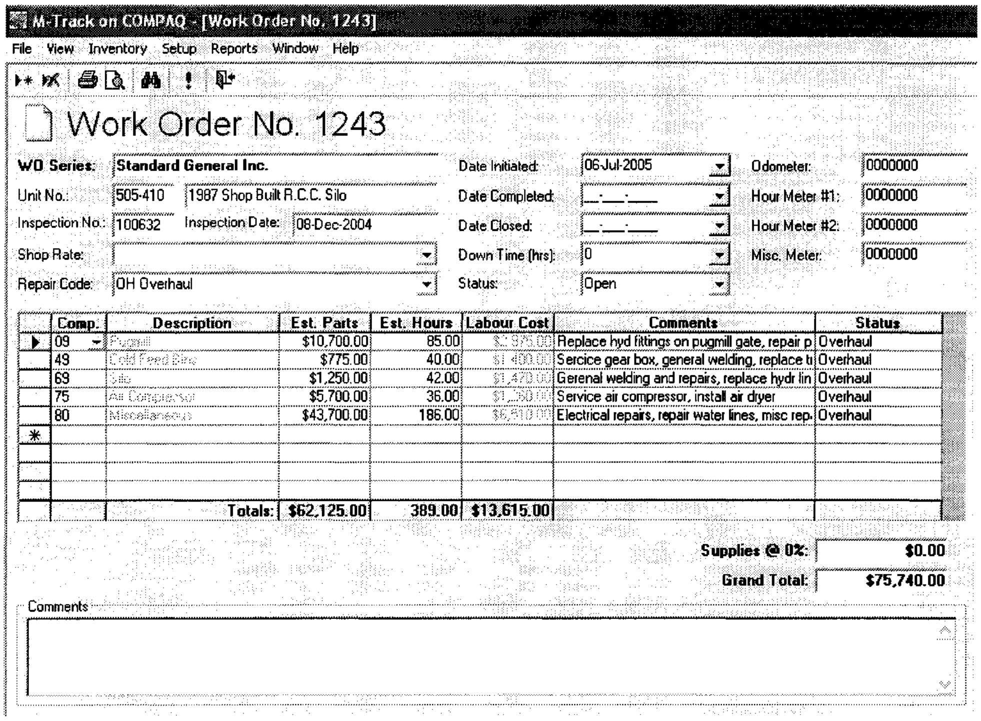
Decision Support in Construction Equipment Management: A Contractor's Perspective

The collaborating contractor, Standard General Inc. (hereinafter called "the contractor"), boasts a large construction fleet including over 1300 pieces of heavy construction equipment in earthmoving, asphalt construction, concrete production and supply to support its road building and maintenance business in Alberta, Canada. The contractor has used an equipment information management system called M-Track since

1997 to facilitate the management of fleet repair, maintenance, and operations, and has achieved great success.

M-Track was developed in 1997 and later upgraded in 2001 by the NSERC/Alberta Construction Industry Research Chair (hereinafter called "the chair"). As a customized development, M-Track modeling and design is based on the equipment management process of the contractor. All business operations related to equipment management are seamlessly integrated into a single multi-faceted system. After taking input data from various management activities, the system stores data in a M-Track equipment database and is capable of generating over 48 types of reports. These reports are either standardized reports for accounting and management support, or customized parameter-driven reports for decision analysis. Figure 1-1 shows the "work order processing" module in M-Track, which handles work order management. When the superintendent places an estimated work order through M-Track, the work order can be retrieved, viewed, or printed in the shop for processing. A completed work order is recorded by mechanics in M-Track indicating actually incurred repair items and costs. Through the reporting module of M-Track, the work orders can be summarized and printed out, along with comparisons between the estimated and completed work orders.

Preventive Maintenance (PM) is a "smart" module designed for automatic scheduling of equipment maintenance. PM events for each piece of equipment are scheduled based on multiple criteria. In addition to the manufacturer-recommended maintenance intervals, PM events are also scheduled based on service meter readings, elapsed time, and fuel consumption. These criteria are programmed into the system for weekly fleet maintenance scheduling and control.


M-Track on COMPAQ - [Work Order No. 1243]
File View Inventory Setup Reports Window Help

Work Order No. 1243

WO Series:	Standard General Inc.	Date Initiated:	06-Jul-2005	Odometer:	0000000
Unit No.:	505-410 1987 Shop Built R.C.C. Silo	Date Completed:	__-___-____	Hour Meter #1:	0000000
Inspection No.:	100632 Inspection Date: 08-Dec-2004	Date Closed:	__-___-____	Hour Meter #2:	0000000
Shop Rate:		Down Time (hrs):	0	Misc. Meter:	0000000
Repair Code:	OH Overhaul	Status:	Open		

Comp.	Description	Est. Parts	Est. Hours	Labour Cost	Comments	Status
09	Pugmill	$10,700.00	85.00	$2,975.00	Replace hyd fittings on pugmill gate, repair p	Overhaul
49	Cold Feed Bins	$775.00	40.00	$1,400.00	Sercice gear box, general welding, replace ti	Overhaul
69	Silo	$1,250.00	42.00	$1,470.00	Gerenal welding and repairs, replace hydr lin	Overhaul
75	Air Compressor	$5,700.00	36.00	$1,260.00	Service air compressor, install air dryer	Overhaul
80	Miscellaneous	$43,700.00	186.00	$6,510.00	Electrical repairs, repair water lines, misc rep.	Overhaul
*						
	Totals:	$62,125.00	389.00	$13,615.00		

Supplies @ 0%: $0.00

Grand Total: $75,740.00

Comments

Figure 1- 1. Work Order Processing Module in M-Track

Through nine years of equipment management using M-Track, the contractor has successfully operated its equipment fleet for construction needs and tracked the history of its equipment fleet. While equipment operations using M-Track have been a great success, the contractor found that using M-Track for decision analysis in equipment management become more difficult with large amounts of data, with increased complexity in data analysis. To satisfy the emerging needs of the contractor in equipment decision analysis, the chair had to keep adding new modules and equipment report types as per the requests from the contractor. The reported problems or areas for improvement on M-Track are mostly focused on the output side of the system, as summarized below:

Firstly, the equipment data are located in three different sources. Equipment repair and maintenance data before 2001 are stored in the previous version of M-Track backed by a Microsoft Access database, while those acquired after 2001 are located in the current version of M-Track backed by a Microsoft SQL Server database with a different database model and design. Besides, equipment usage data have been tracked in an accounting system developed by the Explorer Software Inc. The disparate equipment data made it difficult to perform trend analysis across the year 2001 as it would involve data extraction from both versions of M-Track databases, or to perform equipment analysis involving both equipment usage and operating cost needs data from M-Track data and accounting data, respectively. These two types of analysis cannot be performed by the contractor without resorting to the database specialists in the chair for help.

Secondly, the many report types still cannot satisfy the needs for equipment data analysis. Data analysis for decision support in equipment management is usually an explorative process involving intense interaction between the decision makers and M-Track. The decision makers need to look at equipment data from different viewpoints depending on the decision support problems encountered. Standardized reports or the report with limited user control cannot be used to answer varying business questions. A large number of decision support questions, which include the identification of equipment cost overrun, the identification of equipment underutilization, the relationship between productivity and operating costs, and comparisons of equipment cost with competitor and industrial benchmarks, all involve multiple unanticipated views of data. As a result, the report-style ad-hoc views of equipment data do not provide the flexibility needed for decision analysis.

Thirdly, the interpretation of data is difficult due to a lack of suitable computer tools. Although the equipment data depict daily equipment operations and contain hidden trends, rules, and patterns, they cannot be directly used for explanatory or predictive analysis tasks, such as the following:

- What are the major factors impacting the operating costs of equipment in the company? How can we reduce the costs?
- Based on the financial data available on equipment maintenance and repair over the past few years, what will be the likely costs over the next few months?
- Classify the field usage reports of construction equipment as credible, or non-credible.
- What is the likely market price of a piece of equipment if it is sold?

All these questions can be answered based on the trends, rules, and patterns uncovered through in-depth data analysis; however, this process is difficult or impossible to accomplish by human interpretation alone or simple statistical analysis without relying on powerful computer tools.

A lack of decision support from the equipment management system is not a unique problem to this contractor, but is typical in the construction industry. There are escalating needs for fast information retrieval, explorative interactive analysis, and automated knowledge generation in construction equipment management.

Emerging Data Warehousing Technology

A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data supporting management decisions (Inmon 2005). It has been applied in a wide variety of industries including retail, banking, and telecommunications for strategic decision support with great success. In contrast to the relational database used in most of the current information systems for operational support, data warehouse is built separately for decision support using data from various operational or historical data sources. Data warehousing technology provides the following decision support capabilities which cannot be attained using relational databases:

- Data are validated, transformed, and scrubbed before loading into the data warehouse from other sources, which ensures the data are of high quality;
- Data are pulled from various operational data sources on a regular basis, which leads to integrated data, i.e. all the data needed for analysis are centralized into a single source;

- Data in the data warehouse are organized by subjects using multidimensional data models, which allows the decision makers to choose a subject, and to perform data analysis from different perspectives at different levels of details; and

- Online Analytical Processing (OLAP) of data warehouse facilitates user-directed information retrieval and interactive analysis through visual operations of drill-down, roll-up, slicing and dicing, and pivoting, or through multidimensional query language.

The major differences between the multidimensional data warehouse and the relational database are summarized by Bain et al. (2001) as follows:

- *Design* – data warehouse is subject-oriented and models certain subjects of the business; database is application-oriented, which models a particular business process or functionality;

- *Level of Detail* – data in the data warehouse are summarized or refined using complex calculations; data in the database are detailed data of business operations;

- *Data structure* – data warehouse structure is dynamic insofar as more contents can be added as needed; database structure is usually static with less changes;

- *Target* – data warehouse targets decision making people; database targets data-entry people;

- *Changes* – data in the data warehouse are nonvolatile after loading; data in the database are subjected to frequent changes at any time; and
- *Timeline* – data warehouse reflects historical situation over time; database reflects current situation.

The fundamental difference between a data warehouse and a database is that they use different data models for different purposes. A relational database mimics the business processes using relational data models for fast data updating; A data warehouse depicts the analytical subjects using multidimensional data models for fast information retrieval. Data warehouse is designed to overcome the inefficiency of using a relational database for decision analysis. The benefits of using a data warehouse for decision analysis can be illustrated by the multidimensional data model for "equipment revenue" depicted in Figure 1-2. The multidimensional data model contains a central fact table which stores all the details of facts on the revenue generated by equipment usage, including hours, rate, and earned revenue. The other five tables represent the dimensions which describe the revenue facts, with each dimension containing multiple conceptual levels in a hierarchal structure. This multidimensional data structure allows decision makers to conduct analysis on revenue data from any perspective, and at any level of detail, in order to answer decision support questions regarding which, where, who and when of the revenues earned through the use of construction equipment.

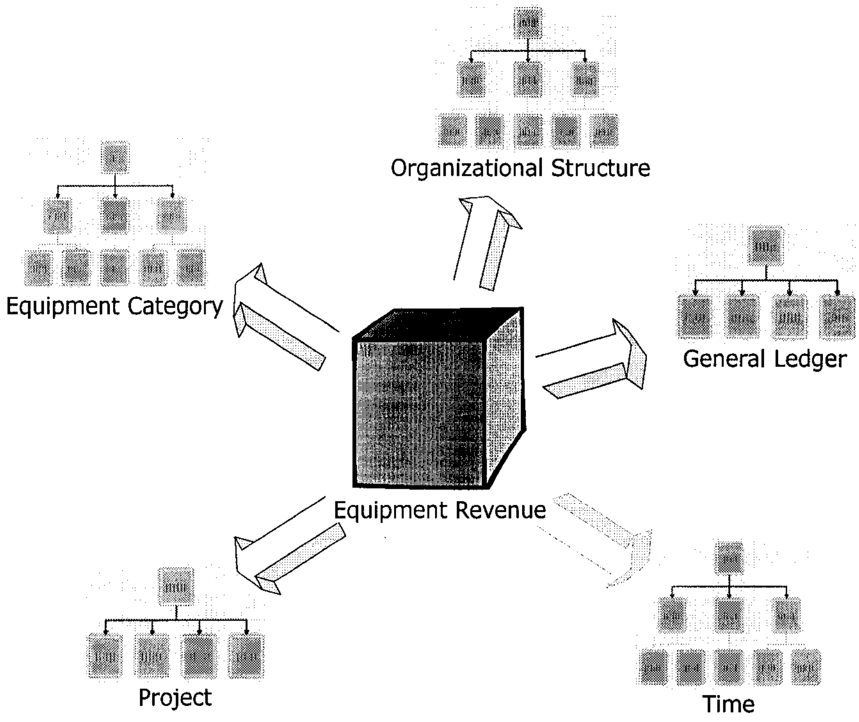


Figure 1- 2. Multidimensional Data Model for Equipment Revenue

Large amounts of financial resources and efforts were involved in the implementation of data warehousing in the past, which made it only affordable to large multinational corporations. Some obstacles to its implementation include problems such as application incompatibility, difficult data extraction, transformation and loading, and the high costs of hardware. Recent developments in data warehousing technology and supporting hardware have greatly simplified the data warehousing design and reduced costs in its implementation. The major database management system developers, such as Oracle, IBM, and Microsoft, have provided data warehouse and OLAP tools in addition to their traditional relational database systems. The design and implementation of a data warehouse becomes technically and economically feasible within media to large size heavy construction contractors.

Emerging Data Mining Technology

Data mining refers to the discovery of previously unknown, non-trivial, and potentially useful (Fayyad et al. 1996) knowledge from large amounts of data using complex data mining algorithms. Data mining research and development was motivated by the needs in the commercial industry for fraud detection, customer relation management, and market segmentation. Currently, the data mining technology is widely applied in science and engineering for such applications as new drug testing and development, science exploration, product design and production control, and environmental protection. Data mining has emerged as a solution for the automatic or semi-automatic discovery of knowledge from the large collections of data available in most business operations to replace human interpretation of data.

Data mining draws theories and methods from multiple disciplines including database system, statistics, artificial intelligence, information theory, and visualization. Data mining technology thereby enables knowledge discovery from data with many decision-friendly features:

- A wide variety of data mining algorithms are available for different decision support tasks, including association analysis, prediction, classification, time series forecasting, clustering, outlier detection, and text analysis;

- Many data mining models can be visualized for human interpretation and judgment;

- Data mining models can be designed as an automatic process and integrated into a decision support system for knowledge discovery, validation, updating, and exploitation for predictive or explanatory analysis; and

- Many data mining models are able to explain the inference process while being used for predictive analysis, which enables the users to make informed decisions.

Data mining algorithms play the role of generating primitive trends, rules, or patterns from data; therefore, tremendous efforts have been devoted to the development of novel data mining algorithms in data mining research in recent years to improve mining quality and efficiency. Depending on their purposes, data mining algorithms can be either descriptive data mining algorithms or predicative data mining algorithms. The former helps to better understand the data by uncovering the relationships and patterns in the data, whereas the latter is used to generate data-driven models for predication, classification, and forecasting.

Table 1-1 shows some general data mining tasks and the kinds of knowledge that can be uncovered from equipment data for different decision support tasks.

Table 1- 1. Data Mining Tasks versus Exemplar Decision Support Tasks in Equipment Management

Data Mining		**Equipment Management**	
Tasks	**Objectives**	**Tasks**	**Objectives**
Association rules mining	Identify events which most likely occur concurrently or sequentially	The association between major component breakdowns of construction equipment	Proactive repair/ replacement of equipment
Clustering	Group the data into clusters so that intra-cluster data are similar and inter-cluster data are dissimilar	Identification of equipment groups with common interesting features	Decision support in equipment allocation and operation
Outlier mining	Find data objects which deviate significantly from general patterns in data	Identify pieces of equipment with abnormal cost of maintenance/repair	Equipment rebuild/replacement/ disposal
Classification	Predict a categorical attribute based on the other attributes of known values	Classify the equipment field usage report as "Credible" or "Not credible"	Identify misuse, or misreport of equipment
Prediction	Predict a numerical attribute based on the other attributes of known values	Predict the equipment market value based on known parameters	Life cycle costing analysis
Forecasting	Forecast time series data into the future	Forecast the future equipment maintenance and repair cost for a piece of equipment, or a group of equipment	Budgetary analysis or life cycle costing analysis

Mining knowledge from data for decision support also differs from the traditional statistical methods, decision tools, and expert systems. The data mining models that represent discovered knowledge have two unique features. Firstly, the models are data-driven; that is, data mining models are obtained from data using complex computer algorithms, meaning that the models are based more on inferred facts than personal judgment or expert knowledge. Secondly, the data mining models may become the only viable solution when the system is too complex to be described by other models. Figure 1-3 describes the major differences between the data mining model and traditional statistical, mathematical, or simulation models.

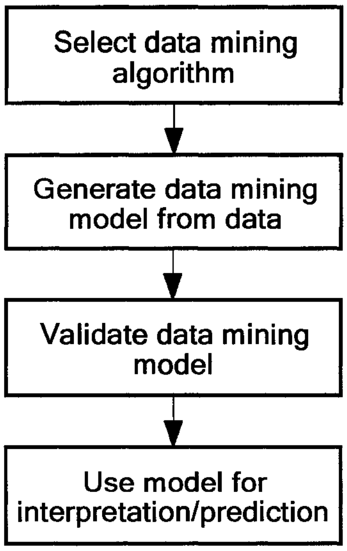
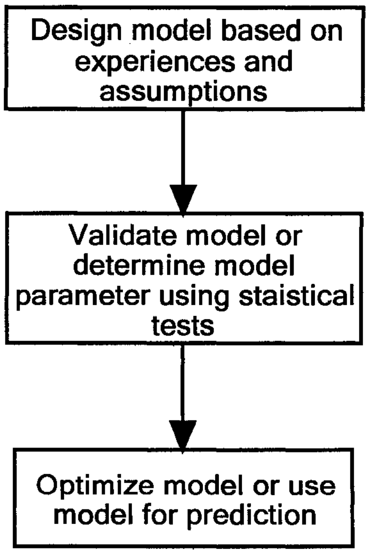
	Data Mining Model	Statistical, Mathematical, or Simulation Model
Methodology	Model Inference From Data	Hypothesis-and-Testing
Process	 Select data mining algorithm Generate data mining model from data Validate data mining model Use model for interpretation/prediction	 Design model based on experiences and assumptions Validate model or determine model parameter using staistical tests Optimize model or use model for prediction

Figure 1- 3. Comparison between Data Mining Model and Statistical, Mathematical or Simulation Model

Though data mining has been well-researched and applied in various industries, applying data mining techniques to construction equipment management faces some special challenges, such as noisy data, inconsistent data, and a lack of prior knowledge of data features. Many researches in construction industry explored the application of data mining techniques for data analysis and decision support. Soibelman and Kim (2002) conducted a systematic research on data preparation and the entire Knowledge Discovery in Database (KDD) process for construction knowledge generation; as a case study, the researchers applied the decision tree algorithm C4.5 for evaluation of construction delays in pipeline installation. Caldas et al. (2002) proposed an automated approach for classification of construction documents through the integration of Support Vector Machine (SVM) with a model-based information system. Lu and AbouRizk applied artificial neural network for estimating construction productivity (Lu and AbouRizk 2000); Wilmot and Mei conducted research on highway cost estimation with neural network (Wilmot and Mei 2005); Lee et al. investigated the application of decision tree to classify and quantify cumulative impact of change orders on productivity (Lee et al. 2004).

Not all the data mining algorithms can be readily applied without thoughtfully considering the specific application scenario and data features. For example, one of the most interesting data mining tasks in equipment management is outlier mining. An outlier is defined as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1986). Searching, sorting and ranking outliers in an equipment database can identify problems in equipment field operations, equipment performance, and management decisions. Yet neither traditional statistical methods nor current outlier mining algorithms can provide flexible and reliable solutions when applied to equipment datasets due to their stringent pre-assumptions on data distributions or sensitivity of outlier mining results to the input parameters. An outlier mining algorithm that caters to the special features of equipment data is required.

RESEARCH OBJECTIVES

The research will attain the following objectives:

- Design and implement a prototype intelligent equipment management system, and make recommendations on system planning and design
- Build a prototype equipment data warehouse and data warehousing system for interactive data analysis and need-based fast information retrieval
- Build automated decision support modules using data mining techniques for knowledge generation, validation, sharing, and utilization in an equipment information management system
- Design and test a novel outlier mining algorithm for anomaly detection from equipment database

RESEARCH METHODOLOGIES

To attain the aforementioned research objectives, the following methodologies will be utilized throughout the research:

Literature Review for Problem Identification

A comprehensive review will be conducted on the recent research, development, and applications that focus on three areas. First, new enabling technologies in construction equipment operations and information management will be reviewed. The focus will be on publications from construction equipment research, current commercial construction equipment information management systems, and current commercial products for data collection and delivery; Second, the area of data warehousing. The literature review will cover the state-of-the-art, application reports, and research publications related to the application of data warehousing for decision support in various industries, especially science and engineering, as well as current research in conceptual modeling for data warehouse and data warehousing systems; Third, this research will review literature in the area of data mining. Various data mining algorithms, especially predictive data mining along with their applications and limitations, will be reviewed. Regarding outlier mining, a thorough review will be conducted on the current outlier algorithms, the special needs, and the challenges found in general engineering applications.

Site Visit for Fact Finding

Both short-term and occasional visits to the collaborating contractor's job-sites are planned in this research for the purpose of fact finding. Site visits will identify the contractors' current practice in: (1) the use of new technologies in equipment data collection, dissemination and exploitation; (2) the computerized construction equipment management; and (3) the decision support for equipment management. The findings will serve as the basic foundation and motivation for this research.

Interview and Meetings with Domain Experts for Problem Formulation

Interviews and meetings with the various levels of decision support people are set up to identify their problems and needs for decision support. Interviews with equipment manager, manager's assistant, superintendents, foremen, and mechanics are arranged to clarify problems in decision making at various levels and to formulate the research problem in order to bridge the gap between the current and expected IT architecture and tools.

Computer Modeling and Programming for Design, Experimentation, and Validation

Computer modeling and programming is implemented using appropriate computer tools, theories and methodologies in data warehousing and data mining.

Computer tools required

The research involves a number of computer tools used for design, implementation and testing at the conceptual level, or for programming, experimentation, and validation at a detailed, algorithmatic level. The computer tools needed for the research fall into two categories:

Development environment and platform – Microsoft SQL Server 2005 and Microsoft Visual Studio 2005 are used for design, development, and testing in this research. Microsoft SQL Server 2005 is a comprehensive database tools comprised of database management system, data warehousing system, and data mining. It will be used for development of the proposed equipment data warehouse and to design intelligent data mining modules. Microsoft Visual Studio 2005 is an integrated development environment based on .NET framework 2.0 for quick application development; and

Programming languages – Visual C# and Visual C++ are used, respectively, for application development and algorithm design. In addition, specialized query languages are used for data retrieval, information processing and knowledge discovery through various Application Programming Interfaces (API). Specifically, Structural Query Language (SQL) is used for data retrieval and manipulation; Multidimensional Expressions (MDX), for information processing from multidimensional data cubes; and Data Mining Extensions to SQL (DMX) for data mining definition and implementation.

Modeling and design of equipment data warehouse

The current equipment data which can be found in three heterogeneous sources will be integrated into an equipment data warehouse through the data Extraction, Transformation/Translation, and Loading (ETL) process. The data in the equipment data warehouse are organized into subjects, such as equipment cost, inventory, and fuel consumption, and exposed using the OLAP server for report generation, interactive analysis, knowledge extraction, or other analytic applications, as shown in Figure 1-4.

Modeling and design of intelligent modules

Building intelligent data mining modules is an iterative process, as illustrated in Figure 1-5. There are five coherent and repetitive steps, which need to be followed, extended, and documented in mining knowledge from equipment data.

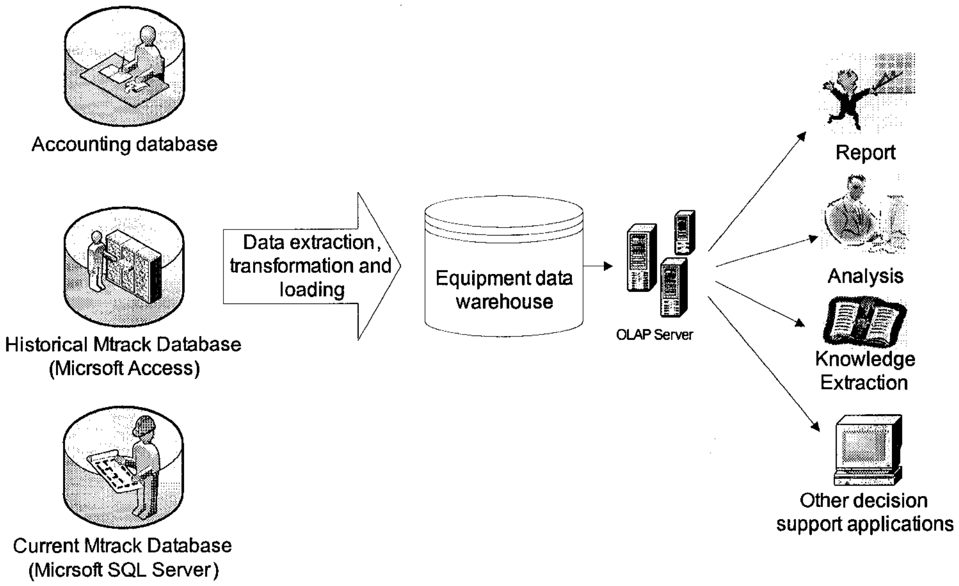


Figure 1- 4. Equipment Data Warehousing for Decision Support

(Use equipment information architecture of the collaborating contractor as an example)

- *Problem definition* – Decision support problem must be clearly defined and translated into a category of data mining problems, whether to summarize and characterize data, or to learn from data for predictive analysis;

- *Data preparation* – Prepare data to support the defined data mining problem, the data should contain sufficient information, represented in appropriate formats, and pre-processed with good quality;

- *Modeling* – Select one or more data mining algorithms to sift through data for discovering primitive knowledge. If meaningful knowledge cannot be generated from data, check if data are well prepared, and rectify problems in feature selection and representation, and data quantity and quality;

- *Evaluation* – The mined knowledge needs to be evaluated to be effective, whether by human interpretation, or independent validation tests. Otherwise modify or redefine the data mining problem; and finally,

- *Deployment* – The data mining models are used for generating reports on findings or integrated into an information system for intelligent decision support.

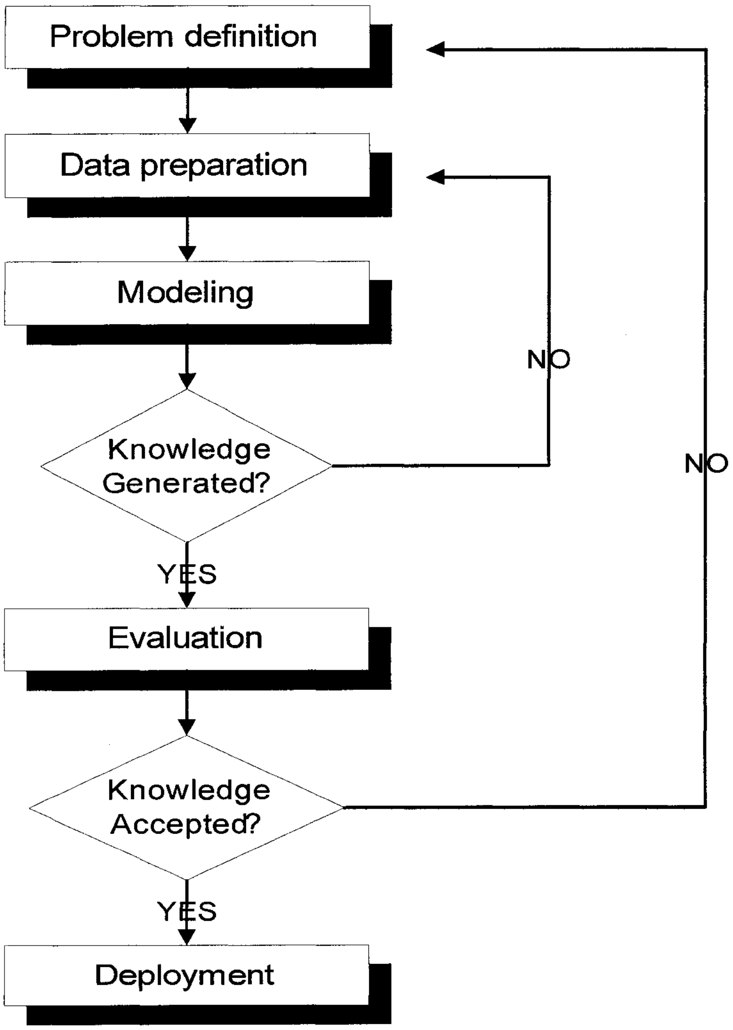


Figure 1- 5. Procedures for Building Data Mining Modules

Design and test of a novel outlier mining algorithm for engineering applications

The proposed outlier mining algorithm is based on the concept of "resolution change," which is used by Foss and Zaiane (2002) in a nonparametric clustering algorithm. For a dataset containing both in-cluster data and outliers, each data point exhibits different behaviour when the dataset is scaled up (multiplied by a resolution factor larger than 1), or scaled down (multiplied by a resolution factor smaller than 1). From the outlier view of a dataset, all the data objects in the dataset are outliers when the resolution is high enough to warrant no neighbours (by the measure of distance) for any data objects in the dataset. Meanwhile, all the data objects are "inliers" when the resolution is low enough to have all the data objects close-packed into a single cluster. If the resolution of a dataset changes, different outliers demonstrate different cluster-related behaviours during resolution change. Objects that are more isolated, with fewer neighbours, and further away from large data communities are more liable to be outliers. On the other hand, the top outliers will be merged into a cluster later when the resolution is decreased. As a result, the accumulated cluster-related properties collected on one object can be used to measure its degree of outlying relative to its close neighbourhood and community.

A factor for measuring the degree of outlying will be defined in the research. Based on this definition, and by consecutively changing the resolution of the dataset, an outlier mining algorithm is proposed, designed, and implemented for detecting top-listed outliers from the equipment dataset.

ACADEMIC CONTRIBUTIONS

By solving the defined problem through the proposed methodologies, this research makes original contributions to the body of knowledge in construction equipment management in the following aspects:

- Proposing a framework for intelligent decision support in construction equipment management based on the current information architecture of heavy construction contractors. The framework supports intelligent construction equipment management by turning current and historical equipment data into as-needed information and knowledge using data warehousing and data mining technologies; The proposed framework is introduced in Chapter 3.

- Systematically studying the necessity and feasibility of applying the data warehousing technology for decision support in construction equipment management. Through a case study on the collaborating contractor with regard to its current practice in equipment operations and its needs of improved decision support in construction equipment management, this research investigates the opportunities and challenges in designing and implementing a conceptual "equipment data warehouse" for an integrated centralized data source, interactive information retrieval and data analysis. In specific, the research proposed the general methodology of identifying and modeling subjects based on construction equipment management processes so that all the interested subjects as well as underlying facts and dimensions are represented appropriately and sufficiently in the data warehouse model. These contributions are iterated in Chapter 4.

- Assessing different data mining techniques, especially predictive data mining, for application in construction equipment for fact-based decision support. By means of extensive testing and experimentation in the computer, the research addresses two important issues in mining equipment data for decision support: (i) how to adapt the current data mining algorithms for knowledge generation from equipment data. Recommendations and suggestions are made with regard to algorithm evaluation, parameter tuning, and the validation of mined knowledge; and (ii) how to integrate the data mining modules into the current information system. A best practice is proposed regarding the design and deployment of data mining modules for automated predictive analysis, the design of Graphical User Interface (GUI) for man-machine interaction through the model visualization, inference explanation, and flexible user control. These contributions are detailed in Chapter 5.

- Proposing a novel outlier mining algorithm suited for engineering applications for generic problem detection in equipment data or data preprocessing. The proposed outlier mining algorithm has features such as no input parameters, being capable of handling datasets containing data clusters of arbitrary shapes, mining, and ranking the top-*n* most interesting outliers. Chapter 6 introduces the proposed algorithm and its contributions to the body of knowledge.

INDUSTRIAL CONTRIBUTIONS

As a research project based on a real life case of the collaborating contractor, the findings and recommendations made in the research are expected to contribute to the construction industry through improved data sharing and exploitation, as-needed information and knowledge, and the proposed concept of intelligent equipment management. The research is expected to bridge the gap between the practical need for actionable information and knowledge, and the inability to provide needed information and knowledge using the current equipment information management architecture, specifically:

- Scattered equipment data located in different systems in different formats, which are difficult to retrieve for complex decision analysis, are integrated into a centralized repository using consistent formats and an analysis-oriented data structure;

- The static reports/displays or those generated with limited user flexibility for equipment management are replaced with an interactive system over which the user has complete control in information retrieval, i.e., it can retrieve information from any perspective of the equipment operations and at various levels of granularity; and

- Integration of data mining into the current information system automates the process of knowledge discovery from data and the utilization of the mined knowledge for predictive and explanatory analysis. Some error-prone, experience-dependent decision-making tasks can be modeled as data-driven computer modules and incorporated into the current equipment information management system.

CONCLUSIONS

Based on the current industrial practice of construction equipment management, and the wide applications of communication and information technologies, this research addresses the common decision support problems existing among large heavy construction contractors. These problems include scattered and inconsistent data, the lack of interactive analysis tools for flexible information retrieval, and the lack of tools for automated knowledge generation from equipment data. The research addresses decision support problems in construction equipment from both the higher levels of information technology architecture, and the lower levels of data mining algorithms. A conceptual framework is proposed in the research for intelligent construction equipment management based on the current information architecture. The Application of current data mining algorithms for decision support is illustrated using real-life application examples. A novel data mining algorithm is proposed in this research to detect outliers effectively from engineering data. The findings, recommendations and methods in this research can be applied to other areas of construction management outside construction equipment management to improve the effectiveness and efficiency of decision support based on large amounts of collected construction data.

REFERENCES

Bain, T., Benkovich, M., Dewson, R., Ferguson S., Graves, C., Joubert, T.J., Lee, D., Scott, M., Skoglund, R., Turley, P., and Youness, S. (2001). *Professional SQL server 2000: data warehousing with analysis services*, Wrox Press, Birmingham, AL.

Caldas, C. H., Soibelman, L. and Han J. (2002) "Automated Classification of Construction Project Documents." *J. Comput. Civ. Eng.*, ASCE, 16(4), 234-243

Russell, J. S., and Kim, S. K. (2003). *Chapter 6: Construction automation,* In *The Civil engineering handbook*, 2nd Ed., , CRC Press, Boca Raton, FL.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery: an overview." In *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA.

Foss, A., and Zaiane, O. (2002). "A parameter-less method for efficiently discovering clusters of arbitrary shape in large datasets." *Proc., International Conference on Data Mining*, IEEE, Maebashi City, Japan

Gransberg, D., Popescu, C. M., and Ryan, R. C. (2006). *Construction equipment management for engineers, estimators, and owners*, CRC Press, Boca Raton, FL.

Hawkins, D. (1980). *Identification of Outliers*, Chapman and Hall, London, U.K.

Inmon, W. H. (2005). *Building the data warehouse*, 4th Ed., Wiley Publishing, Indianapolis, IN.

Lee, M., Hanna, A.S. and Loh, W.Y. (2004). "Decision Tree Approach to Classify and Quantify cumulative Impact of Change Orders on Productivity." *J. Comput. Civ. Eng.*, ASCE, 18(2), 132

Lu, M., AbouRizk, S.M. and Hermann U.H. (2002). "Estimating labor productivity using probability inference neural network." *J. Comput. Civ. Eng.*, ASCE, 14(4), 241-248

Reed (2007). *2007 Annual Report and Forecast.* Supplement to *Construction Equipment*, January 2007 issue, Reed Business Information, New York, NY.

Soibelman, L and Kim, H. (2002). "Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases." *J. Comput. Civ. Eng.*, ASCE, 16(1), 39-48

Vorster, M. C. and Livermore, M. E. (1994). "Executive development for equipment managers." *Proc., Equipment Resource Management into the 21st Century,* ASCE, Nashville, TN., 87-95.

Wilmot C.G. and Mei, B. (2005). "Neural Network Modeling of Highway Construction Costs." *J. Constr. Engrg. and Mgmt.*, ASCE, 131(7),765-771

CHAPTER 2. CONSTRUCTION EQUIPMENT MANAGEMENT: STATE-OF-THE-PRACTICE

INTRODUCTION

Great changes have taken place in construction equipment management due to the wide application of automation control, automated data collection and distribution, as well as computer information management systems. Although the basic principles and guidelines within the decision making process of construction equipment management remain the same, the methods and tools used for the daily operation of equipment management and decision support have changed drastically. The equipment management team has alleviated itself from the burdens of book-keeping and data crunching tasks including operational tracking, scheduling, and reporting with the implementation of various computer controls and information management systems among large equipment-owning organizations and contractors. As a result, their capabilities in data collection, data processing, and information dissemination and exploitation have improved substantially. This chapter provides an overview on the state-of-the-practice of construction equipment management from the aspects of data management, computerized equipment management, as well as decision support practices. The chapter serves a three-fold purpose in the thesis: firstly, it surveys the current practice of equipment data gathering, distribution, and exploitation, as the precondition for this research; secondly, it summarizes the decision support practices from a technological point of view, as the premise and basic platform for this research; and thirdly, the identified problems supplement the research motivations in Chapter 1 as the instigation for the research within the thesis.

CONSTRUCTION EQUIPMENT DATA MANAGEMENT

Construction equipment data are recorded facts in different facets of equipment management. The contractors make every effort to collect a full suite of equipment data in daily operations in order to reflect the history and current status-quo of equipment acquisition, maintenance, repairs, rebuilds, operations, usage, disposal and other management activities. The development of digital controls, sensor techniques, and computer technology coupled with the dramatically reduced costs of implementation in recent years has greatly stimulated the investment and application of these technologies by large equipment-owning organizations in order to increase their capabilities in equipment data collection, storage, and distribution. Traditional paper-based data have gradually become obsolete due to prevalence of electronic data used by contractors for equipment management. Nevertheless, with the flooding of large amounts of data gathered from different business aspects, most contractors encounter data problems such as poor quality, scattered sources, and underutilization due to lack of a coherent forward-looking policy on equipment data management.

The ability of companies to exploit their intangible assets is far more decisive than their ability to exploit their physical assets (Tiwana, 2002). As part of the valuable intangible assets of an equipment owner, equipment data do not merely serve the purpose of data crunching in accounting reports, but more importantly they need to be interpreted for decision support at various levels of management. A sound data management policy ensures the data are collected in the right format and delivered to the right places wherever and whenever they are needed. Data management is defined as "the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise" (DAMA 1995), which emphasizes the top-level general architecture of data management, policies, practices, and procedures guiding its implementation in the enterprise. Though equipment data management is an important area in the overall construction equipment management, it is not effectively addressed among large contractors due to the progressive nature of technology implementation. The contractors are usually conservative in investing in information technology for equipment management because of the uncertainties and rapid changes in technological evolution in recent years. Contractors accept and adopt new tools and techniques only if they are proven to be cost-effective in both the short and long term. As a result, the contractor's equipment data management is usually fragmented and ineffective, with the following symptoms:

Data quality problems – Equipment data tend to suffer from poor quality, such as redundancies, typing errors, inconsistency in structures and formats, etc. Data quality in an outdated system is usually poorer than in recent systems where vigilant data validation mechanisms are introduced.

Scattered, heterogeneous data – Equipment data may be gathered from and located in different systems. For example, the same category of data can be found in obsolete and current systems which use different mechanisms for data storage; the data describing the same equipment management problem may reside in different systems managing different business activities (e.g. for the collaborating contractor, the equipment usage data are located in the Explorer accounting system whereas the equipment fueling, repair and maintenance data are stored in the M-Track equipment management system).

Data communication and sharing problems – Some data may be captured in an isolated application which confines the data to a personal computer or mobile device. The data in this type of system have to be fetched and delivered manually to the other applications in case they are needed. This approach not only leads to inefficiency, but can also cause delays, uninformed decisions, etc.

There are four coherent components in equipment data management, as stated below:

Data Acquisition

Construction equipment data are acquired through various methods when different management activities are taking place; such as equipment maintenance and repair, logistic supplies, and equipment allocation and operations. Traditional manual approaches for record-keeping have been gradually replaced by automatic methods for improved efficiency and effectiveness. The major categories of data acquisition methods include:

Scanning techniques

Many identification techniques are currently used for the automatic tracking of fueling history, inventory control of parts, etc. For example, magnetic cards are used for either fueling at gas stations (e.g. SuperPass® issued by Petro-Canada (2007)) or on the job-site using fueling vehicles. Later the fuel management data can be automatically transferred to a web server for downloading through the internet; Another example is the bar coding widely used for the inventory control of parts/fluids in the warehouse, through which the inventory information is updated through an inventory management system in real time; The Radio Frequency Identification (RFID) technique is another example of an identification technique used and a recent development capable of providing the topnotch features of speed identification and bulk processing in product tracking. There is a belief that the RFID technique will eventually replace bar-coding in all industries, including the construction industry.

On-board digital control and sensors

Modern construction machines are equipped with sophisticated controls, monitoring and diagnostic systems. Large construction equipment manufacturers and independent developers offer electronic data collection products (e.g. Product Link®

(Caterpillar, 2007), KOMTRAX® (Komatsu, 2006), GlobalTRACS® (QUALCOMM, 2006)) which are either pre-installed in the construction equipment, or can be purchased for installation on current equipment in order to collect real time data on machine health conditions, operation status, and utilization, including:

- Current fuel and fluid levels
- Health data on engine, transmission and hydraulics
- Critical temperatures and pressures
- Up-to-the-minute engine running time, including idle time, production time, as well as working time under different engine-loading conditions
- Diagnostic results

Data can also be collected automatically on the project or from the work environment through the construction equipment. For example, Bomag Gmbh (2006) designed VARIOCONTROL® in the roller for the Intelligent Compaction (IC) of soil or asphalt concrete. A set of accelerometers are installed on the roller to collect data on the stiffness characteristics of the material it is compacting by measuring both the horizontal and vertical reaction of the drum. The stiffness of the compaction material is then used for record keeping and achieving the optimum compaction over the work surface by adjusting the energy output of the vibrating roller.

GPS locating

The Global Positioning System (GPS) is a satellite navigation system including a total of 24 GPS satellites divided into six orbits with four satellites distributed equally on each orbit, as well as ground control stations and GPS receivers (Tsui 1995). The location (longitude and latitude) and level (altitude) of a point indicated by a GPS receiver is determined by measuring the distances from the receiver to three or more satellites and computing results based on the principle of triangulation. The GPS locating technique is used in construction equipment for the purposes of:

- Tracking the machine location in real time, for security monitoring, maintenance and repair, or deployment
- Controlling the operations of the machine. Accurate GPS locating makes it possible to control machine operations based on the design configuration. For example, the AccuGrade® GPS (Caterpillar, 2007b) makes it possible to determine the real time blade location of a dozer through two GPS receivers installed on each side of the blade. The blade motion parameters are determined automatically in the control system or by the operator based on the on-screen display of both the current and designed ground contours.

Manual approach

Equipment data are also collected manually in many situations. Using personal computers, or laptops, the equipment management inputs data on equipment scheduling, work order placement, maintenance and repair records, etc. into the equipment information management system. Using portable electronic devices, such as a pocket PC, or smart phone, people on the job-sites keep records of machine usage and labor hours in electronic form. Record-keeping using speech recognition through handheld devices have had some applications reported, but is still in the research stage. This technology will enable data collection through human speech in certain special work conditions.

Data Delivery and Distribution

As soon as the data are acquired, they should be delivered and distributed to the end users for utilization. Equipment data are not only delivered to human users through various applications, including equipment information management system, accounting system, notification services, but also to various automation control systems, or are transferred from one application to another. The technologies for data delivery and distribution include:

Cellular data transmission – The equipment data can be transmitted through the current analog cellular telephone system using a circuit-switched or packet-switched cellular modem. For example, a smart phone or pocket PC (through the cellular phone modem) can be used to retrieve data, fill out a form, or place an order at a remote site.

The Internet – Equipment data can be transmitted over wired or wireless Internet using a TCP/IP encoding/decoding standard. The rapid development of the Internet makes it possible to transmit data over the Internet at a high speed in a global context. Web-based equipment management has gained industry-wide acceptance for equipment management due to unique Internet features, such as good accessibility and a centralized system. In addition, wireless Internet provides the added benefit of mobility.

Local Area Network – A Local Area Network (LAN) is made up of interconnected computers which share data, applications, and other network resources in a confined area, such as office buildings or an equipment deployment and service center. A LAN has a high speed of data exchange but has limited coverage. In the small job-sites, a wireless LAN, along with a shared internet connection, is a good option for data transfer and information exchange.

Satellite Communications – Data communications through Low-Earth-Orbit (LEO) satellites provide the opportunity to exchange for construction equipment in remote sites with poor accessibility. A constellation of LEO satellites provides full coverage to the surface of the earth so that data can be relayed freely through the LEO satellites from one place to another without geographical boundaries.

Data delivery and distribution was once a serious hurdle in the path of attaining effective construction equipment management due to highly scattered job-sites and poor communication in the past; however the recent development and application of various data communication techniques have dramatically altered this situation. KOMTRAX® (Komatsu 2006) is a representative application that illustrates how data communication technologies change the way construction equipment is managed. This equipment monitoring and management system uses the data exchange infrastructure shown in Figure 2-1 to collect, distribute and utilize the data in near real-time, with the following uninterrupted flow of data in the system:

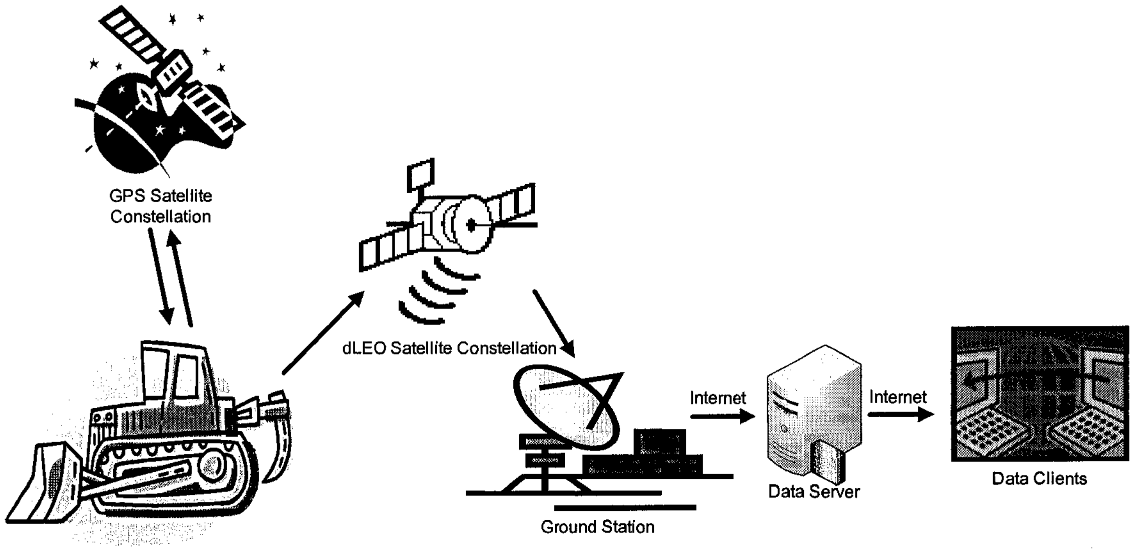


Figure 2- 1. KOMTRAX® Data Communication System for Equipment Management

1. The GPS receiver installed in the equipment collects the machine location data using the GPS navigation system;

2. KOMTRAX® collects data on equipment health, critical machine information, and operational statues;

3. All the equipment data acquired are relayed to a ground station through ORBCOMM® LEO satellites (ORBCOMM 2007);

4. The equipment data are transmitted to KOMTRAX® global data server;

5. The equipment data are distributed to KOMTRAX® local data server;

6. The equipment data are published on the Internet through a web-based equipment management system for decision analysis.

Data Storage

Depending on the applications where the data are generated, equipment data are stored in different media and appear in different formats. Most of the equipment data are physically stored in the hard disks of large server systems or personal computers, but they may also be stored in portable storage media (e.g. CD, DVD) or handheld electronic devices (e.g. pocket PC). On the other hand, the data can also be stored in different formats, such as a relational database, applications, spreadsheets, text files, or even unstructured documents and emails.

Data Exploitation

Most equipment data are collected in a specific business process of construction equipment management, but are used for multiple purposes. For example, equipment maintenance and repair data are kept as historical records for each piece of equipment. Besides the accounting, the data are also utilized for decisions on equipment repair/overhaul, purchase/replacement, rental/purchase decisions, financial budgeting, and the estimation of equipment cost in bidding documents. Generally speaking, the utilization of equipment data falls into the following three categories:

Automation control

Equipment automation control, being an important part of construction automation, relies on the collected equipment data (e.g. location data, machine status data, image data) for programmed equipment operations and control. Examples of this application include construction robots, intelligent compaction, and automatic earthmoving control systems.

Operation support

Operation support is the primary driving force for data collection. Equipment data are collected to track and control the daily operations of construction equipment, such as inventory control, equipment utilization, work order/purchase order placement, and accounting. Different business activities of equipment management generate different data depicting different facets of equipment management over time, which collectively give a panoramic and multidimensional view of the contractors' equipment management process.

Decision support

Decision support is an important yet underemphasized utilization of equipment data. Neither researchers nor industry practitioners have dedicated enough effort to the decision support in equipment management from a data perspective. Operational data in their simplest forms merely describe what happens in daily equipment management, but do not explain why things happen, how to improve situations in the future, or what will likely happen next. The retrieval of needed information and the generation of knowledge (trends, patterns, rules) from operational data for decision support are only partially addressed in the current equipment management IT architecture and tools, and not in a systematic manner. The current practice of decision support relies more on experience and personal interpretation of available data and less on actionable information and knowledge which are embedded in data and explicit without the aid of powerful computer tools.

The importance of construction equipment data management has gained wide recognition among large contractors thanks to the drastically increased capability of collecting large amounts of equipment data. The development of information and communication technologies has enabled the delivery, distribution, storage, and exploitation of collected data in near real-time without geographical limitations. Contractors need to shift their attention from collecting all the needed data to the timely delivery/distribution of these data to where they are needed, as well as the use of these data for fact-based decision support. To fulfill the objective of data management (i.e. data integrity, data sharing, data transformation and exploitation), a long term and far-sighted data management policy and IT architecture needs to be developed and put in place.

COMPUTERIZED CONSTRUCTION EQUIPMENT MANAGEMENT

Computerized Construction Equipment Management (CCEM) refers to the practice of managing construction equipment using computer software solutions. Computerized construction equipment software was first applied in the 1990s to replace repetitive error-prone book keeping functions and to generate preformatted screen display and equipment reports in the construction industry. Figure 2-2 shows the typical functional design of a computerized equipment information management system. The major inputs and outputs of a CCEM system include:

Equipment tracking – all the equipment specifications, attachments, current locations, maintenance and repair states, and service meter readings, are logged in the computer for up-to-date information on every piece of equipment in the fleet. The equipment can be organized by divisions and departments or categories and classes;

Preventive Maintenance – Preventive Maintenance (PM) is scheduled based on product specifications and suggestions from the manufacturer;

Fuel, Oil, and Grease Tracking – the consumption of Fuel, Oil, and Grease (F.O.G.) on each piece of equipment are logged for tracking;

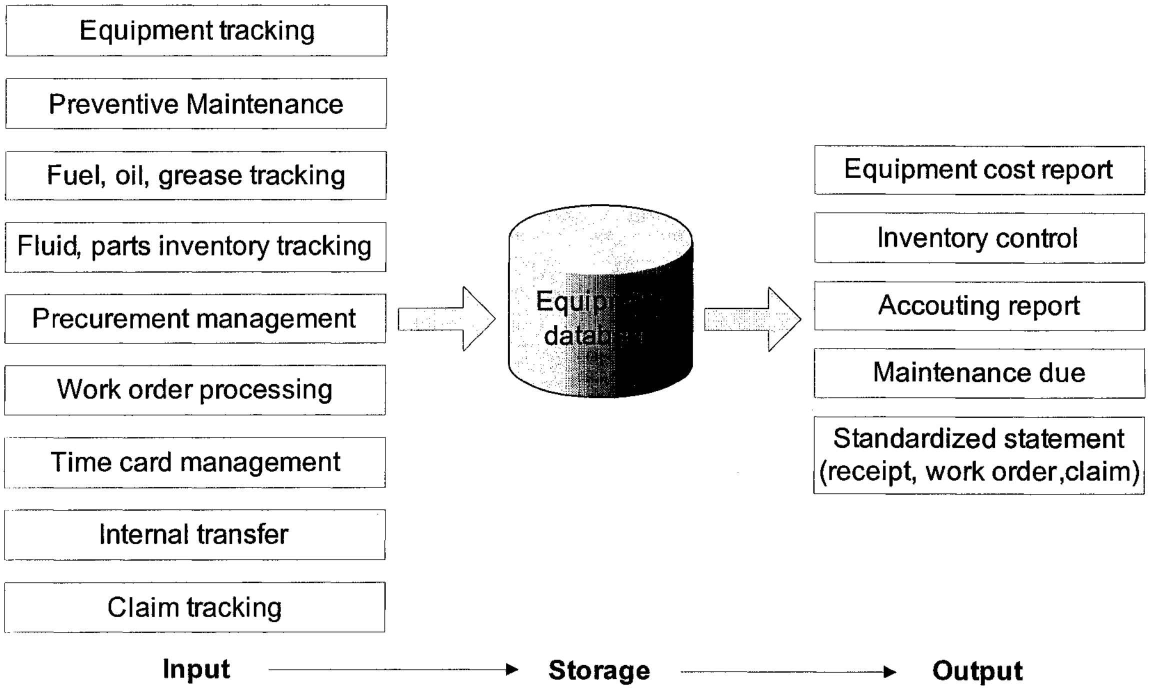


Figure 2- 2. Typical Functional Design of a CCEM System

Procurement management – Cash purchases, or purchase orders for equipment logistics can be managed through computers, including order placement, receiving, and storage;

Work order processing – Work orders for equipment repair or overhaul can be placed in a computer with estimated parts and labor hours. After processing, a report is generated on the completed work order and is compared with the previously estimated costs;

Time card entry – The daily working records are entered by the employee and approved by the superintendent;

Equipment cost report – All the equipment costs can be reported based on the report types and parameters selected and defined by the user.

Computer solutions for equipment management have eliminated most of the paper work needed for book keeping and data crunching, and have greatly improved efficiency in the tracking and managing of equipment operations. However, the basic input-output computer model of the earlier systems (taking in data and generating reports of these data) did not significantly improve the decision making capabilities of the equipment management team.

With the rapid development in information management technology and the World Wide Web (WWW), the computerized equipment management solutions have been advanced in many aspects, such as enhanced functionality, application integration, and other decision support features. A literature review was conducted in this research on the current commercial or custom-developed CCEM systems, with the objective of identifying the current status of computerized equipment management solutions and needs for research and development. The representative CCEM systems reviewed include:

- EQUIPMENT MANAGER; a web-based equipment information management system developed by Caterpillar Inc. (Caterpillar 2007a).

- GlobalTRACS® equipment management system; a web-based equipment management system developed by QUALCOMM Inc. (QUALCOMM 2006).

- eCMS™ ; An entire suite of web-based construction management system developed by Computer Guidance Corporation (Computer Guidance 2006). An equipment management module is included within eCMS™.

- Explorer Computerized Maintenance Management System; a computer solution for equipment management and control developed by the Explorer Software Inc. (Explorer 2007).

- M-Track; a multifaceted equipment management system custom-developed by the NSERC/Alberta Construction Industry Research Chair for the operations of Standard General Inc. (NSERC/Alberta Construction Industry Research Chair 2007).

The literature review found out that, in addition to the traditional modules supporting daily equipment operations, the current CCEM systems have some new plausible features, including:

Web-based solutions – while traditional CCEM solutions are windows-based applications requiring installation on each personal computer, some of the recent systems are web-based. A web-based CCEM system is deployed into a central web server so that it can be accessed online from anywhere, with any device, at any time provided there is a web browser and internet connection. Web-based CCEM solutions fit better into the environment of construction equipment by providing a centralized data repository, better accessibility, and reduced software maintenance. Such examples include the EQUIPMENT MANAGER (Caterpillar 2007), and eCMS™ (Computer Guidance 2005).

Product Integration – The data collection devices, such as Product Link (Caterpillar, 2006), KOMTRAX® (Komasu 2006), and GlobalTRACS® (QUALCOMM 2007) can be installed on-board and connected to a machine's engine for the collection of data on equipment status, operations, etc. Developers provide CCEM solutions along with these hardware products in order to fully harness the power of collected data for near real time equipment management.

Application Integration – Construction equipment management is an integral part of construction management as equipment management activities are inter-dependent in dealing with other missions in project construction, such as accounting, progress, and cost control. Many developers provide CCEM solutions as one of the core components of an entire suite of project management solutions. Examples include eCMS™ (Computer Guidance 2007), and the Explorer Computerized Maintenance Management System (Explorer 2005).

Information retrieval and interactive analysis – while the traditional equipment reporting and operation support remain as the primary system outputs, new features are added in the current systems for improved decision support, such as

- Flexible information retrieval, wherein multiple pre-designed views of equipment data are possible;

- Information dashboards that include critical indicators of equipment operations, such as accumulated fleet costs, equipment utilization profiling, inventory status, are presented and updated on requests or at pre-defined time intervals;

- Interactive data analysis where users are not limited to static reports. Interactive features such as sorting, filtering, roll-up/drill-down, and parameter-driven report changes, are added to allow the user to manipulate data in flexible ways;

- Predictive maintenance wherein the Preventive Maintenance (PM) program is integrated with oil sampling analysis reports to adjust the intervals of PM or replace worn parts proactively before they cause major damages;

- Notification services where data are used to trigger a decision event and notify the concerned parties through a pager, cellular phone, email, etc. Examples include a PM event being due, unauthorized use of the machine, theft-warning by geo-fencing, abnormal equipment conditions, or improper usage.

DECISION SUPPORT IN CONSTRUCTION EQUIPMENT MANAGEMENT

There are two levels of decision support in construction equipment management: the operational level and the strategic level. Decisions at the operational level keep the equipment in service (i.e. maintenance and repair, logistics supply, allocation, and utilization) in order to satisfy the needs of the construction project; whereas, decisions at the strategic level are made for equipment acquisition, financing, disposal and replacement, and life cycle cost analysis, in order to maximize the return on the investment.

Both the aforementioned equipment data management and computerized equipment management address the decision support problem of how the construction equipment fleet is to be managed and managed cost-effectively. The data management emphasizes the data sharing and timely access and the computerized equipment management is concerned with the IT infrastructure in order to present data to decision makers in interpretable forms (i.e. information and knowledge). Generally, the information and knowledge needed for decision support can be obtained by the equipment management team in two ways: push and pull. Push means the information and knowledge is delivered to the decision makers in the correct, readily usable form at the proper time; whereas pull means the decision makers are provided with the data and tools to extract the needed information and knowledge through interactive, explorative data analysis and automation techniques.

The state-of-the-practice in decision support of construction equipment management is summarized from the perspectives of pushed and pulled information and knowledge:

Pushed Information and Knowledge

Some decision support features of the current IT architecture, as explained previously, are designed to satisfy the needs of decision makers. Most of this pre-configured information and knowledge is pushed information and knowledge, regardless of the ways they are delivered (e.g. web-platform, automated notification, and information dashboards). Pushed information and knowledge is highly preferred for construction equipment because it is readily available for decision making purposes. However, this is only feasible if the equipment data can be turned into useful information and knowledge in an automatic approach. For example, data are processed to trigger a decision support event, or generate pre-configured strategic information. Currently the decision support at the operational level is well-supported by pushed information and knowledge.

Pulled Information and Knowledge

Not all the equipment decision making activities are well supported by pushed information and knowledge. In many cases, one question posted by the decision maker has to be answered by looking at the equipment operational data from different viewpoints in unanticipated ways in order to arrive at a conclusion. Many problems in decision making activities cannot be predefined and are constantly changing with the uncertainties in internal operations and market conditions. Pulled information and knowledge supplements pushed information and knowledge from a technical perspective and can be obtained by the decision makers in their own way based on specific situation and information needs.

The current IT infrastructure of the contractors provides some support, but to limited degree, for the generation of pulled information and knowledge. Some examples of tools or platforms include: (i) the interactive features in some of the current equipment information management systems. The decision makers can look at the equipment cost data in multiple views or perform a “what-if” analysis in the life cycle cost analysis module. (ii) The current systems support the downloading of data to other commercial tools (e.g. spreadsheets, statistical package, fleet selection and optimization software, and simulation tools) for in-depth data analysis.

If the tools for the generation of pulled information and knowledge are not available, the equipment manager and the management team have to interpret the operational data and printed reports through personal judgment or meetings. As a matter of fact, the current practice of decision support relies heavily on personal know-how and experience in decision making.

CONCLUSIONS

A comprehensive literature review and analysis is conducted in this chapter on the state-of-the-practice of construction equipment management from three pragmatic points of view: equipment data management, computerized equipment management, and decision support. The focus is on the changes brought to the current practice of construction equipment by the industry-wide application of automated data collection, automation control, and information technology. While indicating the positive changes in equipment management through a review of recent technologies, commercial systems and solutions, the chapter also identifies the needs and new challenges for decision support in equipment management. To put it in another way, although the current IT architecture generates and delivers some readily-usable pushed information and knowledge for decision support, there is also an emerging need for the generation of pulled information and knowledge based on the powerful interactive data analysis tools and automation techniques.

REFERENCES

BOMAG (2007). "Systems for soil compaction." BOMAG GmbH, <http://www.bomag.com/worldwide/index.aspx?fm=%2fworldwide%2fproducts%2f variocontrol.aspx%3fflash%3dfalse&DID=100000000&Lang=10000 > (May 7, 2007).

Caterpillar (2007). "Product Link system." Caterpillar Inc., <http://www.cat.com/cda/layout?m=37506&x=7 > (May 7, 2007).

Caterpillar (2007a). "EQUIPMENT MANAGER." Caterpillar Inc., <http://www.cat.com/cda/layout?m=44501&x=7 > (May 7, 2007).

Caterpillar (2007b). "AccuGrade® grade control system." Caterpillar Inc., <http://www.cat.com/cda/layout?m=37483&x=7> (May 7, 2007).

Computer Guidance (2005). "Financial management." Computer Guidance Corporation, <http://www.computerguidance.com/products/financial_ management.aspx> (May 7, 2007).

DAMA (1995). *A model for data resource management guidelines*, 2nd Ed., Chicago IL DAMA Chapter, Data Management Association, Lutz, FL.

Explorer (2006). “Computerized maintenance management system.” Explorer Software Inc., <http://www.explorer-software.com/prod_equipment_ control.shtml> (May 7, 2007).

KOMTRAX (2007). “KOMTRAX.” Komatsu America Corp., <http://www.komatsuamerica.com/komtrax-home.asp> (May 7, 2007).

NSERC/Alberta Construction Industry Research Chair (2007). “M-Track.” NSERC/Alberta Construction Industry Research Chair, <http://www.construction.ualberta.ca/forum.shtml> (May 7, 2007).

ORBCOMM (2007). “Heavy equipment.” ORBCOMM Inc., <http://www.orbcomm.com/solutions/heavyEquipment.htm> (May 7, 2007).

Petro Canada (2007). “SuperPass® Card.” Petro Canada Inc., <http://www.petro-canada.ca/en/productsandservices/257.aspx> (May 7, 2007).

QUALCOMM (2007). “GlobalTRACS® equipment management system.” QUALCOMM Inc., <http://www.qualcomm.com/technology/assetmanagement/platforms/globaltracs.html> (May 7, 2007).

Tiwana, A. (2002). *The Knowledge management toolkit*, 2nd Ed., Prentice Hall, Upper Saddle River, NJ.

Tsui, J. B. (1995). *Fundamentals of global positioning system receivers*, 2nd Ed., John Wiley & Sons, New York, NJ.

CHAPTER 3. BUILDING INTELLIGENT APPLICATIONS FOR CONSTRUCTION EQUIPMENT MANAGEMENT[1]

INTRODUCTION

Construction equipment management involves the management of equipment resources to achieve the maximum return on fixed assets, and to satisfy the equipment needs of projects for large contractors in a timely and cost effective manner. Owning and operating a large fleet comprised of hundreds or even thousands of pieces of heavy equipment is common for large contractors. As such, decisions regarding common equipment management tasks, including equipment acquisition, maintenance, repair, allocation, operation, and replacement/disposal, are not trivial and may imply a

[1] This chapter is based on the following two published conference papers:

- H., Fan, S., AbouRizk and H., Kim (2007). "Building intelligent applications for construction equipment management." Proceedings of the ASCE International Workshop on Computing in Civil Engineering, July 25-27, 2007, Pittsburg, PA.
- H., Fan, H., Kim and S., AbouRizk (2006). "Model-based recommendation system in support of construction equipment management: a case study." Proceedings of the Joint International Conference on Computing and Decision Making in Civil and Building Engineering, June 14-16, 2006, Montreal, QC. , 1997-2006

The writer of this thesis is the primary contributor and writer of these two papers, the other authors provided supervisory works.

significant financial impact on contractors. Due to the unique, dynamic nature of construction equipment management, these decisions rely primarily on the expertise and experience of the equipment manager and equipment management team.

The wide use of various construction equipment information management systems has simplified most of the daily chores involved in equipment management, including equipment operation data collection, inventory control, equipment tracking, maintenance planning, invoicing and cost reporting. Construction equipment data management serves as an excellent example to illustrate this innovation. Currently, most of the data on equipment conditions, operations, maintenance, and repair can be collected, stored, and transferred in an electronic format with the use of on-board computer control, Global Positioning System (GPS), scanning devices, wireless networks, as well as hand-held electronic devices. The ability to collect an entire suite of equipment data allows a contractor to keep full records of equipment fleet operations and management. However, the value of the data repository cannot be realized until the data are interpreted and converted into actionable information and knowledge. A review of the related literature indicates that most commercial solutions for construction equipment management focus on the functionality of taking data as input and generating reports as output, with a lack of exploratory data analysis and knowledge generation capability.

Intelligent applications for construction equipment management are knowledge-based systems designed to enhance and enrich the current information systems in decision support. An intelligent system is defined as a knowledge-based system, computational intelligence-based, or their hybrids (Hopgood 2001). This chapter introduces a conceptual framework for building intelligent applications for equipment management based on both pre-defined knowledge and computational intelligence. Knowledge-based response or inference is similar to an expert or rule-based system, with decision rules embedded into the system to generate a response or an inference. The rules are usually static and explicit. Examples include scheduling of preventive maintenance events triggered by multiple rules, or recommendations on component replacement based on trend analysis of oil sampling tests. Computational Intelligence embeds data mining modules into the system which is designed to sift through data to obtain rules or knowledge for interpretation or prediction. The knowledge obtained using the data mining techniques is usually dynamic and implicit. For example, data mining is used to uncover the relationship between equipment residual value and its influencing factors, such as age, make, model, and microeconomic indicators.

The use of pre-defined rules for decision support in equipment management is introduced in related research (such as Berzonsky (1990)) and has been adopted in many current equipment information management systems. Therefore, the chapter introduces the general framework for a proposed intelligent system, but with a focus on the incorporation of data mining based computational intelligence for decision support. A case study on automated evaluation of work orders is conducted to illustrate how data mining is used for automated knowledge generation and utilization in a construction equipment information management system.

LITERATURE REVIEW

Research and development in construction equipment management is mainly focused on automation and robotic techniques, real-time data communications, and information processing. The AccuGrade® grade control system, from Caterpillar Inc., facilitates automatic control of the finish grade in earthmoving operations (Caterpillar 2006); the GlobalTRACS® equipment management system, from QUALCOMM Inc., integrates fleet data with the back office information system, and enables the equipment conditions and operations data to be transmitted and monitored in real time (QUALCOMM 2006). However, from a decision support perspective, these processes merely involve simple data interpretation for automation control, or information processing for decision support, but no generation of new knowledge.

Advanced data analysis and knowledge generation in equipment management are heavily dependent on spreadsheets and statistical tools. Examples include: identifying statistical relations between equipment maintenance and repair cost and fuel consumption for replace/repair decisions on equipment (Gillespie and Hyde 2004); and finding relations between equipment residual value and its factors of impact for the prediction of equipment residual value (Lucko et al. 2006).

Decision support based on pre-defined rules has been employed in computerized equipment management. M-Track (NSERC/Alberta Construction Industry Research Chair 2007) allows decision makers to set up multiple rules for the scheduling of preventive maintenance events. The Vital Information Management System (Caterpillar 2006a) alerts and instructs the operator to take appropriate action based on a detected impediment or abnormal equipment conditions. Berzonsky (1990) introduces a rule-based electrical diagnostic system for maintenance of mining equipment.

Intelligent decision support has been intensely researched in the area of computerized maintenance management, which deals with the maintenance operations such facilities as power plants, industrial plants, and military establishments. Due to the unexpected failure of major components, the incorporation of intelligent data analysis modules into a computerized maintenance management system is essential. The objective of this endeavor is to predict the deterioration of conditions based on monitoring data so as to adjust the maintenance schedule proactively or to repair/replace the components prior to failure. These intelligent systems utilize tools in artificial intelligence, such as Artificial Neural Network (ANN) (Fu et al. 2004); Recursive Neural Network (RNN) (Yam et al. 2001); and Bayesian Network (Zhang et al. 1997), for decision support based on decision models generated from the operational records.

CONSTRUCTION EQUIPMENT MANAGEMENT VERSUS DATA MINING

The equipment management has a dual responsibility: long-term decisions in equipment selection, finance, and life cycle costing are made at the corporate level, whereas short-term decisions on logistics, maintenance, and repair are made at the operational level (Voster and Livermore 1994). The fluid, dynamic characteristics of equipment management are highly influenced by uncertainties in areas such as construction projects, equipment reliability, technological progress, and microeconomic environments. The expertise and knowledge required for decision making at either the corporate or operational level can only be accumulated through professional training and many years of hands-on experience. Converting this expertise and knowledge into computer models constitutes a major challenge in the research and development for system automation and decision support in equipment management.

Data mining is an interdisciplinary field with the confluence of statistics, machine learning, database technology, information science, etc., and is capable of extracting non-trivial, implicit, previously unknown, and potentially useful information from large amounts of data (Frawley et al. 1992). Data mining uses sophisticated algorithms to sift through large amounts of data in order to automatically reveal patterns or trends hidden in seemingly chaotic data.

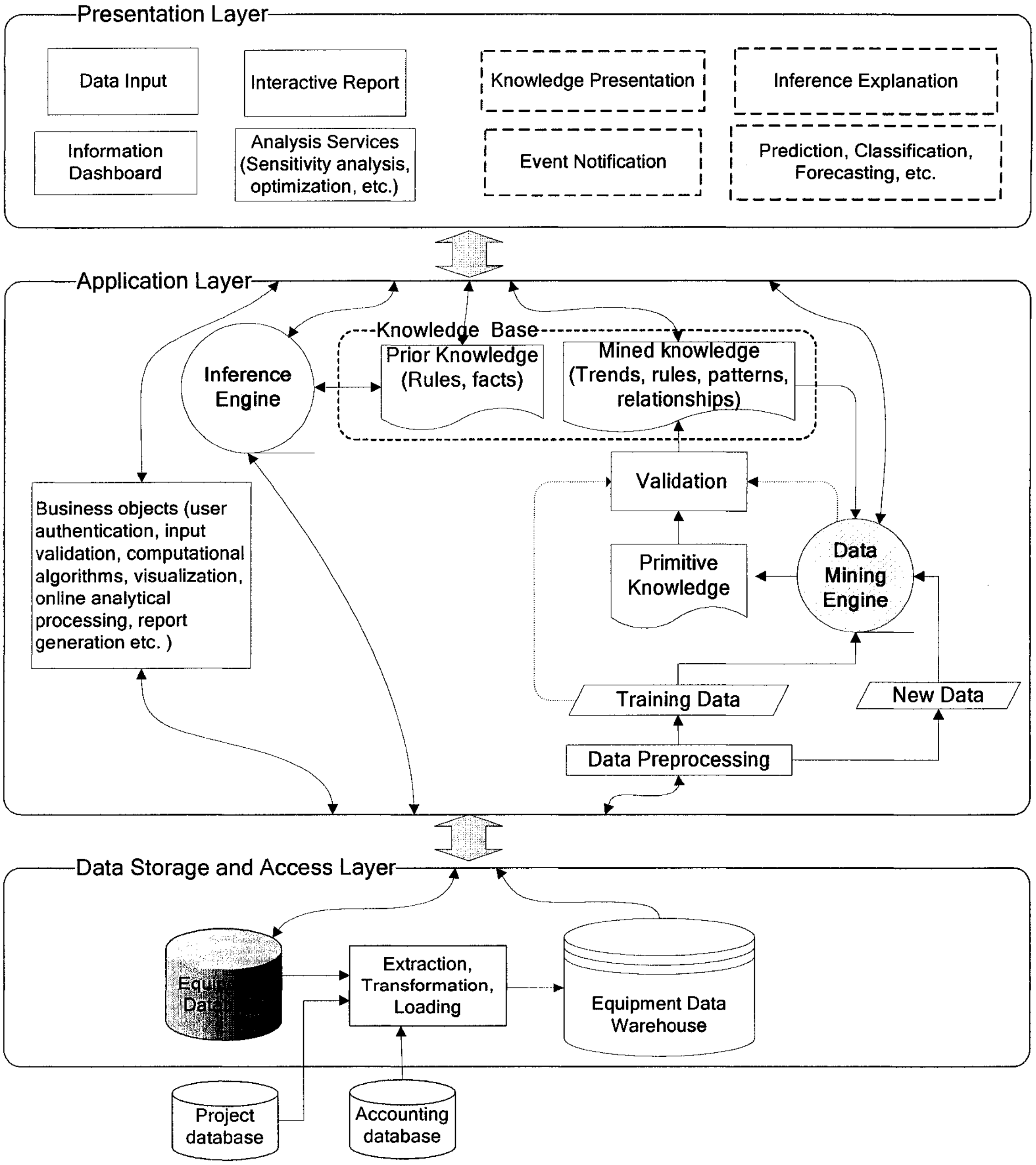
Data mining creates two broad categories of computer models to represent the new knowledge obtained: descriptive and predictive models. Descriptive models uncover patterns or correlations in the data. For example, clustering and outlier detection are two opposite data mining tasks; the former clusters data objects into groups based on their similarities while the latter identifies data objects which deviate from general patterns. Outlier detection can be used to identify pieces of equipment which call for special attention in terms of costs and benefits, or to identify the misuse/misreport of equipment on the job-site. Predictive data mining models are used for prediction based on the discovered rules or trends. For instance, in order to predict the residual value of heavy construction equipment for financial budgeting and life cycle costing analysis, a data mining algorithm is used to search out the relationship between the residual value and the set of potential influencing factors, creating a computer input/output model as a representation. Due to the versatility of data structures and algorithms in a computer model, this representation supersedes mathematical or statistical models in reflecting the inherent complex rules or trends that exist in the real world. Secondly, the integration of computer graphics in data mining allows for the data mining models to be explicitly displayed and visualized in tree structures or other graphical charts, which provides a transparent model of knowledge representation for review and analysis. Lastly, the mining model based knowledge is derived from data and can be updated in real-time, as a live connection can be established between the model and the underlying data source; thus, knowledge can be updated if such need arises. The intrinsic, dynamic, complex features of knowledge used for construction equipment management make that knowledge difficult to explore or predict using mathematical and statistical models. Data mining based modeling overcomes these difficulties and allows for the possibility of knowledge representation under a fluid environment, such as equipment management. Subsequently, the knowledge derived using data mining can be incorporated into the equipment information management system for fact-based decision support.

SYSTEM ARCHITECTURE FOR INTELLIGENT CONSTRUCTION EQUIPMENT MANAGEMENT

Intelligent applications for construction equipment management incorporate both pre-defined knowledge and derived knowledge from data into the current IT environment to automate or facilitate the decision support process. Figure 3-1 shows the three-layer architecture of the proposed intelligent equipment management system and how data mining models are added into the current knowledge base as new categories of knowledge.

Compared with the current IT architecture for decision support in equipment management, the proposed three-layer system has a unique design in terms of its support of intelligent decision-making:

- *Data storage and access layer* – All of the equipment data, whether located in an equipment database, a project database, or an accounting database, can be used for decision support because of "read-only" operations on equipment data in decision analysis. However, an equipment data warehouse is also included in the system to serve as a superior source for information retrieval and knowledge generation. In the equipment data warehouse, all of the pertinent data from disparate sources are integrated into a centralized data warehouse through an automated process of data Extraction, Transformation and Loading (ETL); data in the equipment data warehouse are repackaged in subject-oriented multidimensional data models which promote efficiency of information retrieval from data. Using equipment data warehouses as a data source allows decision-makers to discover knowledge from different perspectives at various levels of granularity under the different business subjects.


Presentation Layer
Data Input
Interactive Report
Knowledge Presentation
Inference Explanation
Information Dashboard
Analysis Services (Sensitivity analysis, optimization, etc.)
Event Notification
Prediction, Classification, Forecasting, etc.
Application Layer
Inference Engine
Knowledge Base
Prior Knowledge (Rules, facts)
Mined knowledge (Trends, rules, patterns, relationships)
Validation
Business objects (user authentication, input validation, computational algorithms, visualization, online analytical processing, report generation etc.)
Primitive Knowledge
Data Mining Engine
Training Data
New Data
Data Preprocessing
Data Storage and Access Layer
Equi Data
Extraction, Transformation, Loading
Equipment Data Warehouse
Project database
Accounting database

- *Application layer* – Both rule-based decision modules and data mining modules are incorporated into the system for decision support. The data mining engine includes various data mining algorithms for knowledge discovery and utilization. Following the iterative data mining process, the validated data mining models are created and stored in the knowledge base to represent the newly discovered knowledge.

- *Presentation layer* – The system provides a user-friendly graphical interface for interactive information retrieval, visual presentation of knowledge, explanation of inference process, event notification, predictive and explorative analysis. Most data mining algorithms are able to generate transparent, easy-to-interpret data mining models. The decision maker is confronted with newly derived knowledge in a visual format which he or she can judge according to interest-level and usefulness. The new knowledge, having been verified to be effective, is then used for classification, prediction, forecasting.

The proposed system architecture facilitates automatic knowledge discovery and utilization in addition to rule-based decision support for construction equipment management. The data mining processes run parallel to the rule-based decision modules in the application layer, but generate previously unknown and dynamic knowledge for decision support.

Although data mining fits into the overall architecture to empower the equipment management decision support system in an automatic approach, the data mining process must be planned, designed, and integrated into the system with expertise and knowledge. As shown in Figure 3-1, design of data mining modules in the intelligent system includes several major steps, including data pre-processing, generation of primitive knowledge, validation of knowledge, inclusion of mined knowledge in the knowledge base, and use of knowledge for various decision support tasks. The details of a typical data mining process, from problem formulation, data preparation, and modeling to deployment, are introduced in the industry-neural data mining standard, CRISP-DM (Chapman 2000).

INTELLIGENT FEATURES OF THE PROPOSED FRAMEWORK

The objective of the proposed framework is to mimic the decision making process of decision makers in order to automate, either fully or partially, the decisions which are otherwise performed manually with tremendous time, effort, and expertise. Incorporation of both prior knowledge from experience and derived knowledge from data enables the system to think, analyze, and act like a human decision maker or provide the decision makers with the right facts and tools for decision making. The proposed system has the following intelligent features for decision support:

Rich data sources, improved data quality and data structure – The system is backed by operational and warehouse data. The data in the equipment data warehouse provide better decision support because of improved data quality, a centralized data source, and subject-oriented data structure. In addition to the standardized reports generated from operational systems, the decision makers can interactively retrieve information from the system based on the needs required for various decision support tasks. Knowledge generated from the equipment data warehouse through the data mining process is more reliable and credible because of improved data quality and structure.

A wide spectrum of knowledge represented within the system framework – To supplement the traditional rule-based decision support modules, the system allows for knowledge to be derived from data and added to the knowledge base. Although both the prior knowledge in rule-based modules and the derived knowledge in data mining modules represent the facts and rules in the system, the two categories of knowledge differ from each other in many aspects, as described in Table 3-1. When the explicit representation of knowledge is not possible in such cases as unconfirmed knowledge, dynamic knowledge, or exploratory knowledge, data mining is a powerful tool for discovery, validation, and deployment of such knowledge in the system.

The methods for generating knowledge models are fundamentally different in rule-based decision modules and in data mining based decision modules. For example, the decision tree model and Bayesian belief network model are used by both types of modules to represent decision models with certainty and uncertainly respectively. However, in rule-based modules, the models are handcrafted meticulously by domain experts, whereas in data mining modules the models are constructed automatically by computer algorithms based on training data.

Table 3- 1. Differences between Prior Knowledge and Derived Knowledge

	Prior Knowledge	**Derived Knowledge**
Scope and source	Explicit known knowledge, based on personal and expert knowledge	Implicit knowledge, potentially existent and useful, derived from historical data
Application	Rule-based decision support modules	Data mining modules
Storage	Knowledge base and inference engine: The former stores raw facts and rules, and the latter combines the facts and rules to make inferences	Data mining models. The data mining models are generated, validated and deployed using data mining engine
Nature	Simple; relatively static	Complex; relatively dynamic
Knowledge representation	Facts and rules: Examples include crafted decision tables, decision trees, and decision graphs	Rules, trends, patterns, relationships: Examples include generated decision trees, visualized data patterns and characteristics, and summarized trends
Knowledge updating	The knowledge base and inference engine are updated manually	The data mining models are updated automatically by rescanning the dataset and revalidating
Knowledge utilization	Interpretation, diagnosis, prediction, and control	Classification, prediction, forecasting, clustering, outlier detection

Data mining can also be applied in the system for unsupervised learning of knowledge from equipment data, including the identification of interesting data clusters, anomalous data, and highly related events. Knowledge of this nature can only be obtained through exploratory analysis in data mining modules.

Enhanced user experience in decision support – The proposed framework enhances user experience in decision support with respect to interactive information retrieval, visual representation of knowledge, automated prediction, inference explanation, and event notification, as described below:

- *Interactive information retrieval* – The interactive features of Online Analytical Processing (OLAP) allow the decision maker to generate information as needed from the warehouse data in flexible ways, and to visually represent the data in pivot tables and charts for decision support.

- *Visual representation of knowledge* – Both prior knowledge and mined knowledge are represented in visual formats, such as decision trees, mathematical formulas, and graphical charts. The decision makers can not only view the knowledge in visually intuitive formats, but also manipulate the display of knowledge through visual operations.

- *Automated prediction* – Rules and trends in the knowledge base can be used in the prediction of new, unknown, or future cases. Prediction results help decision makers to take appropriate action.

- *Inference explanation* – Inference explanation allows for informed decisions by explaining the reasoning process; i.e., given the series of factual data, which rules are triggered, in what sequence, and why is the stated conclusion drawn.

- *Event notification* – Through emails, portable electronic devices, or dedicated system portals, the decision makers are notified of any important events, such as successes, failures, or milestones in important processes.

CASE STUDY: WORK ORDER EVALUATION

This application is designed to automate the evaluation of work orders in M-Track for the collaborating contractor. While minor repairs and maintenance are usually performed on the spot, the major repair, maintenance, or overhaul of heavy construction equipment will be assigned to one of the geographically distributed shops. The general procedures for work order management are listed below:

- The superintendent makes a preliminary inspection and fills out a work order on the equipment through the equipment management system; the work order includes the details of estimated labour hours, parts, and the deadlines;

- Adjustments are made to the maintenance/repair work plan based on the work order estimation, resources, and the priority level of the project;

- Maintenance/repair works are performed according to the schedule, and the actual labour hours and parts are input into the system;

- Reports are generated on estimated work order, actual incurred items, and costs.

Although the current equipment management system provides a work order entry and manipulation module, and generates multiple reports on work order estimation and implementation, the "lessons" learned from the inaccurate estimation of work orders must be explored by the equipment management team. As a matter of fact, the accuracy of work order estimates is generally not unsatisfactory, and its indirect impact on the project from the unavailability of major equipment is often linked to an inappropriate maintenance/repair schedule generated by virtue of inaccurate work order estimation.

Problem identification – The accuracy of estimated labour hours on work order items is a major concern. Many factors are cited as having a potential impact, including equipment division, department, category, class, make, year of manufacturing, component, and parts. In order to improve the accuracy of the work order estimation, the following questions should be investigated: which groups of equipment or equipment show a high accuracy of estimation? Which groups of equipment have underestimated labour hours, and by how much? What are the leading features of impact? These questions can only be answered after a comparative analysis is conducted on both the estimated and actual work orders. And furthermore, the answers to these questions are changing dynamically with time.

Data preparation – The objective of work order evaluation is to predict the accuracy of estimated labour hours for the work order item based on the known factors of impact. Twelve attributes were identified as potential factors of impact: division, department, category, class, unit, manufacturer, equipment age, component, parts, repair type, estimated labour hours, and estimated parts. A total of 952 cases were collected from the equipment database, after removing some obvious outliers caused by data entry error.

Modeling – AutoRegressive Tree (ART) is a non-linear regression tree algorithm proposed by Meek et al. (2002). By learning from the historical cases, the algorithm builds up a top-down decision tree structure with a multiple linear regression model grown at each leaf node. Using the training data, the ART algorithm partitions the data space into sub-regions where linear regression models exist for subsets of data. When the model is used for prediction on a new case, a decision path is found by answering a series of questions along the path from the root node of the ART Model, down through a number of decision nodes, and terminating at one of the leaf nodes where the regression model is used for predicting the target variable using the set of known predictor variables. Based on the prepared dataset, the ART model for work order evaluation is trained by using 80% of the cases randomly selected from the set, with the remaining 20% held back for validation tests. Figure 3-2 shows partially the derived ART model for evaluation of work orders.

Model validation – The predicted deviations and actual deviations were tabulated for comparison and it was found that 87% of the predicted values fell within the acceptable range according to expert opinion. The scatter plot is also used for visually assessing the accuracy of the prediction. As shown in Figure 3-3, predicted deviations of the labour hours are plotted against the actual deviations, indicating that only a small percentage of cases are incorrectly predicted.

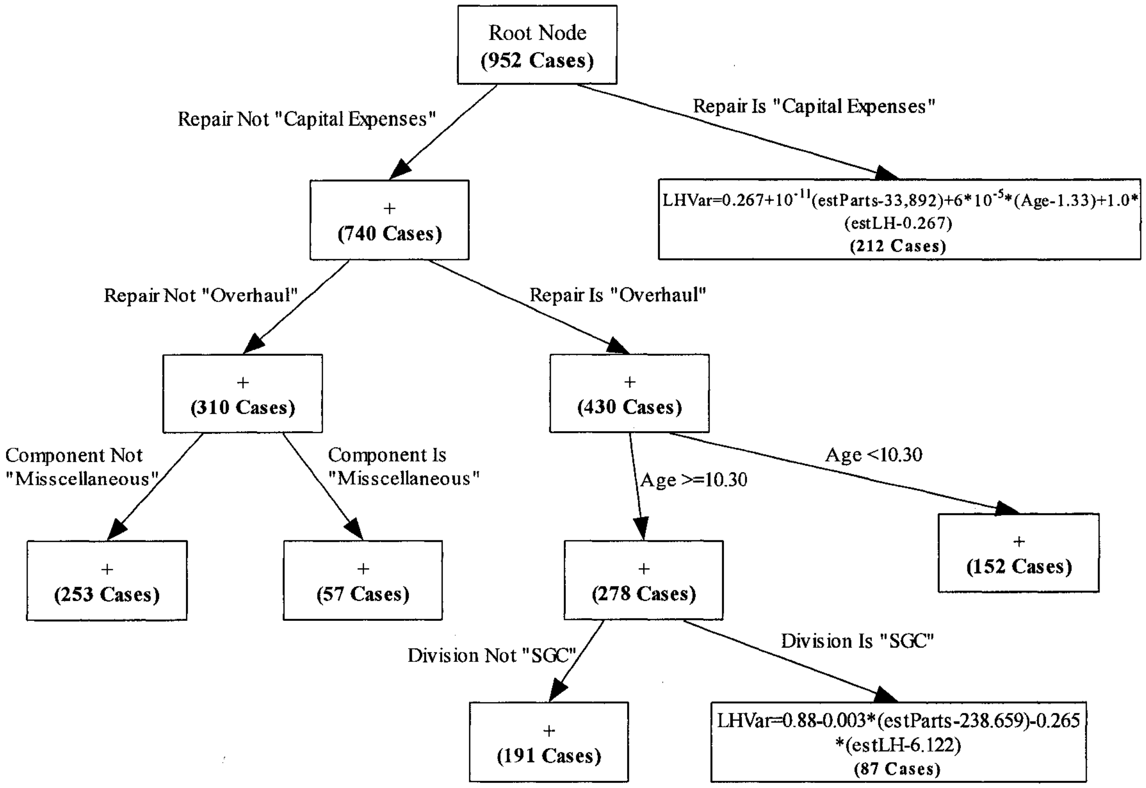


Figure 3- 2. A Portion of the Induced Work Order Evaluation ART Model

(Decision nodes are denoted with an expansible plus sign and leaf nodes with a regression formula)

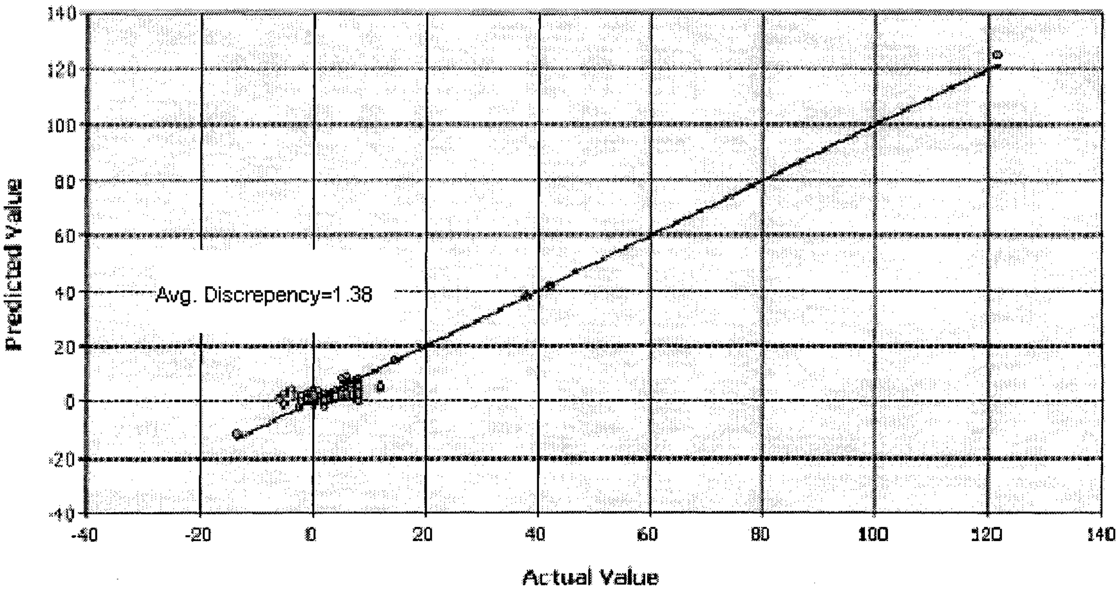
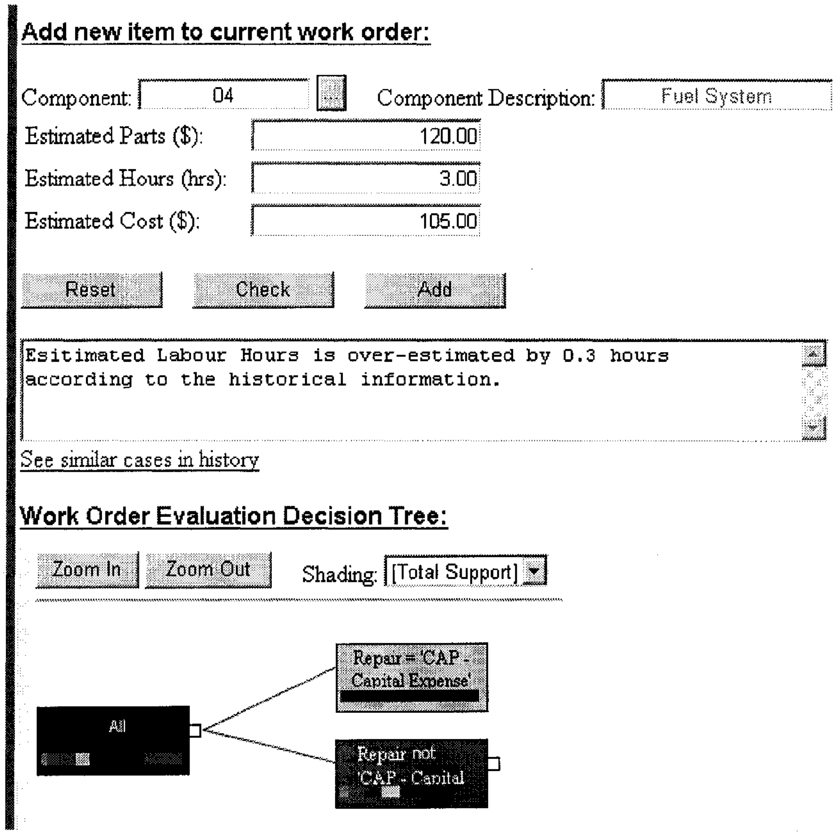


Figure 3- 3. Scatter Plot of Actual Deviations vs. Predicted Deviations Using the Test Dataset

Model deployment – The validated work order of the ART model is hosted using the Analysis Services of Microsoft SQL Server 2005, with a live connection to the equipment database and the current equipment information management system. As shown in Figure 3-4, after the user enters the estimated labour hours as a work order item, the model is capable of validating the input and making recommendations based on historical information. If the user raises any doubts about the "recommendation" from the ART model, he or she can visually analyze the decision tree structure and track those cases which follow the same decision path. The data-mining model is also updated automatically on a regular basis to reflect recent changes in the data depository.



Add new item to current work order:

Component: 04 Component Description: Fuel System

Estimated Parts ($): 120.00

Estimated Hours (hrs): 3.00

Estimated Cost ($): 105.00

Esitimated Labour Hours is over-estimated by 0.3 hours according to the historical information.

See similar cases in history

Work Order Evaluation Decision Tree:

Shading: [Total Support]

Figure 3- 4. Recommendation on Labour Hour Estimate from Embedded ART Model

CONCLUSIONS

This chapter proposes an intelligent system for construction equipment management with the incorporation of data mining techniques. As a supplement to the current rule-based equipment management decision support system, data mining enhances its decision support capability by incorporating previously tacit and dynamic knowledge to the knowledge base. Data mining algorithms sift through data to find useful patterns or rules and to present as data mining models for decision support. A large number of data mining models, such as the regression tree, appear as transparent "white box" models that are easy to visualize and interpret. Similar to the rule-based system, the data mining models, when used for prediction, are capable of explaining the reasoning process to decision makers. Considering the fluid dynamic nature of construction equipment management, the data mining driven decision support system shows advantages which otherwise cannot be achieved using a rule-based or statistics-based decision support approach.

REFERENCES

Berzonsky, B.E. (1990). “A knowledge-based electrical diagnostic system for mining machine maintenance.” *IEEE Transactions on Industry Applications*, 26(2).

Caterpillar (2006). “AccuGrade® grade control system.” Caterpillar Inc., <http://www.cat.com/cda/layout? m=37483&x=7> (January 8, 2007).

Caterpillar (2006a). “Vital information management system.” Caterpillar Inc., <http://www.cat.com/cda/ layout?m=37498&x=7> (January 8, 2007).

Chapman, P.,Clinton, J., Kerber, R.,Khabaza, C. and Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.* The CRISP-DM consortium.

Frawley, W., Piatetsky-Shapiro, G., and Matheus C. (1992). "Knowledge discovery in databases: an overview". *AI Magazine*, Fall, 1992, 213-228.

Fu, C., Ye, L., Liu, Y., Yu, R. Iung, B. Cheng, Y. and Zeng, Y. (2004). “Predictive Maintenance in Intelligent-Control-Maintenance-Management System for Hydroelectric Generating Unit.” *IEEE Transactions on Energy Conversion*, 19(1).

Gillerspie, J.S. and Hyde A. S. (2004). *The replacement/repair decision for heavy equipment*, Report VTRC 05-R8, Virginia Transportation Research Council, November 2004.

Hopgood, A. (2001). *Intelligent systems for engineers and scientists*, CRC Press, Boca Raton, FL.

Lucko, G., Anderson-Cook, C.M. and Vorster M. C. (2006). “Statistical considerations for predicting residual value of heavy equipment.” *J. Constr. Eng. Manage,* ASCE, 132(7), 723-732.

Meek, C., Chickering, D.M. and Heckerman, D. (2002). “Autoregressive tree models for time-series analysis.” *Proc., 2nd SIAM International Conference on Data Mining,* Arlington, VA.

NSERC/Alberta Construction Industry Research Chair (2007). “M-Track.” NSERC/Alberta Construction Industry Research Chair, <http://www.construction.ualberta.ca/forum.shtml> (May 7, 2007).

QUALCOMM (2006). “GlobalTRACS® equipment management system.” QUALCOMM Inc., <http://www.qualcomm.com/qwbs/solutions/prodserv/globtracs.shtml> (January 8, 2007).

Vorster, M. C. and Livermore, M. E. (1994). "Executive development for equipment managers." *Proc., Equipment Resource Management into the 21st Century,* ASCE, Nashville, TN., 87-95.

Yam, R.C.M., Tsei, P. W., Li, L. and Tu, P. (2001). "Intelligent Predictive Decision Support System for Condition-Based Maintenance." Int. J. Adv. Manuf. Technol, Springer-Verlag London Ltd., 17, 383-391.

Zhang, J., Tu, Y. and Yeung, E.H.H. (1997) "Intelligent Decision Support System for Equipment Diagnosis and Maintenance Management." *Proc., Innovation in Technology Management – The Key to Global Leadership, International Conference on Management and Technology*, Portland, OR., 733.

CHAPTER 4. DATA WAREHOUSING FOR CONSTRUCTION EQUIPMENT MANAGEMENT[2]

INTRODUCTION

In recent years, most large contractors have increased their investment in maintaining, updating, and replacing their equipment fleet to satisfy the needs of project construction (Stewart 2000). Equipment managers in charge of a large fleet need both timely information and tools to make strategic decisions pertaining to resource allocation and equipment maintenance, repair, and replacement. The current trend in the industry is for contractors to redirect the responsibilities of routine equipment maintenance and repair from the equipment manager of the company to project managers (Stewart 2004). The project managers rely on a computerized equipment management system to automate the process of making records on daily equipment operations, preventative maintenance, and repair. However, this does not provide a substantial benefit to the equipment manager: although most current equipment management systems can provide canned well-formatted reports, analysis of the equipment data to make decisions turns out to be a nontrivial process.

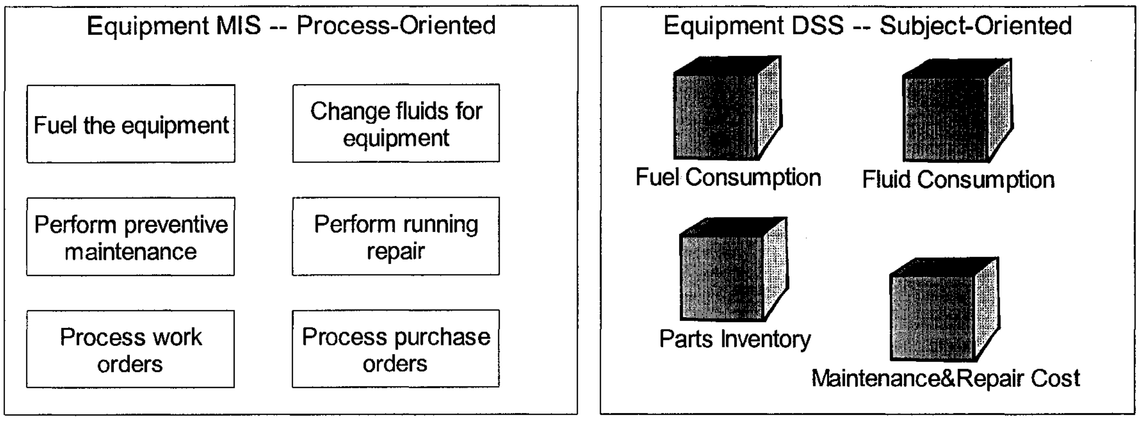
[2] This chapter is a journal paper co-authored with Kim, H. (Department of Civil Engineering, Yonsei University, S. Korea), and Zaiane, O. (Department of Computing Science, University of Alberta), and published in Canadian Journal of Civil Engineering, 33(12), December 2006. The writer of this thesis is the primary contributor and writer of this journal paper, the other authors provided supervisory works.

One of the partners in this research is a major construction contractor in Alberta, Canada. Its current equipment Information Management System (IMS), was developed in 1997 in collaboration with the Natural Sciences and Engineering Research Council of Canada /Alberta Construction Industry Research Chair, partly because of the increase of the contractor's business in road construction and equipment rental. Application of the equipment IMS across the company successfully replaced the traditional, error-prone, paper-based bookkeeping and reporting chores needed to manage the construction equipment. By capturing management data on the daily operations of the more than 3000 pieces of equipment in the fleet, the equipment IMS maintains a parts, fluids, and fuel inventory service and also stores a historical record of all servicing performed on each piece of equipment. The equipment IMS is also capable of producing standardized work orders and 48 types of reports on equipment usage, maintenance, and repair. Although the equipment IMS tracks daily operations successfully, the equipment manager found that the large amounts of data accumulated over the years made it inefficient to answer simple questions, such as comparing one make of earth-moving equipment with another in terms of total maintenance and repair costs over the last 5 years. The 48 reports in the equipment IMS do provide some degree of flexibility by allowing the user to change a few parameters such as cost account and time, but the report functions are restricted to a limited number of query parameters and customized formats. Furthermore, some reports are not easy to read, become they can run up to several hundred pages in length. Given limited number of queries and the fact that the equipment manager must look over the data from a variety of perspectives; these canned, customized reports cannot answer all the constantly changing questions required for decision-making.

The inability to provide efficient decision support is a well-recognized problem with transaction-processing systems in computer science. Based on a relational database model, a transactional system, such as an equipment IMS, is designed for the efficient capture of operational data. The relational database model is built upon various business processes for daily operations. The Equipment IMS mimics purchase order processing, preventive maintenance, fueling, etc., for the collaborating contractor's equipment fleet. The process-oriented equipment IMS guarantees that these data are added and updated efficiently during daily operations; however, sophisticated data analyses are not performed well. The management operations of this process-oriented equipment IMS are structured and repetitive, with isolated transactions. The operations require detailed and up-to-date data, and normally the IMS automates the clerical data processing tasks. For the sake of decision support, however, queries become very complex and require data consolidated from different data sources. Extracting data from a transactional database requires that complex queries be built across different business objects, which can only be accomplished by database specialists. Moreover, extracting data would add considerable unnecessary over-load to operational databases. Using an operational system for decision support becomes even more inefficient with today's increasing volume and complexity of data. For decision support, when real-time ad-hoc complex queries are required for interactive analysis, a multidimensional structure that contains historical data, aggregates relevant information from various sources, and is maintained separately from operational databases is needed. This structure is known as a data warehouse. It allows data to be summarized and analytically processed.

Codd et al. (1993) first tackled this data analysis problem by introducing online analytic processing (OLAP) for decision support. Online analytical processing repackages the data from online transactional processing systems (OLTP) into a data warehouse and presents the data in a multidimensional structure. In addition to its multidimensional nature, a distinct feature of the data warehouse that facilitates dynamic decision support is subject orientation. Subject orientation means that the data model and presentation are centered around individual subjects, such as fuel consumption, parts inventory, and maintenance and repair cost. The fundamental difference between OLTP-based IMS and an OLAP-based decision support system (DSS) is that the former breaks the system down into managerial functions or business processes, whereas the latter breaks the system down into subjects of interest. Figure 4-1 compares the process orientation of an equipment IMS with the subject orientation of a DSS. In an equipment DSS, data from different OLTP systems are integrated into a single repository and re-structured for OLAP. The multidimensional structure of a data warehouse enables data analysis to be performed along any combination of descriptive attributes and at various granularities (levels of detail) for each subject. In brief, the warehouse data comes from the original transactional databases but are cleaned up, integrated, and optimized for analyses and reports. The differences between OLAP and OLTP have been described in many references (Codd et al. 1993; Bain et al. 2001).

An equipment DSS using data warehousing technology can provide high-level decision support for equipment managers. In this study, we developed an equipment data warehouse using the historical and current equipment databases of the collaborating contractor. We also designed a Web-based application to expose the data warehouse to an interactive Web site, which would permit remote access and online access. Use of this prototype shows that this equipment DSS provides a flexible and powerful environment for equipment managers to analyze equipment data with clear advantages over the current equipment IMS. This chapter summarizes the methodology and findings and the challenges encountered in designing and implementing the equipment DSS, with focuses on the modeling and design of the equipment data warehouse.



Note: only a portion of each system is shown for illustration purpose.

Figure 4- 1. Comparison between Process-oriented View of Equipment IMS and Subject-oriented View of Equipment DSS

The next section is a literature review of related research. This is followed by a description of the data architecture for the proposed equipment data warehouse. The chapter then explains the building of the equipment data warehouse and the implementation of the warehouse in a Web-based equipment DSS. The benefits of the data warehousing in equipment management are reiterated, and the chapter concludes with a summary.

LITERATURE REVIEW

Chau et al. (2002) conducted a research project and an investigation into the application of data warehousing technology in construction. The authors built a DSS based on OLAP to manage the inventory of construction materials. After successful application in a residential building project, the authors concluded that data warehousing technology could produce more intuitive, multi-view information from the data depository than the traditional OLTP ad hoc reports provided. In another case, Ma et al. (2005) applied the data warehousing technique to exchange electronic documents between project participants. In their research project, useful information are extracted from electronic documents and loaded into a data warehouse for an in-depth data analysis by the contractor, the owner, and the resident engineer. The authors concluded that the data warehousing enabled all parties to identify critical issues related to their problems.

Microsoft Corporation integrated data warehousing technology into the Portfolio Analyzer of its Microsoft® Office Project Server, which allows the project participants to analyze the project resources and performance data in the form of multidimensional OLAP cubes. Portfolio Analyzer demonstrates that for complex data structures in project management a multidimensional view of the project data provides superior performance over traditional methods of data analysis and information delivery.

Enterprise resource planning (ERP) is an information technology that integrates a firm's applications and functions into a single computer system with a shared database that is accessible across the enterprise. Enterprise resource planning systems primarily use a relational database. Enterprise recourse planning vendors started merging data warehouse and OLAP technology into their product only in recent years. For example, SAP's general platform NetWeaver® (SAP 2006) can be used in the engineering, construction and operations industry. However, there are few reported cases of ERP systems being implemented by construction contractors; the reasons for reluctance to use ERP systems include the high costs of planning, design, implementation, and training, as well as the delay in return on investment (Shi and Halpin 2003; Ahmed et al. 2003).

ARCHITECTURAL DESIGN OF EQUIPMENT DATA WAREHOUSE

The data warehouse provides a consolidated view of enterprise data, allowing users to answer various business questions by browsing through the data from different perspectives. To this end, planning and design of the data warehouse need to address two important issues: (i) identification of all the relevant subjects from the business processes; and (ii) for each subject, identification of a set of measurable facts (numerical measures) and related dimensions (textual entities to describe facts). The dimensions are not designed independently for each subject but are usually shared across subjects. For instances, as shown in Figure 4-2, the dimensions “time”, “equipment group”, and “equipment owner” can all be used for the subjects “fuel consumption”, “purchase order”, and “maintenance and repair cost”. The time dimension is used for all subjects because of the need to evaluate the changes in measures over time.

Kimball and Ross (2002) proposed data warehouse bus (DWB) architecture for the design of a consistent data warehouse: a bus matrix depicting the entire data warehouse is used to identify subjects for operational processes within the company and to obtain a master suite of standardized (conformed) dimensions and facts that are uniformly interpreted across the enterprise. Kimball and Rose defined the conformed dimensions as either identical to or strict mathematical subsets of the most granular (detailed) dimension. We used this approach to design the bus matrix for the equipment data warehouse shown in Figure 4-2. The matrix rows show the business processes in equipment management, of which the identified subject represents a data mart or a data cube. Note that these data marts are at the basic level. Data marts at a higher level, if required, can be created by combining the facts from different processes without much effort. Currently, we include 10 basic-level data marts in the warehouse. Each data mart contains the numerical measurements used as equipment management performance indicators. The matrix columns in Figure 4-2 shows the conformed dimensions that are shared across various equipment management processes. Obviously, the dimensions shared by of the most data marts should have highest priority in the design and deserve the most attention.

The identification and design of dimensions are critical to determining what kinds of questions can be asked regarding each subject. All the appropriate dimensions should be identified for each data mart before participating in the bus matrix. In most cases, each dimension can be naturally organized in one or more hierarchies, from a higher aggregated level to a lower detailed level, in parent-child relationship: an example is equipment category→equipment class→unit in the equipment group dimension. Therefore, the dimension can be represented as a comprehensive entity organizing similar textual descriptions into hierarchical levels; members at each level can also have their own properties. Depending on the problem domain, and how frequently dimensions are shared and used, selection of dimensions may vary across different data warehousing applications. For example, in the equipment data warehouse, the hierarchical structure of the equipment group dimension consists of category, class, and unit in an increasing order of granularity (detail). Although "manufacturer" is a descriptive attribute of the equipment, we chose to build a separate manufacturer dimension rather than use manufacturer as an attribute of equipment dimension, because the manufacturer is commonly shared in the system.

	Common dimensions												
Business processes	Time	Equipment Group	Equipment Owner	Account	Supplier	Fuel Type	Fluid Type	Parts	Employee	Manufacturer	Component	Cost Item	Cost Category
Fuel Consumption	X	X	X	X	X	X				X			
Fuel Inventory	X		X	X	X	X							
Fluid Consumption	X	X	X	X	X		X			X			
Fluid Inventory	X		X	X	X		X						
Parts Consumption	X	X	X	X	X			X		X			
Parts Inventory	X		X	X	X			X					
Purchase Order	X	X	X	X	X			X	X		X		
Work Order	X	X	X							X	X	X	
M&R cost	X	X	X							X			X
Human Resource	X		X	X					X				

Figure 4- 2. Bus Matrix for Equipment Data Warehouse

There are many advantages to modeling the equipment data warehouse using DWB architecture. First, all the processes with measures of interest are identified. The processes and data marts are clearly identified in the rows of the bus matrix. Second, all the common dimensions are identified. These dimensions are standardized and shared among various processes, so the dimension design will give due consideration to every process involved. If well designed, the shared dimensions will guarantee the consistent structure and content of a dimension and make it readily usable for different data marts. Third, the conformed dimensions also facilitate data staging by avoiding the repeated pulling of data from the same source. As a result, the data extraction, transformation, and loading (ETL) efforts can be minimized. Finally, a bus matrix can serve as a management and communication tool for a data warehouse (Kimball and Ross 2002).

MULTIDIMENSIONAL MODELING

In the DWB matrix illustrated in Figure 4-2, each row contains a set of measures for the corresponding business process and the associated common dimensions. This data structure can be best represented by a star schema with a fact table at the center and all related dimensions arranged around it. In a data warehouse, the star schema models each subject as a multidimensional data cube, with all the numerical measurements in the central fact table and all the descriptive entities in the surrounding dimension tables. The star-shaped data structure makes it possible to analyze data in the fact table along one or any combination of descriptive dimensions at various granularities. Questions of when, where, who, which, and so on can be answered after the schema is transformed into a multidimensional data cube. Proper modeling of each data cube, with its underlying fact table and dimension tables, enables comprehensive data analysis of an individual subject. All the cubes in the system collectively provide an integrated view of equipment management performance.

In the equipment data warehouse, 10 data cubes were designed on the basis of DWB architecture to model the different facets of equipment management. Figure 4-3 shows the star schema for the data cube for maintenance and repair cost, one of the most important subjects in the system. The schema includes one fact table and six dimension tables. The fact table for maintenance and repair cost includes measures of the number of hours spent on maintenance and repair, the labor cost in dollars, the parts cost in dollars, and the total maintenance and repair cost in dollars. The six dimension tables are time, equipment group, equipment owner, cost category, account, and manufacturer dimensions:

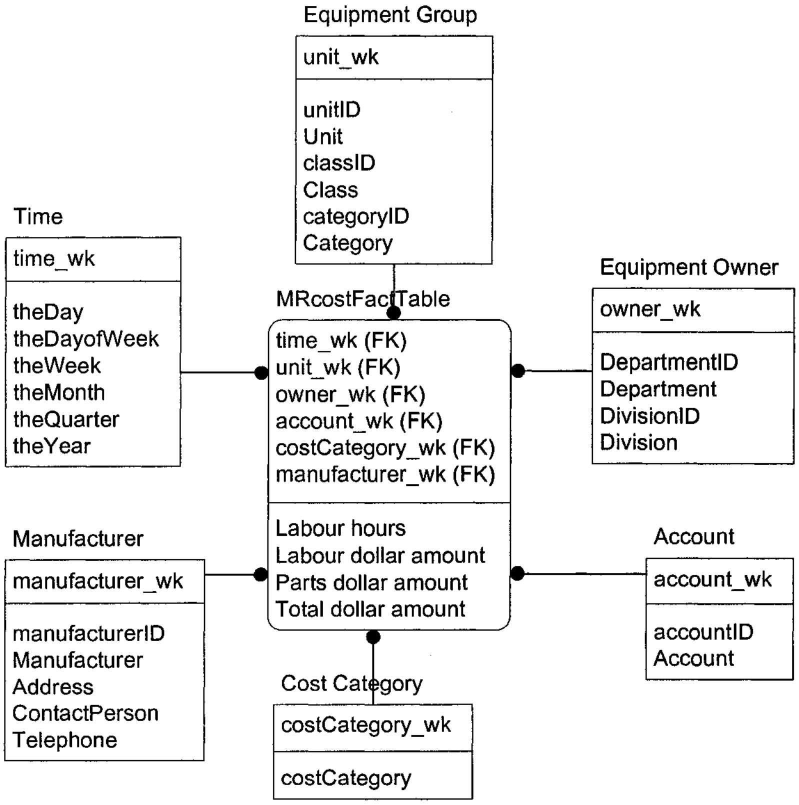


Figure 4- 3. Dimensional Model for Maintenance and Repair Cost Data Cube

- The time dimension has two hierarchies: year→quarter→month→day; and year→week→day.

- The equipment group dimension has one hierarchy: equipment category → equipment class →individual unit.

- The equipment owner dimension has one hierarchy: organization owning the equipment → regional division → department.

- The cost category dimension has three categories for cost occurrence: preventive maintenance, running repair, and work order.

- The account dimension contains the financial account descriptions of cost items.

The manufacturer dimension contains information about the manufacturer of the equipment, including address, contact person, and phone number. With a multidimensional data cube, all the reports on maintenance and repair costs contained in the previous equipment IMS are now integrated into a single destination. Each data cube can be browsed visually in the data warehouse system. Some important data manipulation and data view functions include: drill down, roll up, slice and dice, and pivot, as described below.

- *Dill down* – The drill down operation allows data at a more detailed level to be accessed from a more general category along a dimension. For example, a user could obtain monthly equipment performance data by drilling down through quarterly data .

- *Roll up* – The roll up operation is the opposite of the drill down. These data are manipulated for presentation on a higher level with summarized information along a dimension. For example, monthly data are summarized to get quarterly reports. In the data warehouse system, the numerical facts in data cubes are usually pre-aggregated along each dimension at each level to improve query performance.

- *Slice and dice* – The slice function projects the data on one dimension, whereas the dice function projects the data in two or more dimensions;

- *Pivot* – The pivot function rotates the data cube axes to change the user's view of the data.

With the visual tools in the data warehouse system, the user can browse, drill down, roll up, and slice and dice the cost data in the maintenance and repair cost cube along or across any combination of dimensions, at different levels of detail. Figure 4-4 shows some examples of OLAP operations on a portion of a maintenance and repair cost data cube containing cost facts along with the dimensions of equipment owner, equipment group, and time (minor changes were made on the dimension members for illustration purpose).

The visual browsing tools are further supplemented by a multidimensional query language that allows the user to get answers to all equipment management questions. The following are examples of multidimensional queries: (i) Show the repair costs of all earthmoving equipment for 2001. (ii) Compare the repair costs of the same class of equipment with those of different models and manufacturers for a particular period. (iii) Show the equipment with the top-*n* repair costs in a particular equipment class.

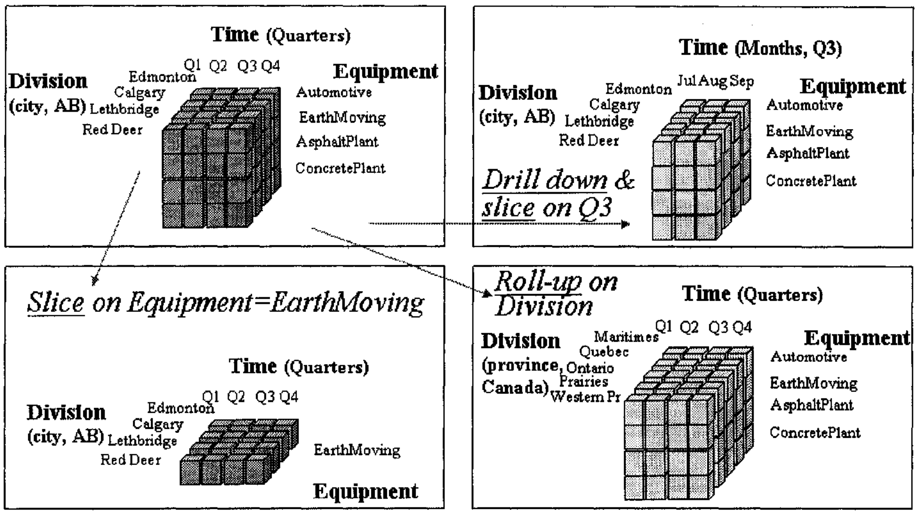


Figure 4- 4. Examples of OLAP Operations on the Maintenance and Repair Cost Data Cube (drill down and slice, slice, roll up).

We encountered several issues that should be addressed when designing an equipment data warehousing system, to guarantee usefulness, efficiency, and consistency.

- *Levels of detail in modeling* – The dimensional tables should focus primarily on the most granular level of daily operations. Design of the fact tables should also follow this principle, but not strictly, depending on practical needs and on storage capacity and performance considerations. The data warehouse differs from a database in that it delivers summarized information rather than specific transaction details. The data at the atomic level allows for a higher level view through the roll-up operation; while data modeling at a higher level makes it impossible to drill down to get detailed performance data.

- *Measures in the fact table* – All potential measures of interest related to a given subject should be included in its fact table. The data warehouse is designed to be used by different management personnel in the company: the more extensive the measures, the more the diverse needs of users can be satisfied.

- *Rules of aggregation* – Some semi-additive measures must be identified. For example, the parts inventory volume in the parts inventory fact table cannot be summed along the time dimension. The cost variation percentage in the work order fact table cannot be added up along any dimension. Therefore, in these two types of processes, we shall choose *average* rather than *addition* for roll up operations.

EQUIPMENT DATA WAREHOUSING

Data warehousing refers to all the processes needed to build up and implement the data warehouse. The procedures for data warehousing include (i) identifying data sources, such as the operational systems, applications, or flat files; (ii) staging the data, which usually involves data extraction, transformation, and loading (ETL) from heterogeneous sources to a consolidated data warehouse; and (iii) using data access tools to present the multidimensional data for reports, interactive analysis, knowledge discovery, etc. Figure 4-5 shows the equipment data warehousing processes implemented in this research using the Analysis Services (a component of Microsoft® SQL Server 2000) and a web client tool designed for user interaction with the underlying equipment data warehouse.

Data Sources

One advantage of using a data warehouse for decision support is that all the data distributed in different systems across the company can be integrated into a single destination. Therefore, all of the data sources containing useful equipment management information needs be identified or the data in the data warehouse will be incomplete and not suitable for strategic decision-making. The data sources in most cases are heterogeneous, which means that they reside in operational systems (such as in an equipment IMS), applications (such as spreadsheets), and flat text files. The data warehouse system can extract data from these disparate sources and copy them to the data warehouse after preprocessing.

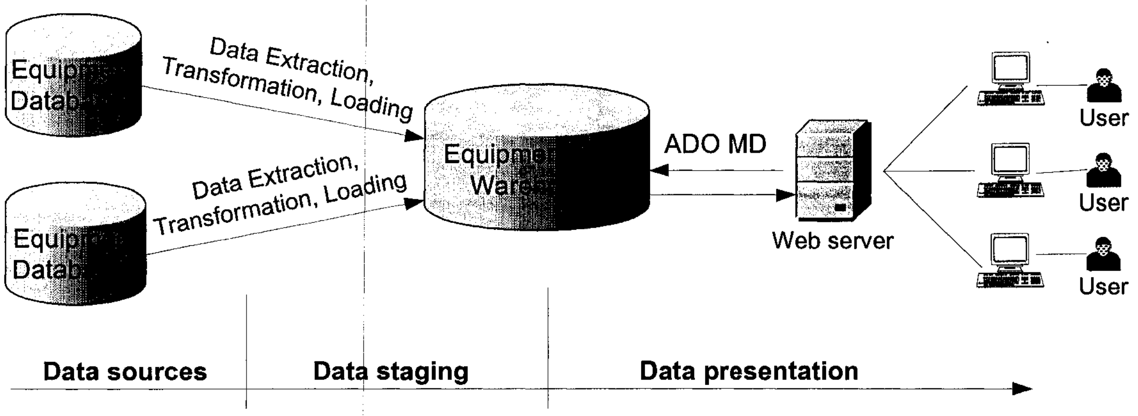


Figure 4- 5. Equipment Data Warehousing and Data Access

The equipment data warehouse in this research has two data sources: (i) the current Microsoft® SQL Server database; and (ii)the historical Microsoft® Access™ database used in the first release of the equipment IMS, which contains equipment data from 1997 to 2001. The data accumulated over these years of operation are considered valuable information for the contractor and thus are included in the equipment data warehouse. The current equipment database in SQL Server is used across different divisions located in different areas of Alberta. The data collected from different sources can be replicated and synchronized inside the SQL Server database management system. As a result, only one instance of the SQL server database needs to be used as the current active data source for the equipment data warehouse. The two sources of the data for the warehouse (the historical Access™ database and the current SQL Server database) have different data structures because of significant changes in the database model of the SQL Server version of the contractor's equipment IMS.

Data Staging

Data staging mainly comprises three sequential steps: extraction, transformation, and loading from data sources to the data staging area. Kimball and Ross (2002) defined the data staging area as everything between the operational data sources and the data presentation area. The three tasks of data ETL are completed sequentially to transport data from disparate sources to the data warehouse. Data extraction copies data from different sources to a temporary staging database in the SQL Server to minimize interference with the operational system. Data transformation is then responsible for preprocessing the data before loading to the warehouse. This step is regarded as being critical to data quality. Potential flaws, such as out-of-range, erroneous data, duplicate records, and null values, as well as in consistencies in data format and structure, are resolved in this step. Data transformation is known to be a tedious process, as some deficiencies in the source data are difficult to detect because of the large amounts of data involved. Finally, the transformed data are loaded into the database for presentation. Humphries et al. (1998) indicate that data ETL can consume 60% - 80% of the time and effort of the team in data warehousing.

Data ETL is not a one-time process: data in the warehouse need to be synchronized periodically with the updated sources. Therefore, data ETL is usually designed with off-the-shelf software as an application package scheduled to run periodically. For example, we use the Data Transformation Services (DTS) in Microsoft® SQL Server 2000 to implement the data staging process for the equipment data warehouse. Data source definition, data validation, and data extraction and loading, as well as notification of execution results, can be visually designed as a flow of processes in an application package using the complete suite of visual design tools in DTS. Scheduled execution of the package enables periodic updating of the data in the data warehouse.

Data Presentation

The data presentation area is where the data is organized, stored, and made available for direct query by users and report writers, as well as by analytical applications (Kimball and Ross 2002). After the data warehouse has been designed, built, and processed, the multidimensional data cubes are stored in an OLAP Server for utilization.

To expose the data accessible to end users, one of the most commonly used approaches is to use OLAP visual browsing tools for interactive data analysis. In this research, we chose to build the visual browsing components into an interactive Web site and host it on a centralized Web server for user access to avoid repetitive installation of OLAP client tools on desktop computers. Warehouse data presentation and access are illustrated in Figure 4-5. The next section discusses the design of the web-based DSS.

EQUIPMENT DECISION SUPPORT SYSTEM DESIGN

In recent years, many vendors of construction equipment management software have migrated their systems from desktop solutions to web-based management systems (Computer Guidance Corporation 2005; Caterpillar Inc. 2006) because of the highly distributed nature of construction equipment management operations. The Web-based equipment management DSS proposed in this research aims to provide a mechanism for users to browse and communicate with the equipment data warehouse through a Web site.

Designed as a Web-based data warehouse system (abbreviated webhouse by Kimball and Merz (2000)), the equipment DSS allows authorized users to remotely access the equipment warehouse with an Internet browser and perform data analysis online. The equipment Web site is developed and works as a bridge between the front-end user and the equipment data warehouse. Figure 4-6 shows the system architecture, which comprises three logical components.

- *Database server–* In this research project, the Pivot Table Services in Analysis Services of Microsoft® SQL Server 2000 provides a set of client tools for retrieving multidimensional data from the warehouse. The Microsoft® ActiveX® Data Objects (Multidimensional) is a data access adapter that enables communication between the data warehouse and external applications.

- *Web server* – The Web application is hosted in the Internet Information Services Web server as an interactive web site so that users can perform online data analysis after logging in.

- *Web browser and Internet connection* – The Web browser is the only client tool needed for communicating with the equipment data warehouse.

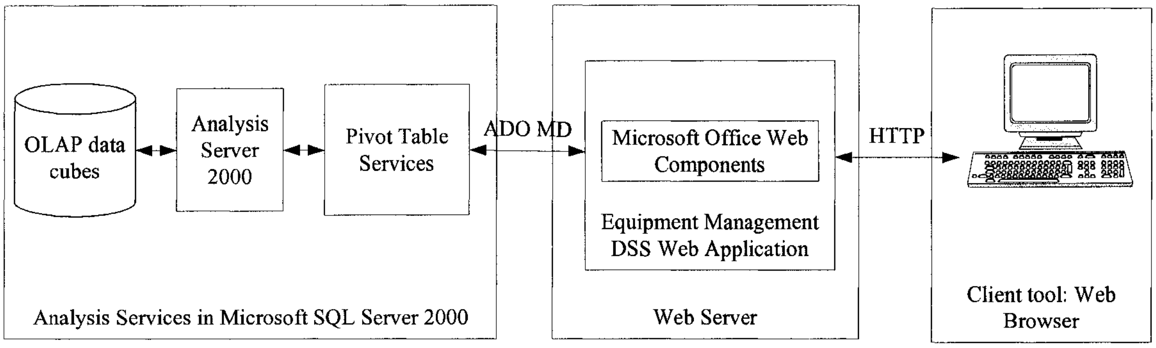
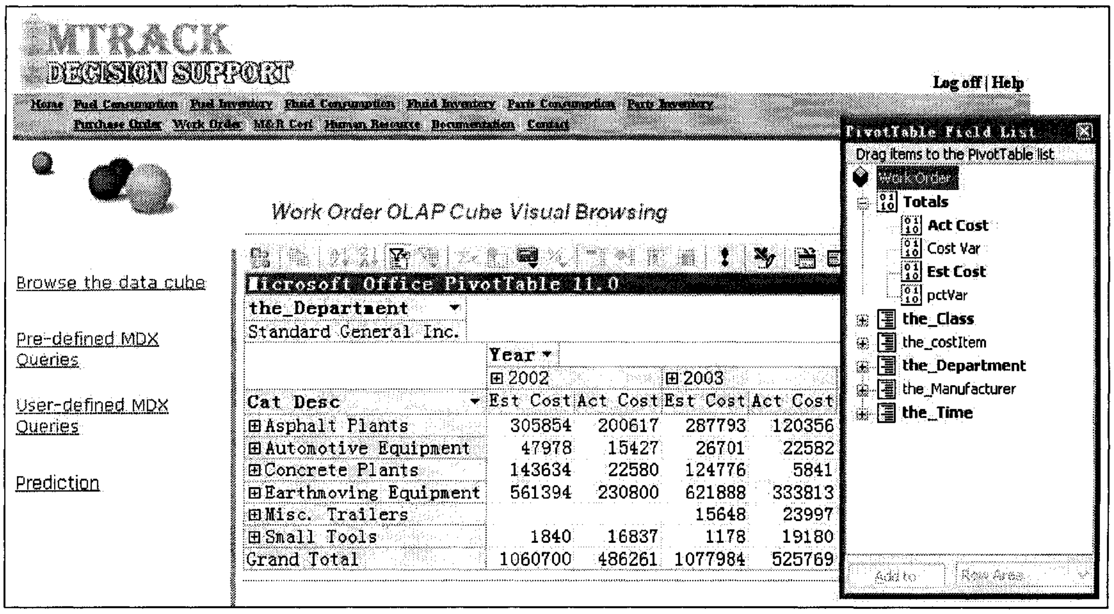

OLAP data cubes
Analysis Server 2000
Pivot Table Services
ADO MD
Microsoft Office Web Components
Equipment Management DSS Web Application
HTTP
Analysis Services in Microsoft SQL Server 2000
Web Server
Client tool: Web Browser

Figure 4-6. Equipment Data Warehouse System Architecture

The equipment DSS Web site is an interface that enables users to send in requests and get results from the data warehouse. Therefore, the Web site design should provide a visually intuitive environment which allows users to perform data manipulation and to view data cubes with minimum assistance. In this research, Microsoft® Office Web Components (OWC), including Pivot Table and Pivot Chart, were embedded in the web application to serve as the client-side visual interface for multidimensional data. Pivot Table contains user interfaces for visual browsing of data cubes, and Pivot Chart displays analysis results in a graphical format.

Figures 4-7 and 4-8 are snapshots from the equipment DSS Web site. Ten multidimensional data cubes are available in the system. Each data cube has three modes for data analysis: cube browsing, preformulated queries, and user-defined queries. Figure 4-7 shows a screen that allows the visual browsing of data cubes through roll up, drill down, and slicing and dicing. All these techniques can be accomplished with mouse clicking and the drag-and-drop technique. The measures and dimensions can be expanded from the field list to show members at different levels. Dragging and dropping selected items into different data view windows will slice and dice a data cube. A cross-tabulated table is generated if multiple items are selected for row or column headers in the data view window. Some queries involving special requests, such as top-*n*, sorting, and complex filtering of a data cube, rely on multidimensional query language designed for querying multidimensional databases.



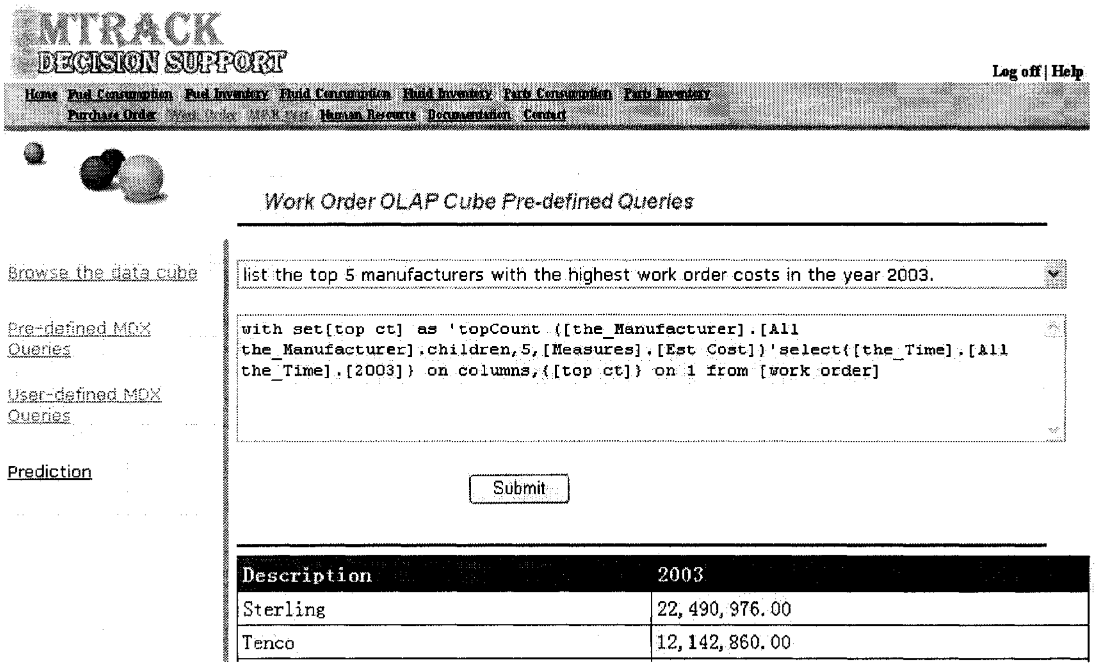
Work Order OLAP Cube Visual Browsing

Microsoft Office PivotTable 11.0

the_Department
Standard General Inc.

	Year			
	⊞ 2002		⊞ 2003	
Cat Desc	Est Cost	Act Cost	Est Cost	Act Cost
⊞ Asphalt Plants	305854	200617	287793	120356
⊞ Automotive Equipment	47978	15427	26701	22582
⊞ Concrete Plants	143634	22580	124776	5841
⊞ Earthmoving Equipment	561394	230800	621888	333813
⊞ Misc. Trailers			15648	23997
⊞ Small Tools	1840	16837	1178	19180
Grand Total	1060700	486261	1077984	525769

Figure 4- 7. Visual Browsing of Data Cubes in Equipment Decision Support System


MTRACK
DECISION SUPPORT
Log off | Help
Home Fuel Consumption Fuel Inventory Fluid Consumption Fluid Inventory Parts Consumption Parts Inventory
Purchase Order Work Order M&R Cost Human Resource Documentation Contact
Browse the data cube
Pre-defined MDX Queries
User-defined MDX Queries
Prediction
Submit

Work Order OLAP Cube Pre-defined Queries

list the top 5 manufacturers with the highest work order costs in the year 2003.

```
with set[top ct] as 'topCount ([the_Manufacturer].[All
the_Manufacturer].children,5,[Measures].[Est Cost])'select{[the_Time].[All
the_Time].[2003]} on columns,{[top ct]} on 1 from [work order]
```

Description	2003
Sterling	22,490,976.00
Tenco	12,142,860.00

Figure 4- 8. Customized MDX Queries in the Equipment Decision Support System

There are no standard query languages for multidimensional databases in the market like the standardized Structural Query Language (SQL) for relational databases; many multidimensional query languages are product dependent. In this research, Multidimensional Expressions (MDX) from Microsoft Corporation (Spoffort 2001) is used. The Web-based system provides two approaches for using MDX: preformulated and user-defined queries. The former is pre-defined in the system to answer the most frequently asked questions. The latter is defined by users on the basis of available templates and can be added to the pool of queries. Figure 4-7 shows one preformulated MDX query and its running results. According to Kimball and Ross (2002), 80~90% of all business users will depend on cube browsing and canned queries for data cube analysis. Therefore, predetermining the most important and used queries is paramount.

The impression seems to be prevalent that multidimensional data cube querying is a complex task requiring a high level of expertise. In fact, the MDX query language is greatly simplified compared with SQL queries used against data tables in relational databases. First, the data cubes are read-only, and thus, MDX is only used for retrieving data in most cases. Second, each data cube has only one fact table, and dimension tables are equally connected to the central fact table. The simplicity of the star schema structure eliminates the need for joining operations from different entities. The purpose of MDX is to define how the user wants to view the subject data: in what perspective and at what level of detail.

The following shows a sample MDX query for comparing the itemized maintenance and repair costs of manufacturer A dump trucks versus those for manufacturer B dump trucks in Division 1 of the contractor's fleet from 2001 to 2004, using the maintenance and repair cost data cube.

```
SELECT
CROSSJOIN (
{[Manufacturer].[All Manufacturer].[Manufacturer B] , [Manufacturer].[All Manufacturer].[Manufacturer
A]} ,
{[costCategory].[All Category].MEMBERS}) ON COLUMNS,
{[Time].[All Time].[2001] : [Time].[All Time].[2004]} ON ROWS
FROM [MRcostCube]
WHERE ([Equipment Group].[All Equipment Group].[Automotive Equipment].[Dump truck or deck (6
wheels)], [Equipment Owner].[All Equipment Owner].[Division 1.], [Measures].[Total])
```

In the example, SELECT is followed by data sources for column and row headers, FROM specifies the data cube, and WHERE defines the filtering criteria. The running results from the system are shown in Table 4-1. Note that the data in the table have been modified for reasons of confidentiality.

To answer the same question with SQL and the original equipment database, a complex SQL script needs to be designed, which would involve at least a dozen tables (normalized) and complex SQL statements. Parameterized SQL-based reports in the original equipment IMS can answer this question provided this problem is addressed in the system design, but the reports cannot cater to the need to answer unanticipated, constantly changing equipment management questions. Furthermore, this sample question can be answered visually using OLAP-supported visual tools, which automatically generate the MDX statements behind the scene. Users do not need to know MDX.

Table 4- 1. Results of Sample MDX Query on Maintenance and Repair Cost Data Cube

	Manufacturer A			Manufacturer B		
	Preventive Maintenance ($CAN)	Running Repair ($CAN)	Work Order ($CAN)	Preventive Maintenance ($CAN)	Running Repair ($CAN)	Work Order ($CAN)
2001	46 794	591 964	91 444	176 154	3 230 614	25 278
2002	86 060	1 190 756	643 683	441 512	3 218 389	1 319 932
2003	78 472	1 193 318	262 688	470 044	1 849 637	2 031 146
2004	92 194	949 521	309 083	561 815	1 547 489	1 828 335

Note: The data in the table have been modified for reasons of confidentiality.

BENEFITS OF EQUIPMENT DATA WAREHOUSE

The developed equipment data warehouse was deployed via a prototype web-based DSS and tested in a local network. It proved that building and deploying a separate equipment data warehouse based on the current equipment data not only eliminated performance and maintenance interference with the original equipment database, but also provided a more flexible and powerful environment for strategic high-level decision support in equipment management. Major advantages over the current system for equipment data analysis include the following:

- *Better user control of data analysis* – Equipment managers usually query the equipment data in broad and unexpected ways to make high-level strategic decisions. The limited flexibility of the current system in presenting equipment data cannot satisfy the various needs of equipment management because of its inherent characteristics of "easy to get data in, difficult to get data out". The equipment data warehouse helps equipment managers accomplish these tasks through the roll up, drill down, and slicing and dicing of different multidimensional cubes without additional assistance from database experts. To answer complex business questions, the multidimensional query language for the generation of datasets from multidimensional data cubes is simpler than SQL for relational databases.

- *A better tool for problem identification and investigation* –The body of equipment data collected for a large fleet on a daily basis is huge, and thus potential problems arising from deteriorated equipment or inappropriate decisions at an operational level are not easy to detect with the current equipment IMS, which relies on the concept of a relational database. However, the equipment data warehouse has the capability to help equipment managers gain valuable insight into these data. The numerical data in the fact table, as performance indicators of the equipment management, can be aggregated along different dimensions at different levels or properties for analysis.

This enables equipment managers to detect problematic areas through interactive exploration of equipment data. For instance, the equipment manager can analyze the labor cost in maintenance and repair works across the company by simply choosing the maintenance and repair cost data cube in the equipment data warehouse and making comparisons over time, according to equipment classes, manufacturers, divisions, departments, and so on. An exceptionally high labor cost in one division for the equipment of a particular manufacturer will be noticeable immediately. The equipment manager may then further investigate the problem by looking at other details. If the system indicates that the equipment from one manufacturer are not being properly maintained, or that the mechanics in the division do not have the expertise for this group of equipment, the equipment manger can then decide, for example, whether to outsource the maintenance and repair work to a subcontractor or local dealer.

- *A better tool for strategic decision making* – Strategic decision-making in equipment management deals with long-term, high-level issues or corporate policies. Strategic decisions answer questions such as "Shall we replace the group of equipment we purchased from this manufacturer in 1985?"; "Shall we outsource the equipment maintenance and repair for this particular project?"; "Which one of two competitive manufacturers shall we buy new equipment from?" Obviously, these questions cannot be answered with a few ad-hoc reports. Comprehensive exploration of the historical data using an equipment data warehouse can assist with these decisions by providing a consolidated view of equipment management data across the company.

- *Improved data sources for knowledge discovery* – Knowledge discovery from database helps the user to detect hidden patterns (common, or unusual) and trends, as well as to predict events. Data in the equipment data warehouse have improved quality, integrated views, and well-organized structure, and therefore can be used as for improved pattern identification and predictive analysis.

SUMMARY AND CONCLUSIONS

Construction equipment management has been drastically simplified because of wide application of equipment management software by large contractors. Large amounts of equipment data collected during daily operations and managements need to be transformed into actionable information for high-level decision support. This chapter summarizes a research project in which a separate equipment data warehouse was built upon the equipment databases that were currently being used by a construction contractor. The data warehouse was deployed with a web server for decision support. In this research project, equipment data from two different equipment databases are consolidated into a single data repository after pre-processing; the equipment data are repackaged into subject-oriented data cubes using multidimensional data models for visual interactive analysis. The equipment data warehouse enables the equipment manager to gain valuable insight into the contractor's equipment data and to answer various equipment management questions in real time. As a result, the decisions on equipment management are based more on facts and less on individual experiences.

The design of the equipment data warehouse was based on bus architecture and multidimensional modeling. The general procedures for designing, implementing, and deploying an equipment data warehouse are introduced in this chapter. The challenges and best practices in building an equipment data warehouse are also discussed.

This research showed that data warehousing concepts and techniques can be applied to computer-assisted construction equipment management to improve the current practice of decision support. This was affirmed by industry practitioners after preliminary use of the system. With the rapid development of computer technology, the high cost of designing and implementing a data warehouse has dropped to a reasonable level. The major providers of relational database products have included data warehouse technology in their database products as either built-in features or add-on products. Therefore, building an equipment data warehouse for decision support in construction equipment management is not only feasible and efficient, but also cost effective.

REFERENCES:

Ahmed, S.M., Ahmad, I., Azhar, S. and Mallikarjuna, S. (2003). "Implementation of enterprise resource planning systems in the construction industry." In Winds of Change: Integration and Innovation of Construction. *Proc., ASCE Construction Research Congress,* Honolulu, Hawaii, 19-21 March 2003. [CD-ROM.] Edited by K.R. Molenaar and P.S. Chinowsky. American Society of Civil Engineers, Reston, Va. dio 10.1061/40671(2003)125.

Bain, T., Benkovich, M., Dewson, R., Ferguson S., Graves, C., Joubert, T., Lee, D., M., Skoglund, R., Turley, P., Youness, S., and Scott, M. (2001). *Professional SQL Server 2000: data warehousing with Analysis Services.* Wrox Press, Birmingham, AL.

Caterpillar (2006). "Caterpillar Equipment Manager." Caterpillar Inc., <http://www.cat.com/cda/layout?m=44501&x=7> (June 6, 2006).

Chau, K.W., Cao, Y., Anson, M., and Zhang J. (2002). "Application of data warehouse and decision support system in construction management." *Automation in Construction,* Elsevier, 12 (2002) 213–224.

Codd, E.F., Codd, S.B., and Salley, C.T. (1993). *Providing on-line analytical processing to user–analysts: an IT mandate,* E.F. Codd and Associates, San Jose, CA.

Computer Guidance (2005). “Construction management system: eCMS.” Computer Guidance Corp., < http://www.computer-guidance.com/products_ construction /financial_management.asp> (June 6, 2006).

Humphries, M., Hawkins, M.W., and Dy M.C. (1998). *Data warehousing architecture and implementation*, Prentice Hall, Upper Saddle River, NJ.

Kimball, R. and Merz, R. (2000). *The data webhouse toolkit: building the web-enabled data warehouse*, John Wiley & Sons, New York, NY.

Kimball, R. and Ross, M. (2002). *The data warehouse toolkit: the complete guide to dimensional modeling*, 2nd Ed., John Wiley & Sons, New York, NY.

Ma, Z., Wong, K.D., Heng, L. and Yang J. (2005). “Utilizing exchanged documents in construction projects for decision support based on data warehousing technique.” *Automation in Construction*, Elsevier, 14 (2005) 405-412.

SAP Inc. (2006). “SAP Netweaver.” SAP Inc., <http://www.sap.com/platform/ netweaver/index.epx> (June 6, 2006)

Shi, J and Halpin, D.W. (2003). “Enterprise resource planning for construction business management.” *J. Constr. Eng. Manage.,* ASCE, 129(2), 214-221.

Spofford, G. (2001). *MDX solutions with Microsoft SQL Server Analysis Services*, John Wiley & Sons, New York, NY.

Stewart, L. (2000). “Giants replace machines to control costs.” *Construction Equipment: Lincolnwood*, 102(3), 62.

Stewart, L. (2004). “Reliability enlists project support for maintenance.” *Construction equipment: Boston*, 107(10), 59.

CHAPTER 5. ASSESSING RESIDUAL VALUE OF HEAVY CONSTRUCTION EQUIPMENT USING PREDICTIVE DATA MINING MODEL[3]

INTRODUCTION

The owning and operation of construction equipment constitutes a significant portion of yearly spending for large contractors engaging in equipment-intensive projects such as earth-moving, highway and industrial installations. According to Stewart (2006), the total construction equipment replacement value in North America of the top 250 construction and mining related companies reached nearly US$ 100 billion in 2006. To minimize the equipment cost per unit of service or maximize the stream of profits generated from the equipment investment, the contactors need to make the right decisions on equipment acquisition, repair/replacement/disposal, and reshuffle their equipment fleet on a regular basis in response to rapid changes in construction markets.

[3] This chapter is a journal paper co-authored with AbouRizk, S. (Department of Civil and Environmental Engineering, University of Alberta), Kim, H. (Department of Civil and Environmental Engineering, Yonsei University, S. Korea), and Zaiane, O. (Department of Computing Science, University of Alberta). Submitted to ASCE *Journal of Computing in Civil Engineering* in April 2007, and accepted in June 2007. The writer of this thesis is the primary contributor and writer of this journal paper, the other authors provided supervisory works.

Among all the factors impacting such decisions, equipment residual value is cited as one of the most important, yet uncertain, with no consensus on the method of determination (Perry and Glyer 1990; Lucko and Vorster 2003). The residual value of construction equipment is the expected selling price in the market at a point of its service life. When the need arises for the determination of equipment residual value for equipment management decisions such as equipment repair, rebuild, disposal, or replacement, the equipment under consideration has not yet been subject to the pricing process in the market (e.g., auction), therefore the market value of a piece of equipment can only be estimated based on experience, historical auction cases, or postulated formulae. Since a wide variety of factors exert impact on the market value of construction equipment, including age, manufacturer, model, intensity of use, care, as well as market supply and demand, it is not surprising that no industrial criteria currently exist for an evaluation on the price of used construction equipment.

Research on estimating the depreciation and residual values of heavy equipment is conducted extensively in the agricultural and forestry industries, where similar or same equipment is generally used, and equipment cost constitutes a significant portion of the total production cost. Previous research efforts tended to use statistical regression approaches in order to establish functional relationships between the residual value of machinery and the known impact factors (McNeill 1979; Reid and Bradford 1983, Cross and Perry 1995; Unterschultz and Mumey 1996). In particular, the Cross and Perry (1995) method made its way to the American Society of Agricultural Engineers (ASAE) standard, "Agricultural Machinery Management Data" (ASAE 2003) for estimating the residual values of major types of agricultural equipment. In the construction industry, Lucko et al. (2006) investigated the effectiveness of applying a similar statistical approach to estimate the residual value of heavy construction equipment based on equipment auction data.

Several issues make it difficult or impossible to find a universal solution using the statistical regression method. First, different equipment categories display different behaviors in depreciation, as pointed out by Cross and Perry (1995) for agricultural machinery. The same observation holds true for construction equipment; large, heavy-duty, special-purpose equipment depreciates faster than small, multi-function equipment. Second, some influencing factors on equipment residual value are dynamic and changing constantly, and their degree of impact may fluctuate over time (e.g., microeconomic indicators), and some other factors are very difficult to quantify as input for the regression model (e.g., technological progress and renovation). Third, the regression model is more appropriate for a specific class of equipment in a narrow range of model series, but this makes it difficult to collect sufficient samples to warrant statistical significance. Lastly, multicollinearity is a widely acknowledged problem for statistical regression that influences both the stability and accuracy of the derived statistical regression model. Some features of data samples (e.g., equipment age versus usage, and equipment age versus condition ratings) for equipment price evaluation have a relatively high coefficient of correlation.

This chapter introduces an approach for predicting the residual value of construction equipment using the data mining technique. Data mining is an interdisciplinary science of statistics, machine learning, database, information theory, visualization, etc., with the objective of discovering valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al. 1996). As a hybrid of multiple disciplines, data mining integrates their perspective plausible features for knowledge extraction, validation, presentation, and deployment. The technique for which we advocate is a predictive data mining algorithm, the AutoRegression Tree (ART) proposed by Meek et al. (2002). This technique is utilized in this research to build a tree-structured non-linear regression model based on large amounts of construction equipment auction data. The ART algorithm is chosen in this research for the building of predictive data mining models in consideration of its high interpretability and accuracy. A prediction model represented as a tree structure is more meaningful to decision makers in construction equipment management, because a tree-type model is more accurate in mimicking the non-deterministic non-linear relationship between the input-output variables and is much simpler to interpret. The logical meaning of the tree-type model is inherently apparent, providing a convincing and transparent reasoning path to the prediction. Simply put, the analysis by means of a tree (or decision tree) is to settle on a set of "if-then" split conditions that allow for accurate predictions via careful data partitioning. After validation, the data mining model is embedded in a construction equipment information system for the prediction of equipment residual value. The integration of a data mining model with the equipment information system provides the user with the capability of browsing through the visualized data mining model, making prediction on the fleet equipment while being informed of the reasoning process. In addition, the automated processing of data mining makes it possible to update the model in real time to reflect recent changes in the equipment auction market. After comparing the data mining approach with the statistical regression approach, the chapter concludes that the data mining model can better capture the complex and dynamic relationship between equipment residual value and its various influencing factors, and is easier to integrate with the current equipment information management system.

LITERATURE REVIEW

Depreciation is defined as the decrease in the residual value of equipment over time. Considering the fact that equipment cost constitutes a significant portion of the total production cost and the different options available for equipment acquisition, accurate evaluation of depreciation is crucial for a successful business in agriculture, mining, or construction. Therefore, much research was devoted to quantifying the depreciation patterns of heavy equipment.

One of the most comprehensive studies on depreciation of agricultural equipment was the one conducted by Cross and Perry (1995). Their research objective was to identify the most fitted mathematical functions for modeling the relationship between equipment Residual Value (RV) and the known explanatory variables as stated in the following general form:

$$RV = f\left(age, usage, care, manufacturer, auctionType, region, microeconomicsVariables\right) \quad (1)$$

Given the fact that different types of equipment display different behaviors of depreciation, and that depreciation normally occurs in a nonlinear way, Cross and Perry (1995) proposed to transform the dependent variable RV and the two most statistically significant explanatory variables *age* and *annual hours of use* using the following Box-Cox transformation:

$$y^{\lambda} = \begin{cases} \dfrac{\left(y^{\lambda} - 1\right)}{\lambda}; & \lambda \neq 0 \\ \log(y); & \lambda = 0 \end{cases} \quad (2)$$

Depending on the transformation parameters λ obtained by maximum likelihood estimation or the Bayesian methods from the data, the Box-Cox transformation enables the regression model on RV to encompass a wide variety of functions, such as linear, exponential, and logarithmic. Using the Box-Cox transformation, Cross and Perry (1995) estimated price models for nine types of machinery and equipment used in agricultural production. Based on their research results, ASAE (2003) recommended a generalized regression formula for the estimation of residual value percentage (residual value divided by original list price) using *equipment age* and *annual hours of use* with different coefficients for different types of equipment.

Other similar research includes that conducted by McNeill (1979), Reid and Bradford (1983), Perry et al. (1990), and Unterschultz and Mumey (1996). To evaluate and compare different research results and functional forms, Dumler et al. (2000) and Wu and Perry (2004) each evaluated and compared different functional forms and debated over their applications in the agricultural industry.

Statistical regression was applied for building prediction models for the residual value of heavy construction equipment (Lucko and Vorster 2003; Lucko et al. 2006; Lucko et al. 2007). Upon identification of influential factors to the residual value of construction equipment, the researchers proposed several forms of multiple linear regression models (plain model, best model and trade model) based on equipment auction data after statistical tests. The research confirmed that while equipment age is indeed the most significant influential factor of equipment residual value, other factors including the manufacturer, condition rating, auction region, and microeconomic indicators also contributed to the "goodness-of-fit" of the prediction model with statistical significance (Lucko et al. 2006). In their work, they used a narrow range of equipment, namely, the track dozers of 74.57-148.39 KW (100-199HP) were selected as an implementation example to illustrate the methodology of applying regression analysis to predict the residual value of heavy construction equipment.

In the area of construction research, there have been many applications involving prediction using inferred models from data. Lee et al. (2004) used the GUIDE regression tree algorithm to quantify the cumulative impact of change orders on productivity; Arditi and Pulket (2005) applied a boosted decision tree for predicting the outcome of construction litigation; Kim et al. (2004) proposed a Neural Network-based classification system for automatic assessment of aggregate quality using laser imaging results.

DATA MINING FOR PREDICTION OF EQUIPMENT RESIDUAL VALUE

Prediction of a numerical value is a common data mining task to infer the most likely value of a response variable based on the known predictor variables, and can be represented in the following generalized form: $y = f\left(x_1, x_2, ..., x_n; r_1, r_2, ..., r_k\right)$, where y is the target variable of the continuous data type, x_i (i=1,2,…,n) are predictor variables of either categorical or continuous data types, and r_i (i=1,2,…,k) are the model parameters. Instead of a mathematical or statistical function as defined traditionally, f( ) stands for a data mining model representing the discovered patterns or rules from observation data by a data mining algorithm. The model parameters r_i are introduced in some algorithms to fine-tune the model structure or incorporate prior knowledge into the model generation process.

Predictive data mining uses complex computer algorithms to search through the data and generalizes the rules and patterns reflecting the relationship between the target variable and predictor variables. The "Divide-and-conquer" and "heuristic" methods are two representative ones for inferring predictive models from data:

- *The Divide-and-Conquer Method* – The algorithm searches over the data space and recursively partitions it into subspace where more pure information or promising relations can be found. For example, the algorithms of a decision tree family use a measurement such as information gain or chi-square test to search for most informative splitting of data space by an input variable as well as a split-on value so that the partitioned data space contains more pure information on the prediction results. Using the prediction problem for equipment residual value as an example, after the algorithm identifies the input variable *equipment age* as the most relevant and informative feature to the residual value, it would have a tendency of splitting the data using *equipment age*. Exemplar algorithms using "divide-and-conquer" method are C4.5 by Quinlan(1993), Categorization and Regression Tree (CART) by Breiman et al. (1984) and AutoRegressive Tree(ART) (Meek et al. 2002).

- *The Heuristic Method* – Both Artificial Neural Network (ANN) (Anderson 1995) and the Support Vector Machine (SVM) (Burges 1998) use a trial and error method to iteratively obtain an optimized predictive model based on predefined error measurement.

Though different algorithms use different methods for model inference, they have some common features which make them excel over traditional statistical regression approaches for predictive modeling on equipment residual value:

- The models are inferred from data (recorded facts) with minimum user input. In contrast to the hypothesis-and-testing approach used in statistical regression, search and generalization is used in predictive data mining for the inference of patterns or rules in training data. The assumptions on statistical distributions or postulated functional forms make a statistical regression model subjective, and vary from one model to another; whereas, a data mining model is derived by an algorithm based on the available data, and involves minimum user interference in model generation.

- Data mining models are represented by a computer model capable of storing complex rules and patterns by utilizing data structures, algorithms, and indexes. Therefore, complex rules and patterns that exist in data can be uncovered and represented.

- Many data mining models, such as the family of decision trees or Bayesian inference, can be visualized in an intuitive manner for human interpretation.

- Data mining models adapt to changes easily. Since data mining can be designed as an automated process in a computer system, a data mining model can be updated in real time after the updating of the data sources.

AUTOREGRESSIVE TREE (ART) ALGORITHM

The AutoRegressive Tree is a data mining algorithm proposed by Meek et al. (2002) to establish a nonlinear relationship between a set of explanatory variables and a target numeric variable through the exploration of the training dataset. Much research (e.g., Chipman et al. 2002) proved that, for a large number of domain problems, the data space can be partitioned into sub-regions where a simple linear regression model exists for each sub-region. Since the partitioning of the data space can be conveniently expressed in a decision tree structure with subsets of data residing in tree leaves where regression models are grown, this type of model is called treed regression (Alexander and Grimshaw 1996). Different approaches have been proposed to induce the decision tree structure with linear regression models at its leaf nodes, such as m5 by Quinlan(1992), RETIS by Karalic (1992), and Bayesian Treed Models by Chipman et al. (2002). Though different algorithms generate similar treed models with a common goal of partitioning data space into subsets in such a way that the overall goodness-of-fit of the model to training data is maximized, different mechanisms and measurements are used to partition the data space and build local linear models. The ART algorithm uses the Bayesian technique to generate the tree structure and model parameters.

The Bayesian updating technique is a statistical inference method for model induction based on both prior assumption and observed facts. In contrast to the traditional statistical method assuming the model parameters are fixed, the Bayesian updating method considers the model parameters as changing variants which can be described by the current statistical distribution. A prior distribution of model parameters is assumed based on past experiences or subjective judgment. However, if the factual information is available, the prior distribution is updated by the likelihood that the observed factual data fall into the prior distribution. This updating process draws prior probability distribution closer to its true distribution and hence the posterior probability of model parameters is more accurately obtained.

The ART algorithm uses this posterior probability of model structure *s* to compare different alternatives of tree topology in terms of their goodness-of-fit to the training dataset. Based on the Bayesian theory:

$$p\left(s\middle|d\right)=\frac{p(s)p(d\mid s)}{p(d)} \tag{3}$$

Where

p(s|d) – Given training data *d*, the probability of fitting model structure *s*

p(s) – Prior probability of model structure *s*

p(d|s) – Marginal probability of observing data *d*, given model structure *s*

p(d) – prior probability of data *d*

Because the prior probability *p(d)* of the training dataset is a constant, Meek et al. (2002) defines *p(s)p(d|s)* as the Bayesian score for the ART model. The first product *p(s)* is the assumed structure prior, which is a subjective judgment on the probability distribution of model parameters, whereas the second product $p(d|s)$ is the marginal likelihood of training data falling into the assumed prior distribution for given structure s.

For each candidate model structure in the data partitioning process, ART builds a normal multi-linear regression model for the subset of data at each leaf node. Assuming the linear model parameters at the leaf nodes are independent from each other, the Bayesian model score is calculated as:

$$score(s) = \prod_{i=1}^{L} LeafScore(l_i) \tag{4}$$

The LeafScore in equation (4) at each leaf node is calculated according to (3), using an assumed prior distribution $p(s)$=$0.1^{|\theta|}$ (θ is the number of model parameters) and Normal likelihood function. See Meek et al. (2002) for details on the Bayesian score calculation for each leaf node.

The ART algorithm uses the "divide-and-conquer" method to partition the data space and builds regression models at each leaf node. The pseudo-code of the algorithm is shown below:

List 1. Pseudo-code of AutoRegressive Tree Algorithm

```
1. #Start with the root node
2. build a linear regression model at the root node
3. calculate Bayesian model score
```

```
#Compare alternative splitting options
For each input attribute A
        #Determine candidate split values split[]
        # in case of categorical attribute
        If input attribute A is a categorical attribute
        set split[]= distinct nominal values of A
        else
        #in case of continuous attribute
        set split[]= 7 splitting points of 8 equal-probability areas assuming the
         input attribute conforms to a Normal distribution
        end if

        # Loop through every split-on value to evaluate current splitting option
        For each value in split[]
                partition the data using current attribute A and split-on value
                build linear regression model for each leaf node
                calculate Bayesian model score
                calculate increase in model score compared with model prior to
                splitting
                store current splitting parameters and model score
        end For

End For

choose attribute and split-on value which leads to highest increase in model
score

#continue with the recursive partitioning
split data using the selected attribute and split-on value
recursively repeat the above process for each subset of data
```

```
27. #Terminate splitting process
28. If      the splitting will not increase the model score or
            the number of cases in the leaf node is less than specified threshold
            value
29. terminate splitting
30. end If
```

To improve the model accuracy, the ART algorithm uses a dynamic splitting method proposed by Chickering et al. (2001) to determine candidate values for data splitting. Instead of determining these values at the beginning of the algorithm and using them for all the subsequent partitioning, the algorithm re-calculates the candidate split-on values for features of the data subset at each step of partitioning:

- For a categorical attribute, it uses distinct nominal values in the subset of data (lines 8-9 in List 1), and
- For a continuous attribute, it uses 7 intermediate points that split the attribute values into 8 equal-probability areas assuming Normal distribution (lines 10-12 in List 1).

ART MODEL FOR PREDICTION OF EQUIPMENT RESIDUAL VALUE

Although it is theoretically possible to build a single predictive data mining model for all types of heavy construction equipment so long as they are fully represented by training data, the model of this scale would be of poor quality and difficult to interpret.

Therefore, separate data mining models are built for each major category of heavy construction equipment. In this section, the data mining process is exemplified by selecting the equipment category of wheel loaders for model building and validation.

Data Sources

The primary data source for model building is Last Bid®, an online construction equipment database covering up-to-date auction results across the U.S. and international markets (Prism Business Media Inc. 2005). The wheel loader auctions across the U.S. and Canada from 1996 to 2005 are selected with the available information on make, model, year of build, auction year, conditions, auction locations, and transaction price. The "usage of equipment" information is missing from this data source because "it is difficult to confirm the data with confidence" (Vorster 2004). Other potential factors of influence on auction results, (i.e. Gross Domestic Product (GDP) and yearly construction investment (CI)) are obtained from the U.S. Bureau of Economic Analysis and Statistics Canada.

Feature Selection

The residual value of a wheel loader is influenced by various features which have a potential impact on its market transaction price. To enable the data mining model to capture the inherent relationship, all the factors of potential influence to the residual value should be fully identified, and some features need to be transformed either to fit the model input or improve the model accuracy. Two examples of feature transformation for this model are *equipment age* and *auction location*. *Equipment age* measures the number of years the equipment has been in service at the time of auction, and has a direct impact on equipment residual value, therefore, it is derived and used together with *auction year* to describe the timeline of the auctioned equipment. For predictor variable *auction location*, the state/province is given in the auction data as a characterization attribute. To better represent the location variable, a simple transformation is conducted to derive the country of auction and the region of auction as two additional candidate attributes for this variable. A calculation of information gain based on the information theory (Shannon 1948) determines that the region of auction is the best attribute among the three (country, region, state/province) to represent the *auction location* because it has the maximum discriminating power on the response variable of *auction price* in the dataset.

The usage of equipment (accumulated operation hours of a wheel loader) is considered an important factor on equipment residual value, but it is not available from the data source. Assuming normal use of equipment in its life time, the age and hours of use have a high coefficient of correlation (e.g., 0.75 in a research conducted by Perry et al. (1990) on farm tractors). Therefore, it is safe to ignore this variable while including the age in years in training data to represent the usage of equipment.

To determine the condition rating of a piece of equipment with minimal bias, evaluation of equipment needs to follow the detailed guidelines set out by the equipment auctioneer, and is usually carried out by accredited equipment appraisers. The determination of condition rating for construction equipment is also explained by Lucko et al. (2006).

Finally, the following features are selected for building the predictive data mining model for equipment residual value:

- Make: Manufacturer of wheel loader;
- Model: Model of wheel loader;
- Horsepower: the rated engine horse power (HP);
- Age in years: obtained based on the year of build and the auction year;
- Auction year: the year at which auction occurred;
- Auction location: the auction region (U.S. Southeast, Southwest, West, Mideast, Northeast, and Canada);
- Condition rating: the rating of equipment in terms of physical conditions (New, Excellent, Very Good, Good, and Fair);
- Annual construction investment: Annual construction investment in U.S. and Canada in US$ million at the year of auction;
- GDP: the Gross Domestic Product in U.S. and Canada in US$ billion at the year of auction.

To measure the equipment price in constant dollars, the response variable *auction price* is indexed to the year 2000 based on the consumer price index obtained from the U.S. Bureau of Labor and Statistics as well as Statistics Canada.

Data Quality Control

Data quality in data mining measures the overall fit of data to knowledge generation. In addition to the general requirements, such as consistent format, no missing values and outliers, for data quality in decision analysis, the data should be representative of all the features in full range and unbiased quantity. If there is a systematic lack of attribute values (e.g., lack of equipment auction cases in a price range, or lack of an auction region in the training data) the predictive model for equipment residual value would have a poor accuracy of prediction for the defined domain problem.

The attribute values of "unknown" for *equipment condition* are considered as missing values, and replaced by the mode value of "good". Some auction cases of transaction prices over US$ 200,000 are removed from the collected data as isolated cases resulting from customized build or special attachments. Finally, a total of 8,589 effective cases are obtained for model generation.

A preliminary check is conducted as to the representation of predictive features and auction results by the training data. Figure 5-1 shows the histogram of the discretized auction price, the auction prices from 6,000 to 200,000 are binned into 14 cohorts based on Sturge's rule (Number of bins=1+3.3*Log(N), N is the number of data points), with no obvious missing data in any price cohorts. The representation data on each predictor variable is checked in the same principle based on its frequency diagram (for categorical variables) or histogram (for continuous variables).

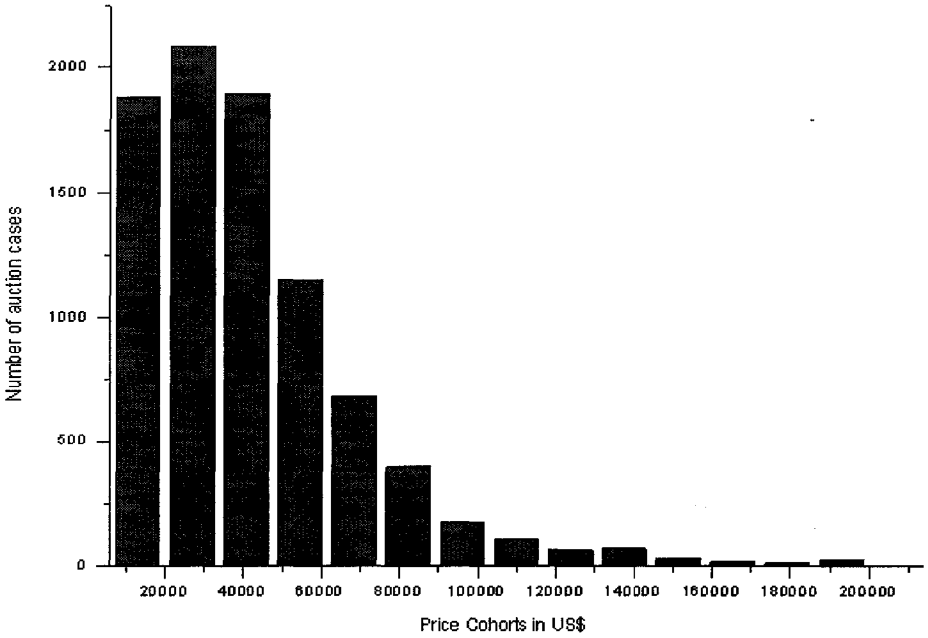


Figure 5- 1. Histogram of Wheel Loader Auction Price

Model Generation and Validation

The ART algorithm includes two parameters for model structure control: one is the coefficient of complexity γ controlling the growth of the tree; a higher value increases the likelihood of node splitting to generate a bushy tree, and the other is the minimum number of cases M in each leaf node. The sensitivity analysis of γ and M on prediction accuracy, using 90% of the data for training and the remaining 10% for validation, found out that prediction accuracy is not sensitive to γ but is sensitive to M. The model parameters γ and M are determined as 0.5 and 20 respectively for the final model generation.

To verify the stability and accuracy of the predictive model for equipment residual value, a 10-fold cross-validation method is used for model generation and validation. The wheel loader training data comprised of 8,589 effective cases is randomly divided into 10 partitions of approximately equal size, each containing around 859 cases. A data mining model is generated and validated for 10 iterations according to the following procedure: hold each partition and use the remaining of the entire dataset as the training data to generate the ART model, and then use the reserved partition as out-of-sample data for the validation test.

Test results indicate that the 10 models are similar in their tree topography and linear regression functions at leaf nodes. Figure 5-2 partially shows the structure of the derived tree model at the first few levels. A comparison of the 10 models found that the top levels of the tree structure starting from the root node are the same, while some nodes at the bottom levels vary slightly.

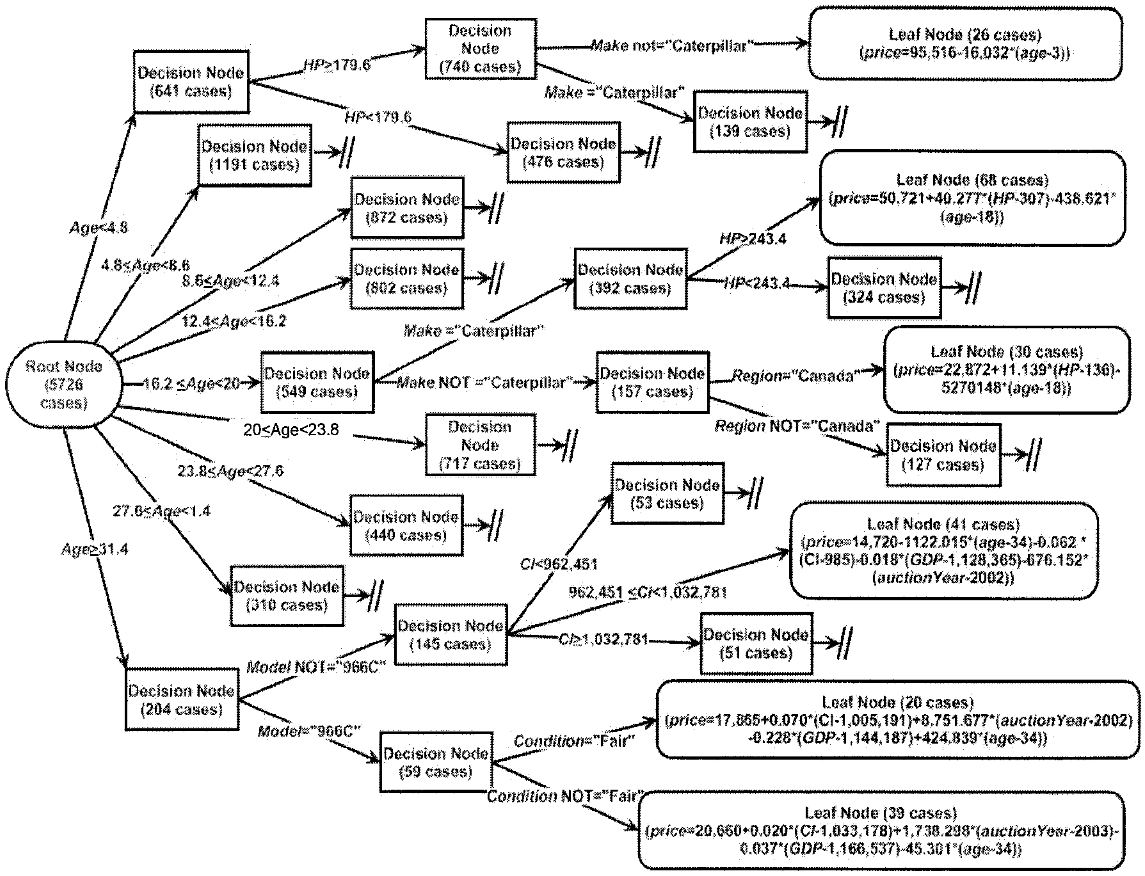

Root Node (5726 cases)
Age<4.8
4.8≤Age<8.6
8.6≤Age<12.4
12.4≤Age<16.2
16.2 ≤Age<20
20≤Age<23.8
23.8≤Age<27.6
27.6≤Age<1.4
Age≥31.4
Decision Node (641 cases)
HP≥179.6
HP<179.6
Decision Node (740 cases)
Make not="Caterpillar"
Leaf Node (26 cases) (price=95,516-16,032*(age-3))
Make ="Caterpillar"
Decision Node (139 cases)
Decision Node (476 cases)
Decision Node (1191 cases)
Decision Node (872 cases)
Decision Node (802 cases)
Decision Node (549 cases)
Make ="Caterpillar"
Decision Node (392 cases)
HP≥243.4
Leaf Node (68 cases) (price=50,721+40.277*(HP-307)-438.621*(age-18))
HP<243.4
Decision Node (324 cases)
Make NOT ="Caterpillar"
Decision Node (157 cases)
Region="Canada"
Leaf Node (30 cases) (price=22,872+11.139*(HP-136)-5270148*(age-18))
Region NOT="Canada"
Decision Node (127 cases)
Decision Node (717 cases)
Decision Node (440 cases)
Decision Node (310 cases)
Decision Node (204 cases)
Model NOT="966C"
Decision Node (145 cases)
CI<962,451
Decision Node (53 cases)
962,451 ≤CI<1,032,781
Leaf Node (41 cases) (price=14,720-1122.015*(age-34)-0.062 * (CI-985)-0.018*(GDP-1,128,365)-676.152* (auctionYear-2002))
CI≥1,032,781
Decision Node (51 cases)
Model="966C"
Decision Node (59 cases)
Condition="Fair"
Leaf Node (20 cases) (price=17,865+0.070*(CI-1,005,191)+8,751.677*(auctionYear-2002) -0.228*(GDP-1,144,187)+424.839*(age-34))
Condition NOT="Fair"
Leaf Node (39 cases) (price=20,660+0.020*(CI-1,033,178)+1,738.298*(auctionYear-2003)-0.037*(GDP-1,166,537)-45.301*(age-34))

Figure 5- 2. Partial ART Data Mining Model for Prediction of Equipment Residual Value (Obtained using Microsoft SQL Server 2005 Analysis Services; for lack of space, most of the decision paths are not shown here)

In addition to the regression tree, the algorithm also generates an output which ranks the predictor variables as per their discriminating power on equipment auction price. The same ranking is generated from all the 10 iterations as following in a decreasing order: *equipment age, horsepower, make, auction year, GDP, Construction Investment, model,* and *condition*. The fact that *equipment age, horsepower*, and *make* are the top three relevant features for prediction can be observed from the tree structure: the three features are most frequently selected at the top levels to partition the data space

(Figure 5-2). Another finding of interest is that the algorithm ranks *equipment model* as one of the least predictive features on equipment residual value even though equipment model is typically one of the decisive factors on its residual value. To explain this result, only the top 50 equipment models with a large number of auction cases accounting for the over 70% of the total cases are selected for the model building, and it turns out that *equipment model* is ranked as one of the powerful features for prediction. In consideration that there are over 300 models of wheel loaders in the training data, the information conveyed by this feature on equipment residual value is noisy, therefore the algorithm cannot identify *equipment model* as a persuasive explanatory variable on equipment residual value.

To cross-validate the prediction accuracy of the data mining model, three measures are used to evaluate prediction errors in each iteration:

1. Relative Squared Error (RSE): Relative Squared Error evaluates the percentage of the total squared error between the predicted value and actual value out of the total squared error if using the average of actual values as a prediction. That is, the total squared error is normalized by dividing it by the total squared error of a simple default predictor using the average of the actual values for prediction.

$$RSE = \frac{\sum_{i=1}^{n} (p_i - a_i)^2}{\sum_{i=1}^{n} (a_i - \overline{a})^2} \qquad (5)$$

Where p_i is the predicted value; a_i is the actual value; and $\bar{a}$ is the average of actual values in the validation sample.

2. Root Relative Squared Error (RRSE): Root Relative Squared Error takes the root of RSE so that the error rate is reduced to the same magnitude as the response variable (i.e. the same dimensions as the quantity being predicted).

3. Mean Absolute Error (MAE): Mean Absolute Error is the average of absolute prediction errors (i.e. without taking into account the sign of the error). Contrary to the Mean-squared error, which tends to exaggerate the effect of outliers, the Mean Absolute Error is not affected by extreme values and all sizes of errors are treated evenly.

The 10-fold cross-validation generated similar test results on RSE, RRSE, and MAE as summarized in Table 5-1. On average, after using the predictive model, the total squared error is reduced by 94.5% compared to using the average value as the prediction. This error reduction in the same magnitude is 76.7% and the mean absolute error of prediction results is US$ 4,248. About 1% of the cases are indicated as "unpredictable" by the model; this is because when a test case falls into a leaf node where no regression is available due to insufficient information for the creation of a regression model, or if the number of cases is less than 20.

To evaluate the dispersions of prediction errors, the prediction errors in US$ are transformed into error percentage rates after being divided by their perspective actual values, and presented in box plots for each iteration. As shown in Figure 5-3, the 10 boxplots show similar characteristic values (i.e. quartile values at 25%, 50% and 75% denoted by the top, middle and bottom lines of the box). The prediction error rates for the middle 50% cases (the cases with prediction errors between upper and lower quartiles) are less than 8.5%, whereas a high degree of dispersion is observed from all the 10 sets of test results. Each set of test results has a small number of outliers with prediction errors outside 1.5 times Inter-Quartile Range (IQR) from the median. The prediction errors of each set follow a bell-shaped distribution, yet with open ends on both sides.

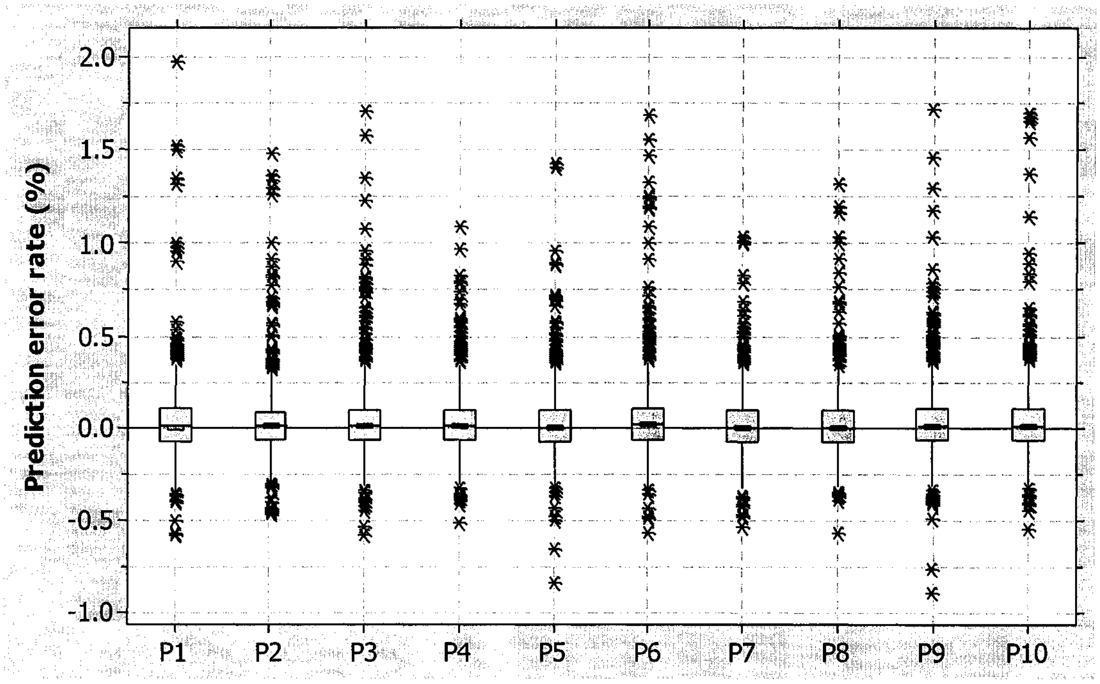


Figure 5- 3. Boxplots for Prediction Error Rates of Cross-Validation Tests

Both the cross-validation tests and boxplots verify the stability of the inferred data mining model. The mean absolute error of US$4,248, representing the deviation of predicted market prices of the equipment from their transaction prices, is less than 10%

of the average equipment transaction price and deemed acceptable for heavy construction equipment.

Table 5- 1. Test Results of 10-fold Cross-Validation

Partition	Missing	Relative Squared Error (RSE)	Root Relative Squared Error (RRSE)	Mean Absolute Error (MAE)
1	11/859	5.3%	23.0%	4,289
2	8/859	5.1%	22.6%	4,105
3	18/859	7.7%	27.7%	4,594
4	17/859	3.9%	19.7%	3,972
5	2/859	6.6%	25.6%	4,064
6	5/859	7.7%	27.7%	4,743
7	0/859	5.7%	23.9%	4,391
8	10/859	3.9%	19.7%	4,211
9	11/859	4.1%	20.1%	4,032
10	2/859	5.4%	23.1%	4,075
Average	8.4/859	5.5%	23.3%	4,248

The dispersion of error distribution shows that a small percentage of cases have a high degree of deviation, which is attributed to the fact that these isolated cases are not well represented in the model; on the other hand, the equipment age being the most important predictor introduces a certain degree of error when being measured in the whole number of years. The predictor accuracy of the current model can be improved if equipment usage data, rather than equipment age, is available to gauge the intensity of equipment usage. To improve the prediction accuracy, the final model is built using the entire dataset.

COMPARISON OF ART WITH ANN AND MLR MODELS

The same prediction problem is also modeled using the Artificial Neural Network (ANN) model and the Multivariate Linear Regression (MLR), and validated with 10% hold-out test. The prediction accuracy is evaluated on the ANN and MLR models using the same three measures for the ART model, as summarized in Table 5-2. It shows that both the ANN and MLR models perform worse than the ART model in explaining the variability of auction price and also generate a higher degree of prediction error.

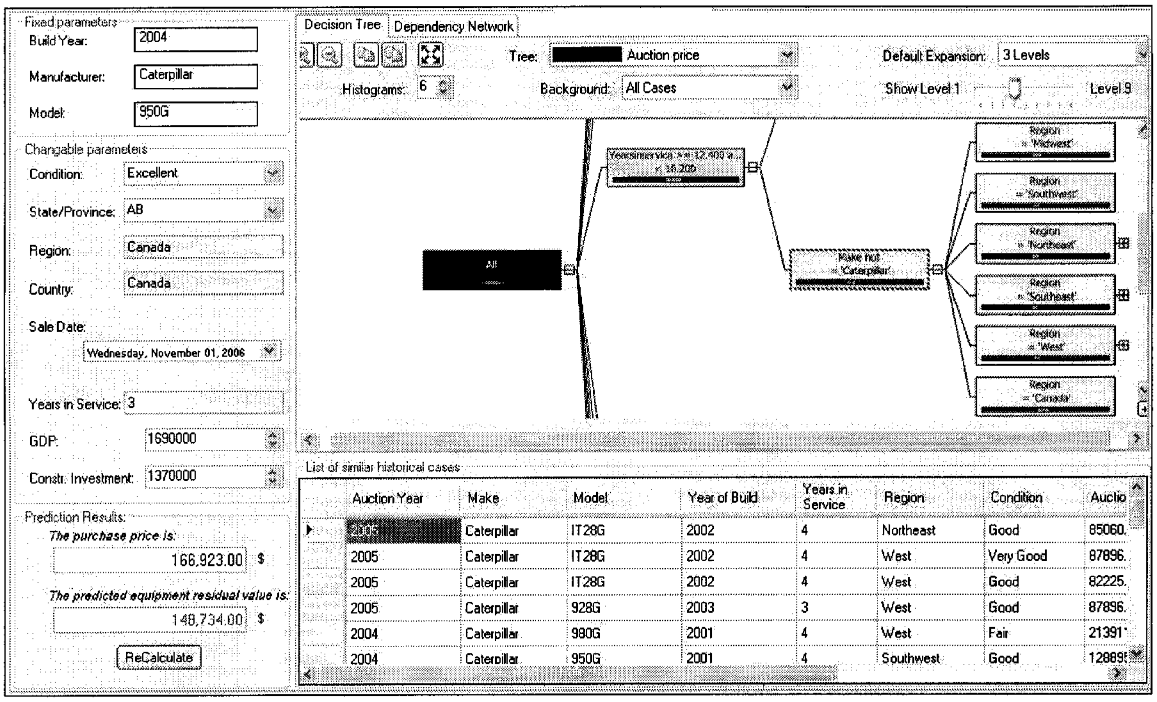
The higher prediction error of the ANN model can be explained by the way it handles the categorical input attributes: each categorical input attribute is encoded into N-1 (N is the number of nominal values for the attribute) binary attribute as model input. A large number of input attributes are created for a predictive model if there are many categorical features as input, such as is the case in this problem. As a result, the large number of input variables quickly decreases the prediction accuracy of the ANN model. The high prediction error of the MLR model is attributed to its over-simplified assumption of the linear relationship between the auction price and its predictor variables using a single statistical regression model.

Table 5- 2. Comparison of ART with ANN and MLE in prediction accuracy

	RSE	RRSE	MAE
ART	5.5%	23.3%	US$ 4,248
ANN	22.3%	47.2%	US$ 7,274
MLR	40.4%	63.6%	US$ 11,069

DEPLOYMENT OF PREDICTIVE DATA MINING MODELS FOR EQUIPMENT RESIDUAL VALUE

The data mining models for the prediction of equipment residual value were built for major categories of equipment and deployed in a construction equipment information management system for testing. Figure 5-4 shows a screenshot from the system with the integrated data mining module. After choosing a piece of equipment from the fleet database, the user gets the projected market price under given conditions. The sensitivity of the auction price on the sale date and location can be conducted on a piece of equipment by changing the input parameters. The user can also browse through the regression tree model by visually expanding the tree structure, and moving the mouse cursor over a leaf node to view the the multivariate linear regression model. The related historical cases in the leaf node used for prediction on the current case are automatically retrieved for comparisons. The visualization of the regression tree model allows the user to reaffirm the predicted transaction price or analyze a doubted prediction.



Fixed parameters

Build Year: 2004

Manufacturer: Caterpillar

Model: 950G

Changable parameters

Condition: Excellent

State/Province: AB

Region: Canada

Country: Canada

Sale Date: Wednesday, November 01, 2006

Years in Service: 3

GDP: 1690000

Constr. Investment: 1370000

Prediction Results:

The purchase price is: 166,923.00 $

The predicted equipment residual value is: 148,734.00 $

Tree: Auction price

Default Expansion: 3 Levels

Histograms: 6

Background: All Cases

Show Level 1 Level 9

List of similar historical cases

Auction Year	Make	Model	Year of Build	Years in Service	Region	Condition	Auctio
2005	Caterpillar	IT28G	2002	4	Northeast	Good	85060.
2005	Caterpillar	IT28G	2002	4	West	Very Good	87896.
2005	Caterpillar	IT28G	2002	4	West	Good	82225.
2005	Caterpillar	928G	2003	3	West	Good	87896.
2004	Caterpillar	980G	2001	4	West	Fair	21391'
2004	Caterpillar	950G	2001	4	Southwest	Good	12889!

Figure 5- 4. Screenshot of the Equipment Information Management System

(Note: the diamond location denotes the mean of auction price, and its width denotes the variance of auction price on the node).

Due to copyright issues, the online equipment auction data source does not allow for a live connection to external applications. Otherwise, the data mining for prediction of equipment residual value can be designed as a fully automated process. Under a fully automated data mining process, all the procedures including feature extraction and transformation, data cleaning, modeling, validation, and deployment can be implemented programmatically in a streamlined process. One recommendation on such a data mining enabled system is to design it as a two-mode application: user mode and developer mode. The user mode interacts directly with the users by exposing the data mining models for browsing, analysis and prediction; whereas, the developer mode implements the data mining process with the visual capability of updating the training data, fine-tuning the model parameters, and validating the model.

DISCUSSIONS

Though data mining serves as a unique modeling approach for predictive analysis in construction equipment management, there are several issues which have to be handled carefully to ensure that the derived model truly reflects the inherent relationships or rules in reality:

- *Problem definition* – The inference capability and storage mechanism of data mining greatly loosens the problem scope definition compared with the traditional statistical methods; however, a data mining problem needs to be defined at an appropriate level to control the model complexity, and model interpretability. This research divides the equipment into major categories for which separate models are built up for prediction. The categorization of equipment can be interpreted as building the first level of a decision tree for all the equipment in a manual process, followed by the automatic tree growing for the remaining part.
- *Data source selection* – The data mining algorithms infer the model structure out of data, hence, the data source used for mining should contain unbiased, fully-

represented information on the data mining problem. The data sources should be sufficient and reliable, with a fair level of data quality.

- *Data preparation* – As demonstrated in this research, the data used for the model inference should be appropriately represented by a set of predictive features and also quality insured. Common methods for preparation are feature transformation and pre-selection, data validation, evaluation of data representation on both input features and output results, etc.

- *Model inference* – Select a data mining algorithm which fits into the data mining problem, is easy to use and simple to interpret. In this application, neither neural network nor CART algorithms were selected because the former generates a model that is difficult to interpret while the latter only predicts a ranged value for the response variable.

- *Model validation* – The data mining model must be validated in order to be effective. Out-of-sample cases should be used for the validation of the model inferred from in-sample data. The multi-fold cross-validation is used in this research to verify that the generated data mining model is stable, representative, and accurate.

Finally, incorporating the prior domain knowledge into the data mining process is both necessary and important, in spite of the fact that data mining is an automated or semi-automated process of discovering knowledge from data. The feature selection and model validation in this research demonstrates the importance of domain knowledge for data mining. It would be difficult to obtain an unbiased accurate predictive model if model features are not represented by the data or the generated model is not validated based on domain knowledge.

SUMMARY AND CONCLUSIONS

This chapter presents a data mining approach for the prediction of construction equipment residual value and its deployment in a construction equipment information system. In summary, the data mining-based solution provides some benefits which cannot possibly be achieved with the current rule-of-thumb or statistical methods:

- The data mining model is capable of capturing the relationships, patterns and rules that exist in a dynamic and complex environment. The prediction of equipment residual value involves a large number of influential factors which are subject to changes over time. Statistical regression tackles this problem by inferring the statistical relationship between the residual value and a few meticulously selected predictors and cannot easily adapt to changes.

- The data mining model is primarily data driven and less dependent on personal experiences. The data mining algorithm searches over the data space to infer a model structure that reflects the relationship between the equipment residual value and its influential factors. Prior knowledge can be incorporated into a data mining process, but the model inference is based more on recorded facts, less on individual experience.

- Many data mining models are transparent and interpretable. Many data mining models, such as the decision tree, and the Bayesian inference, can be visualized for human judgment and analysis, with the reasoning method and process also explained. For a few unpredictable cases, because historical related cases are missing from the training data, or the cases with a high level of deviation, the user is always informed or has an opportunity of making further investigation. Therefore the "white box" data mining models help to make informed decisions in the prediction of equipment residual value.

- The data mining model can be deployed in the equipment information management system with an automated process of modeling and updating.

Using the predictive modeling for the wheel loader residual value as an example, this chapter explains how a typical data mining algorithm, the AutoRegressive Tree, infers knowledge from data. The entire process of data mining from data preparation, model generation, and model validation is illustrated by using the ART model of wheel loaders. Multi-fold cross-validation and boxplots are used for testing the stability and accuracy of the generated data mining model. The chapter also demonstrates the advantages of data mining enabled applications in construction equipment management by deploying the predictive models of equipment residual value in an equipment information management system.

In summary, predictive data mining provides a more accurate, flexible, and interpretable approach for assessing the residual value of heavy construction equipment. Using the embedded predictive modules for equipment residual value, the sellers can determine the best time to sell their machines, the buyers can determine the best time to purchase their required machines, and the equipment owners can perform life cycle analysis on equipment to make decisions on equipment repair, overhaul, disposal and replacement. Other data mining applications in construction equipment management, including outlier mining for problem identification [Fan et al. 2007], and time series forecasting for budget planning are also underway in this research. The data mining captures the complexity and dynamics of construction equipment management by making inferences out of data, which could not be realized by using conjectured mathematical or statistical models. The application of data mining in construction equipment management makes it possible for the management team to gain insight into the large amounts of data collected in construction equipment operations and management, and to make proactive decisions.

REFERENCES

Alexander, W. P., and Grimshaw, S. D. (1996). "Treed regression." *Journal of Computational and Graphical Statistics*, American Statistical Association, 5(2), 156-175.

Anderson, J. A. (1995). *An introduction to neural networks.* The MIT Press, Cambridge, MA.

Arditi, D., and Pulket, T. (2005). "Predicting the outcome of construction litigation using boosted decision trees." *J. Comput. Civ. Eng.,* ASCE, 19(4), 387-393.

ASAE (2003). *Agricultural machinery management data.* American Society of Agricultural Engineers, D497.4, Feb 2003.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*, Wadsworth International Group, Belmont, CA.

Burges, C. J. C. (1998). "A tutorial on Support Vector Machines for pattern recognition." *Data Mining and Knowledge Discovery*, Springer Netherlands, 2, 121-167.

Chickering, D. M., Meek, C., and Rounthwaite R. (2001). "Efficient determination of dynamic split points in a decision tree." *Proc., International Conference on Data Mining 2001*, IEEE, San Jose, CA., 91-98.

Chipman, H. A., George, E. I., and Mcculloch, R. E. (2002). "Bayesian treed models." *Machine Learning*, Springer Netherlands, 48, 299-320.

Cross, T. L. and Perry G. M.(1995). "Depreciation patterns for agricultural machinery." *American Journal of Agricultural Economics*, American Agricultural Economics Association, 77(1), 194-204.

Dumler, T. J., Burton, R. O. and Kastens, T. L. (2000). "Management implications of farm tractor deprecation methods." *American Society of Farm Managers and Rural Appraisers Journal*, 63(1), 3-10.

Fan, H., Kim, H., AbouRizk, S., and Han, S. H. (2007). "Decision support in construction equipment management using a nonparametric outlier mining algorithm." *Expert systems with applications*, Elsevier. In press.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery: an overview." *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA.

Karalic A. (1992). "Linear regression in regression tree leaves." *Proc., ISSEK '92 (International School for Synthesis of Expert Knowledge)*, Bled, Slovenia.

Kim, H., Rauch, A. F., and Haas, C. T. (2004). "Automated quality assessment of stone aggregates based on laser imaging and a Neural Network." *J. Comput. Civ. Eng.,* ASCE, 18(1), 58-64.

Lee, M., Hanna, A. S., and Loh, W. (2004). "Decision tree approach to classify and quantify cumulative impact of change orders on productivity." *J. Comput. Civ. Eng.,* ASCE, 18(2), 132-144.

Lucko, G., Anderson-Cook, C.M., and Vorster, M.C. (2006). "Statistical considerations for predicting residual value of heavy equipment." *J. Constr. Eng. Manage.,* ASCE, 132(7), 723-732.

Lucko, G. and Vorster M.C. (2003). "Predicting the residual value of heavy construction equipment." *Proc., Information Technology 2003 Towards a Vision for Information Technology in Civil Engineering, 4th Joint International Symposium on Information Technology in Civil Engineering*, ASCE, Nashville, Tennessee. Doi:10.1061/40704(2003)49.

Lucko, G., Vorster, M. C., and Anderson-Cook, C. M. (2007). "Unknown element of owning costs-impact of residual value." *J. Constr. Eng. Manage.*, ASCE, 133(1), 3-9.

McNeill, R.C. (1979). "Depreciation of farm tractors in British Columbia." *Canadian Journal of Agricultural Economics*, Canadian Agricultural Economics Society, 27(Feb. 1979), 53–58.

Meek, C., Chickering, D.M. and Heckerman, D. (2002). "Autoregressive tree models for time-series analysis." *Proc., 2nd SIAM International Conference on Data Mining*, Arlington, VA.

Perry, G. M., Bayaner, A., and Nixon, C. J. (1990). "The effect of usage and size on tractor depreciation." *American Journal of Agricultural Economics*, American Agricultural Economics Association, 72(2), 317-325.

Perry, G. M., Glyer, J. D. (1990). "Durable asset depreciation: a reconciliation between hypotheses." *The Review of Economics and Statistics*, The MIT Press, 72(3), 524-529.

Prism Business (2005). "Last Bid auction results." Prism Business Media Inc., <https://www.equipmentwatch.com/Marketing/LB_overview.jsp> (March 27, 2007).

Quinlan, J. R. (1992). "Learning with continuous classes." *Proc., AI'92 (Adams & Sterling, Eds), Singapore,* World Scientific, 343-348.

Quinlan, J. R. (1993). C4.5: *Programs for Machine Learning*. Morgan Kaufmann

Publishers, Mateo, CA.

Reid, D.W. and Bradford.G.L. (1983). "On optimal replacement of farm tractors." *American Journal of Agricultural Economics*, American Agricultural Economics Association, 65(May 1983), 326-31.

Shannon, C. E. (1948). "A mathematical theory of communication." *The Bell System Technical Journal*, 27, 379–423.

Stewart, L. (2006). "Giants 2006." *Construction Equipment, Boston*, 109(9), 47

Unterschultz, J. and Mumey, G. (1996). "Reducing investment risk in tractors and combines with improved terminal asset value forecasts." *Canadian Journal of Agricultural Economics*, Canadian Agricultural Economics Society, 44(1996), 295-309

Vorster, M. C. (2004). “How to estimate market value.” *Construction Equipment. Boston.* 107(6), 64-65.

Wu, J. and Perry G. M. (2004). "Estimating farm equipment depreciation: which functional form is best?" *American Journal of Agricultural Economics*, American Agricultural Economics Association, 86(2), 483-491.

CHAPTER 6. DECISION SUPPORT IN CONSTRUCTION EQUIPMENT MANAGEMENT USING A NONPARAMETRIC OUTLIER MINING ALGORITHM[4]

INTRODUCTION

Large contractors have been steadily increasing their investment in construction equipment to satisfy their needs for grown construction volume in recent years (Steward 2000). In 2004, the Association of Equipment Manufacturers (AEM) recorded an 8.8% boost in the sales of construction equipment worldwide. The rate of increase was forecast at 7.0% by the end of 2005 (AEM 2004). Keeping an equipment fleet in well-maintained and workable conditions and with a high state of availability becomes increasingly important for construction contractors in light of this increased need.

Innovations in construction equipment using cognitive and automation techniques, in conjunction with computerized equipment Information Management System (IMS), have greatly simplified the equipment management process. An onboard computer control system makes it possible to collect data on equipment status and

[4] This chapter is a journal paper co-authored with AbouRizk, S. (Department of Civil and Environmental Engineering, University of Alberta), Kim, H., Han, S. H. (Department of Civil Engineering, Yonsei University, S. Korea). Accepted for publication in *Expert System with Applications*, Elsevier. to appear in Issue 35(4), 2008. The writer of this thesis is the primary contributor and writer of this journal paper, the other authors provided supervisory works.

operations automatically. Equipment IMS and wireless communications allow data to be logged, transmitted, stored and reported timely in an electronic format. However this improved ability of data collection does not imply a proportional increase in the efficiency of equipment management for a large contractor. The company equipment manager finds it difficult to turn these data into actionable information when confronted with large amounts of fleet maintenance, repair, and operational data. Our survey with a local contractor shows that only a small portion of the data (10%~20%) is utilized for decision making when we exclude their contributions through various financial reports.

One pressing issue of the contractor in managing a large fleet is to identify problems in equipment management, at both the operational and corporate levels. The company equipment manager always attaches a higher priority to the question "Has anything gone wrong?" than to "How can we improve?" The answer to the second question should, most probably, be based on the solutions to the first question. The symptoms of deteriorated equipment, improperly used equipment, or other mismanagement practices are predominantly hidden in the collected data. Detecting problematic records can help the company equipment manager to decide what immediate corrective actions should be taken.

Currently, the equipment manager depends on spreadsheets and various equipment reports to perform data analysis, identifying potential problems based on his or her personal experience, which is a tedious and error-prone task. Traditional statistical approaches and canned reports (generated from the equipment database using Structural

Query Language) are not able to reveal some inconsistent yet important information for the following reasons:

- They help to detect out-of-range values, get the statistical results, rank the records based on a criteria of selection, but have difficulty in detecting anomalies of irregular patterns.

- They can handle univariate cases, but cannot detect problematic observations from multidimensional datasets in an efficient manner. In fact, most datasets in problem domains have multiple attributes. In many cases, it is the combination of the attributes that make the data records stand out from the others, instead of a single attribute.

- They do not provide a mechanism of "filtering" so that only potentially interested records are presented for analysis in cases of large datasets.

Hawkins defines an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1980). Outlier detection in statistical science provides a mechanism for differentiating inconsistent observations from the general patterns, or exceptional cases from main data groups. Attempts have been made to detect outliers from both univariate and multivariate datasets. The common methods for univariate outlier detection include the quartile method, least square, as well as various slippage tests; the graphical method is effective for detecting outliers in small two-dimensional datasets. For multidimensional datasets, the outlier detection methods fall into two categories: one approach is to calculate distances (such as Mahalanobis's distance) from each record to their gravity center, where those data points far from the data clouds are classified as outliers; the other is the projection pursuit technique, which converts a multi-variate problem to univariate ones by linearly projecting multivariate datasets in some directions. Yet most multi-variate outlier detection methods in statistics have limited applications due to various presuppositions, such as a probability distribution of data (Fan et al. 2005).

Outlier mining is an important branch of data mining, which is a hybrid of statistics, computer science, artificial intelligence, etc. Because outlier mining techniques eliminate many of the restrictions of statistical approaches, they provide better solutions to multivariate outlier detection when applied to real world datasets. For example, outlier mining algorithms can handle large datasets of high dimensions satisfactorily, irrespective of their data distributions or the presence of multiple clusters.

Among the various data mining techniques, clustering is the closest counterpart of outlier mining. Clustering and outlier mining are two opposite tasks on the same dataset: the former is designed to group the data points into clusters by means of unsupervised learning so that inter-group similarity is maximized and intra-group similarity is minimized, while the latter identifies as outliers the points which deviate significantly from their genuine clusters. Most clustering algorithms treat outliers as noisy data, either eliminating them or minimizing their influence on the clustering results.

One major problem with most of the clustering algorithms has to do with the requirement for domain-dependent parameters, which are usually unknown. Because of the fact that many outlier algorithms are derived from their corresponding clustering algorithms, this well-cited problem is inherited. For example, in k-means or k-medoids distance-based clustering algorithms, the user needs to define the number of clusters before the algorism starts to search and optimize the data clusters. In the distance-based outlier mining algorithm introduced by Knorr and Ng (1998), the values of distance *D* and percentile *p* are required as input parameters. Implementation and tests on a number of datasets in this research proved that the results generated from these outlier mining algorithms are usually sensitive to their input parameters, which can only be justified through trial-and-error with the instillation of domain expert knowledge.

The proposed equipment management system uses the nonparametric outlier mining algorithm proposed by Fan et al. (2006). The rationale behind this algorithm is resolution change. When we scale up or down the data attributes of each record, the data points become farther away from or closer towards each other, and the data clouds either separate or merge based on the hyper-spatial distribution of data points. During this process, we collect the accumulated properties of each data point with respect to its clustering-related behaviors and measure its degree of outlying. The data points can be ranked based on how isolated the point is in relation to its close neighborhood and community; as a result, the outliers of greatest interest in the top list can be detected.

This chapter first conducts a literature review on multi-variate outlier detection techniques in statistical science and data mining, followed by the introduction of the definition of a resolution-based outlier factor and the associated outlier mining algorithm, both suggested by Fan et al. (2006). The outlier mining algorithm is then compared to two popular outlier mining algorithms with respect to both technical and experimental aspects, after which the potential benefits of combining this algorithm with an intelligent equipment analysis system are discussed.

MULTIVARIATE OUTLIER DETECTION

Multivariate outlier detection has great potential in civil engineering applications. For example, it was applied to weight-in-motion truck data in order to automatically detect anomalies such as unlikely vehicles or misclassifications (Raz et al. 2004).

Multivariate outlier detection is regarded as a challenging task in statistics since most statistical methods make assumptions about data distribution. Most approaches need to handle joint distributions, which are regarded as posing practical difficulty in statistics. For example, many methods involve calculating Mahalanobis's distance for each observation using robust estimators of the covariance matrix and the mean vector (Patak 1990; Pena and Prieto 2001). Mahalanobis's distance is defined as:

$$D(X_i) = (X_i - u)^T V^{-1} (X_i - u) \qquad (1)$$

Where V – The covariance matrix between attributes

X_i – The ith observation

u – Mean vector

This definition measures the distance of each point from the gravity center of the dataset, taking the covariance among attributes into account. When the covariance equals 1, the distance becomes the Euclidian distance of the data point from the gravity center of the dataset. The data points located at a considerable distance from the center are classified as outliers.

Other approaches in statistics try to break the multivariate detection problem into a set of univariate detection problems by projecting data attributes into some directions (Gnanadesikan and Kettenring 1972). However, the methods have limited applications because they provide correct solutions only when the outliers are located close to the direction of the principle component (Pena and Prieto 2001).

All of the above-mentioned methods assume the presence of only one data cluster, where each variable of data points conforms to a bell-shaped distribution (for instance, Normal distribution). In practice, there may be multiple clusters in a dataset, and where no distributions can be defined for its variables.

Outlier mining algorithms provide better solutions in handling multi-variate multi-cluster cases. The notion of Distance-Based outliers (abbreviated as DB-outlier) and the DB-outlier mining algorithm were proposed by Knorr and Ng (1998). DB-outlier is defined as follows: "An object *O* in a dataset *T* is a DB(p,D)-outlier if at least a fraction

p of the objects in T lies greater than distance D from O". The authors claim that this notion generalizes the notion of outliers defined in statistical tests for standard distributions. The DB-outlier mining algorithm can detect outliers from large, multidimensional datasets. One problem with this notion is that the outliers are measured across the dataset in global context, with no regarding for the fact that an outlier is deviated from its own genuine cluster and that as such its degree of outlying should be measured against its restricted neighborhood; as a result, the algorithm cannot handle datasets containing clusters of heterogeneous densities.

To overcome this deficiency, Breunig et al. (2000) proposed a definition of Local Outlier Factor (LOF) to measure the degree of outlying for an object with respect to its surrounding neighbors. The definition is density-based, because the LOF value depends on how closely packed the local data points are. Those points deep inside a cluster have an LOF value of approximately 1, while the isolated points have a much higher value. The LOF-outlier mining algorithm can rank the outliers in decreasing order of their LOF values. Other researchers also suggested improvements to the LOF method, such as the Connectivity-based Outlier Factor (COF) proposed by Tang (2002) for cases of low-density patterns in a dataset.

Although the widely used outlier mining algorithms have their advantages, they also pose a number of common problems. First, they require input of unknown parameters, such as distance D and percent p in Distance-based outlier mining algorithm and *Minimum Number of Neighboring Points* in the case of the LOF-based outlier mining algorithm. These parameters are domain-dependent and are usually difficult to determine before hand. Secondly, the detected outliers are regarded as unreliable due to their high sensitivity to the input parameters. Finally, the algorithms cannot satisfactorily handle datasets containing clusters of arbitrary shapes, such as convex or elongated. All of these issues hinder the application to real-life problems of this approach.

RESOLUTION-BASED OUTLIER

TURN*, proposed by Foss and Zaiane (2002), is a nonparametric clustering algorithm which can efficiently cluster a dataset containing clusters of arbitrary shapes. The algorithm works by changing consecutively the resolution of a dataset (i.e. transformation by multiplying by a factor greater than 1 or smaller than 1), and finding a resolution at which clustering achieves optimal results. The algorithm consists of two component algorithms: TURN-RES, which collects clustering results and global statistics at each resolution, and TURN-CUT, which detects the turning point of statistical characteristics in conjunction with the changing of resolutions.

Increasing or decreasing the resolution of a dataset changes its clustering results. According to Foss and Zaiane (2002), this is equivalent to viewing a density plot with a microscope or telescope at a certain magnification. In the case of a scatter plot on a two-dimensional dataset, when one looks closely at the plot, or perhaps with a magnifying glass, one can see every point clearly, in which case every point can be considered a cluster; but if we view the same density plot from a far distance, there appears to be only one cluster.

For outlier detection in a dataset, the same observation holds. When the resolution changes, one fails to identify an outlier when the resolution is very small (multiplied by a factor far smaller than 1), whereas one can identify every point as an outlier when the resolution is very large (multiplied by a factor far greater than 1). When the resolution changes on a dataset and clustering occurs at each resolution, the data points exhibit different behaviors based on their degree of outlying with respect to their close neighborhoods. Therefore, if the properties of a data point with respect to its behavior in clustering are collected during resolution change, its degree of outlying can be measured with an accumulated property value.

Fan et al. (2006) defined the concept of neighborhood and resolution-based outlier factor for classifying data points and measuring the degree of outlying, respectively.

Definition 1 (Neighborhood of Object O) – If an Object O has two nearest neighboring points P_i (i=1,2) along each dimension in *k*-dimensional dataset D, and the distance between P_i (either one of the two neighboring points) and O is less or equal to 1, then P_i is defined as the close neighbor of O; all of the close neighbors of P_i are also classified as the close neighbors of O, and so on. All these connected objects are classified as of the same neighborhood.

The threshold value is taken as 1 in order to measure whether or not two points are close enough to be regarded neighbors. The absolute value of this threshold is not important since the distances between points are only relative measurements during resolution change. The algorithm will first find the maximum resolution S_{max} at which all of the points are far enough from each other as to be non-neighbors, along with the minimum resolution S_{min} at which all of the points are close enough to be neighbors.

The Resolution-Based Outlier Factor (ROF) is defined as to measure the degree of outlying:

Definition 2 (Resolution-Based Outlier Factor (ROF)) – If the resolution of a dataset changes consecutively at a specified ratio between maximum resolution, where all of the points are non-neighbors, and minimum resolution, where all of the points are neighbors, then the resolution-based outlier factor of an object is defined as the accumulated ratio of sizes of clusters containing this object for two consecutive resolutions.

$$ROF = \sum_{r_i=1}^{n} \frac{ClusterSize_{i-1} - 1}{ClusterSize_i} \qquad (2)$$

Where $r_1, r_2 \ldots r_i \ldots r_n$. — Resolution at each step

n —Total number of resolution change steps from S_{max} to S_{min}

CluserSize_{i-1} —Number of objects in the cluster containing object O at the previous resolution

ClusterSize$_i$ —Number of objects in the cluster containing object O at the current resolution

To illustrate how this definition can measure the relative isolativity of each object in a dataset, we use a sample two-dimensional dataset containing three clusters C1, C2 and C3 along with some outliers, as shown in Figure 6-1. First, one should note that, the most isolated objects tend to get smaller ROF values. For outliers in the top list, ROF remains at zero until the object has merged into a cluster. The later it merges, the smaller its ROF value; Secondly, objects with enough neighbors, or those affiliated with large sized clusters (C1) increase their ROF values (approximately equal to 1) faster than the smaller, more isolated clusters (C2, C3). Finally, this definition can measure the degree of outlying for an object against its genuine cluster. This can be explained by comparing O2 with O3 in Figure 6-1. At the previous resolution r_{i-1}, the ROF increase is zero for both points. If O3 is merged simultaneously with a large cluster while O2 is merged with a small one, then ROF increase at r_i is added to ROF for both of them. The increase in O2 is slightly slower than in O3 since it is affiliated with a small cluster; accordingly, O2 is a stronger outlier than O3.

In the ROF formula, the cluster size at previous resolution is deducted by 1 to eliminate the effects of repetitive counting of outliers prior to merging. This will set the ROF of an object to 0 before its merging commences.

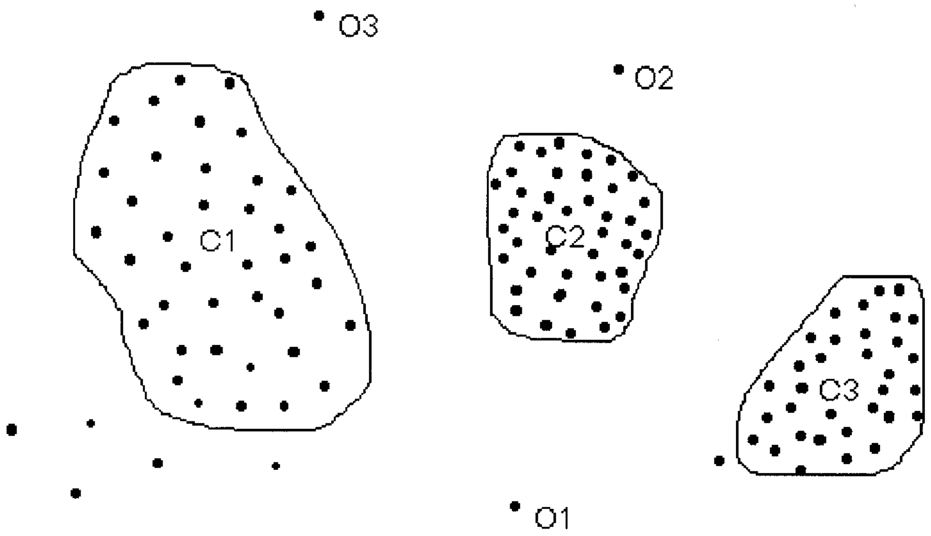


Figure 6- 1. Identification of Resolution-Based Outliers

Finally, one should note that this outlier notion is applicable for multidimensional cases according to the definitions of neighborhood and ROF. A two-dimensional dataset is used in Figure 6-1 merely for its visual intuitiveness.

RB-OUTLIER MINING ALGORITHM

Based on the above definition of RB-outlier, Fan et al. (2006) developed a RB-outlier mining algorithm for mining top outliers based on their ROF.

The algorithm comprises one component algorithm and one overall algorithm: RB-CLUSTER component algorithm handles the merging of data objects at each resolution and collects ROF values; RB-MINE as an overall algorithm steers the mining process by changing resolutions and collecting ROF property values. RB-CLUSTER and RB-MINE are outlined below:

RB-CLUSTER

Given a resolution *r* and a dataset *D* with normalized feature values:

1. Scale the coordinates using current resolution *r*.

   Current Coordinates = Original Coordinates * *r*

2. For each data point, scan all other data points.

   For the data points within a threshold distance of 1, find the closest neighbors in each direction (+/-) along every dimension.

3. Select an unlabeled data point and assign it a new cluster label *C*.

   Initialize its neighborhood chain, *nChain,* and set the size of cluster *C* to 1.

4. Scan this data point's clustered neighbors. For each neighbor:

   If the neighbor is unlabeled, give the neighbor the same cluster label *C*, add the neighbor to *nChain*, and increase the cluster size of *C* by 1.

   If the neighbor has already been labeled as *C'* (where $C' \neq C$), change the label of all points in cluster *C'* to *C*. Increase the cluster size of *C* by the number of points contained in cluster *C'*, then delete records on cluster *C'*.

5. Move to the next data point on *nChain*. Repeat step 4 until all data points on *nChain* are checked.

6. Record the size of cluster *C*.

7. Repeat (3)-(6) until all data points are labeled.

8. For each data point *p*, update the ROF value.

RB-MINE

Given a dataset *D* with normalized feature values and the desired number of outliers *N*

1. Find the maximum resolution, S_{max}, at which no close neighbors can be found for each data point, and the minimum resolution, S_{min}, at which all of the points are close neighbors in the same cluster.
2. Starting at S_{max}, initialize the RB value as 0 for each data point.
3. Update $r_i = r_{i-1} + (S_{max} - S_{min})*\Delta r$, ($\Delta$r is the resolution changing ratio).
4. Run RB-RES to cluster the data at resolution r_i.
5. Update the ROF value for each data point.
6. Rank data points in decreasing order of ROF, obtain top *N* outliers.

The resolution changing ratio must be specified in the algorithm but not considered as an input parameter, since the detected outliers are not sensitive to this parameter. A moderate range of values for this ratio would render satisfactory results for a wide variety of datasets; as such the user need not spend much effort in fine-tuning its value.

COMPARISON OF RB-OUTLIER, DB-OUTLIER AND LOF-OUTLIER MINING ALGORITHMS

Table 6-1 compares RB-outlier, DB-outlier, and LOF-outlier algorithms from the perspectives of outlier definitions, mining algorithms, implementation and application, as well as mining results.

The three outlier mining algorithms are implemented in the same C++ development environment. In order to validate the RB-outlier mining results and compare these with results from DB-outlier and LOF-outlier mining algorithms, experiments are conducted on a number of synthetic datasets. Figure 6-2 shows the top-10 and top-20 outlier mining results on a two-dimensional synthetic dataset containing four distinct clusters and 20 outliers with a total number of 200 points. Analysis of the results indicates that the three algorithms generally detect similar outliers from the same dataset, but with a different list order. Among the top-10 outliers, eight points are unanimously detected as outliers by the three algorithms; all of the 20 outliers are correctly detected in the top-20 list of each algorithm. However, a detailed analysis finds that the commonly identified outliers have different rankings in each list. RB-outlier moves the common outliers to higher rankings in its list, which implies that RB-outlier generates better or equivalent results given a number *N* of top outliers to be scrutinized.

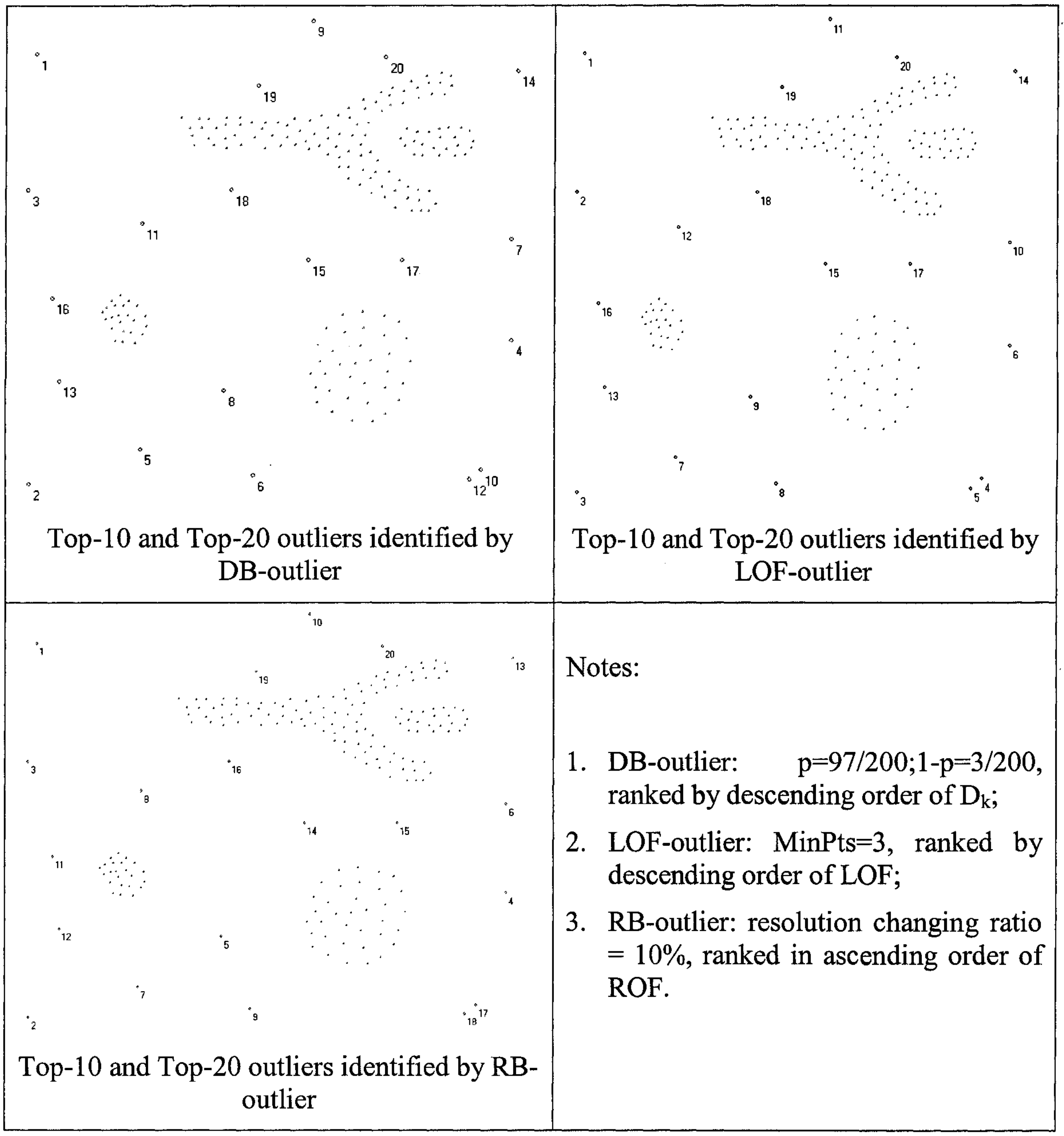
The other two algorithms tend to mark points in small clusters as outliers. For example, in the top-10 lists, the outlier ranked No. 4 in LOF-outlier results is ranked No.

10 in DB-outlier and not included in RB-outlier at all; and the No. 5 outlier in LOF-outlier mining results is not included either in DB-outlier or RB-outlier mining results. Marking data records in a small cluster as outliers and including them in a top *N* list is domain-dependent, and is not referred in most applications.

The results of DB-outlier and LOF-outlier are sensitive to their input parameters. In the case of LOF-outlier, the minimum number of neighboring points defines how large the neighbor is for each data point, and its value has a significant influence on outlier mining results. Figure 6-2 shows only the results from optimized input values for DB-outlier and LOF-outlier. On the other hand, the outlier results are not sensitive to the resolution change ratio defined for the RB-outlier, provided that it falls within a moderate value range between 4% and 20%.

Table 6- 1. Comparison of DB-outlier, LOF-outlier and RB-outlier

	DB-outlier	LOF-outlier	RB-outlier
Outlier notion	Evaluate the degree of outlying of an object by looking for a specified number of nearest objects	Measure how an object has deviated from its "best-guess" cluster	Measure how the object has deviated from its neighborhood with consideration to the surrounding community (reachable neighborhoods)
Outlier mining scheme	Search the specified number of nearest objects to each object	Search the nearest objects and calculate the "local reachability density" of its neighborhood and LOF for each object	Change resolution of the dataset and collect properties of each object with respect to its clustering behavior at each changed resolution.
Implementation and application	Easy to implement, difficult to use	Fair to implement, fair to use.	Easy to implement, easy to use
Outlier mining results	Best suited for dataset of a single cluster. Some local outliers are missed in case of multiple clusters	Good for dataset of multiple clusters with different densities and shapes. Good identification of local outliers.	Good for datasets with multiple clusters of different densities and shapes. Good identification of local outliers with consideration of some global features of a dataset. Satisfactory ranking of the top-listed outliers.



Top-10 and Top-20 outliers identified by DB-outlier

Top-10 and Top-20 outliers identified by LOF-outlier

Top-10 and Top-20 outliers identified by RB-outlier

Notes:

1. DB-outlier: p=97/200;1-p=3/200, ranked by descending order of D_k;
2. LOF-outlier: MinPts=3, ranked by descending order of LOF;
3. RB-outlier: resolution changing ratio = 10%, ranked in ascending order of ROF.

Figure 6- 2. Comparison of Results from DB-outlier, LOF-outlier, and RB-outlier on a 200-tuple Synthetic Dataset

TEST RESULTS ON A THREE-DIMENSIONAL EQUIPMENT DATASET

The three algorithms (RB-outlier, DB-outlier and LOF-outlier mining algorithms) were implemented in the decision support system for construction equipment management. In this system, we run the three algorithms on a construction equipment dataset obtained from a collaborating contractor. The dataset includes 1033 pieces of equipment from the contractor's equipment fleet. In addition to the equipment identification attribute, each unit is described by three other attributes: yearly repair/maintenance cost (Yearly cost), rate of charge, and age, as described below:

- *Yearly repair/maintenance cost (in CAN$ per year)* – This attribute measures the total cost of the repair/maintenance for each unit, divided by the number of years in service. This attribute, as one of the important factors influencing Total Cost of Ownership (TCO), characterizes equipment classes, technical parameters, make and model, usage, etc.

- *Rate of charge (in CAN$ per hour)* – This is the rate (updated on a yearly basis) set by the contractor based on equipment costs in the previous year, current market conditions, etc. This attribute is also a comprehensive measure of equipment value in the market.

- *Age (in year*s) – The annual equipment cost to the owner is also loosely related to its age. The old equipment typically involves higher repair/maintenance (R/M) costs, lower rate of charge, etc.

Figure 6-3 visualizes the dataset in a three-dimensional space where different equipment groups appear as clusters. The purpose of the problem is to detect the irregular equipment of special interests to the equipment manager, and to sort out the equipment based on its degree of "interestingness"; this will identify such equipment as units with irregular combinations of the three attributes as outliers in the top *N* list, or units with exceptionally high or low charge rates or yearly costs.

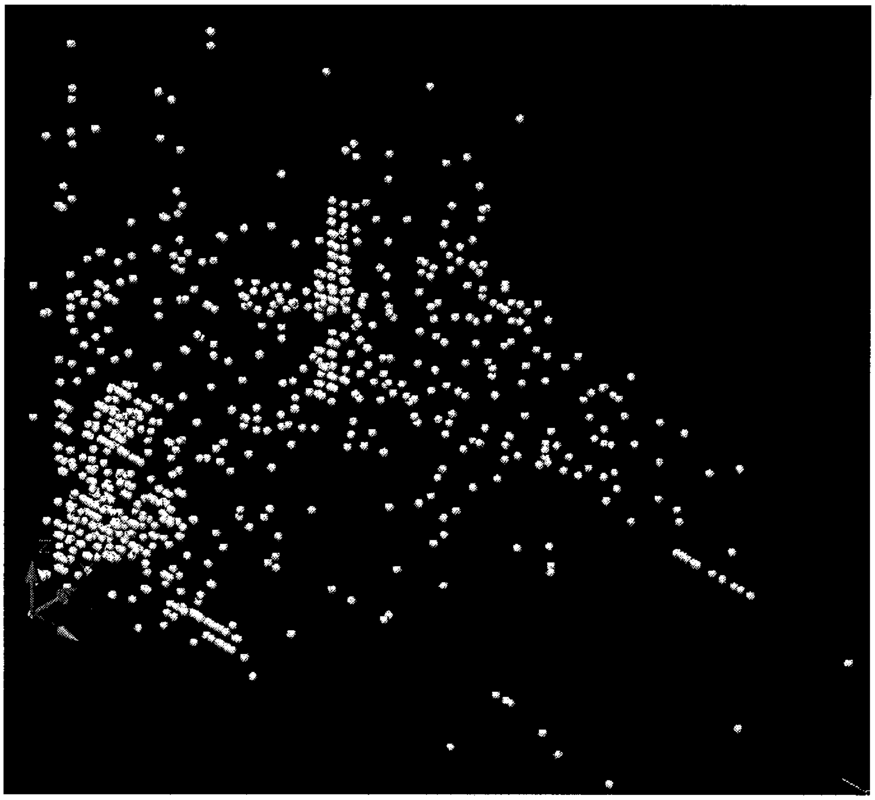


Figure 6- 3. Visualization of the Three-dimensional Equipment Dataset
(x: Cost in CAN$ per year; y: Rental rate in CAN$ per hour; z: Age in years)

In order to compare the outlier results detected by DB-outlier, LOF-outlier and RB-outlier mining algorithms, the three sets of top-20 outliers from the equipment dataset are summarized in Table 6-2.

An analysis of the results shows that 11 out of 20 are unanimously detected as outliers by the three algorithms in their top-20 lists. LOF-outlier generates very similar results to DB-outlier, with 16 of the top-20 being the same. This is not surprising since a large number of isolated objects appear in this dataset. The LOF definition becomes similar to the DB-outlier notion for a sparsely distributed dataset.

RB-outlier moves all of the 11 common units to higher rankings and adds six new units in its top-20 outlier list. Analysis of the six added units against their individual equipment groups and the entire fleet confirmed their interestingness. For example, for unit #138-405, a heavy-duty concrete float, a slice of data finds that its yearly cost is significantly higher than other similar units at comparable rental rates and ages.

Table 6- 2. Top-20 Outliers Identified by DB-outlier, LOF-outlier and RB-outlier Algorithms

No.	**Top-20 outliers identified by DB-outlier**	**Top-20 outliers identified by LOF-outlier**	**Top-20 outliers identified by RB-outlier**
1	**505-401: : Soil Cement Plant (300 to 600 TPH)**	120-446: Dump truck of deck (10 wheels)	120-446: Dump truck of deck (10 wheels)
2	120-446: : Dump truck of deck (10 wheels)	**505-401: Soil Cement Plant (300 to 600 TPH)**	120-445: Dump truck of deck (10 wheels)
3	240-403: : Graders (150 to 225 hp)	120-445: Dump truck of deck (10 wheels)	240-403: Graders (150 to 225 hp)
4	120-793: : Dump truck of deck (10 wheels)	111-570: Distributor Truck	120-402: Dump truck of deck (10 wheels)
5	120-445: : Dump truck of deck (10 wheels)	222-402: Wheel loaders	120-793: Dump truck of deck (10 wheels)
6	222-402: : Wheel loaders	240-403: Graders (150 to 225 hp)	***138-405: Floats***
7	111-570: : Distributor Truck	240-402: Graders (150 to 225 hp)	411-561: Screeners more than 75ton/hour
8	254-700: : Slurry Machine	120-793: Dump truck of deck (10 wheels)	120-748: Dump truck of deck (10 wheels)
9	240-402: : Graders (150 to 225 hp)	**202-401: Tire compactor**	222-402: Wheel loaders
10	120-402: : Dump truck of deck (10 wheels)	**120-438: Dump truck of deck (10 wheels)**	240-402: Graders (150 to 225 hp)
11	**256-406: : Asphalt pavers**	**256-406: Asphalt pavers**	254-700: Slurry Machine
12	120-748: : Dump truck of deck (10 wheels)	**810-328: Snow removal equip**	***223-402: Wheel loaders***
13	411-561: : Screeners more than 75ton/hour	254-700: Slurry Machine	240-706: Graders (150 to 225 hp)
14	**202-401: : Tire compactor**	120-402: Dump truck of deck (10 wheels)	120-456: Dump truck of deck (10 wheels)
15	**120-438: : Dump truck of deck (10 wheels)**	120-456: Dump truck of deck (10 wheels)	***216-405: Rented Equipment***
16	***505-405: : Portable Conveyor Batch***	411-561: Screeners more than 75ton/hour	111-570: Distributor Truck
17	***505-575: : Soil cement***	***120-736: Dump truck of deck (10 wheels)***	***120-737: Dump truck of deck (10 wheels)***
18	**810-328: : Snow removal equip**	***117-700: Truck (6 Wheels) with Drill***	***239-411: Graders (125 to 150 hp)***
19	**115-703: : Dump truck or deck (6 wheels)** *	240-706: Graders (150 to 225 hp)	***240-501: Graders (150 to 225 hp)***
20	***120-428: : Dump truck of deck (10 wheels)***	120-748: Dump truck of deck (10 wheels)	**115-703: Dump truck or deck (6 wheels)** *

Notes:

1. The shaded cells contain units commonly identified by the three algorithms
2. The straight underlined units are commonly identified by DB-outlier and LOF-outlier
3. The squiggly underlined units are commonly identified by LOF-outlier and RB-outlier
4. Units with an asterisk (*) are commonly identified by DB-outlier and RB-outlier

The inability of DB-outlier and LOF-outlier mining algorithms to identify some outliers can be illustrated by their outlier notions. These two algorithms evaluate the degree of outlying by drawing a hyper-sphere around each object, where the number of objects inside the hyper-sphere influences the outlier measurement of the object. The outliers harbored inside the concave portion of a cluster cannot be identified efficiently based on this rationale; subsequently, some outliers are missing in the results if the two algorithms are applied on real-life dataset, which may contain clusters of any arbitrary shapes.

Another problem with DB-outlier and LOF-outlier mining algorithms is their inclusion of isolated units, or units in small clusters, as outliers in their top list: for example, Nos. 1, 16 and 17 outliers in the DB-outlier list and No. 2 in the LOF-outlier list are the only units in their individual equipment classes. No. 18 outlier in the LOF-outlier list designates an equipment class comprising only two units. Obviously, these units are not true "outliers" of particular interest to the equipment management. Their appearance degrades the general quality of the top *N* outlier mining results of DB-outlier and LOF-outlier in real-world applications. On the contrary, all eleven common outlying units and the six additional units unique to the RB-outlier list are inconsistent in relation to their individual equipment groups, and may be considered "true" outliers.

The experiments confirm that the RB-outlier algorithm generates results better than or equivalent to those of the DB-outlier and LOF-outlier mining algorithms on real-world datasets, if we look at the same number of top-listed outliers.

APPLICATION ON CONSTRUCTION EQUIPMENT MANAGEMENT

The resolution-based outlier notion and the nonparametric outlier mining algorithm will potentially benefit construction equipment management by providing flexible solutions to various outlier detection problems. As stated earlier, the company equipment manager is interested in anomalies occurring during fleet operations and management. The current subjective approaches of outlier detection utilizing statistical tools and customized equipment reports are tedious and error-prone. Neither multivariate outlier detection techniques in statistics nor outlier mining algorithms in computing science can provide satisfactory solutions to this problem. In addition, the currently available outlier mining algorithms have difficulty in integrating practically with the equipment information management system due to their domain-dependent input parameters. The RB-outlier mining algorithm used in this research has some major advantages over the current algorithms when applied to equipment management. Some of these advantages are listed below:

- The algorithm needs no parameters of input. This is the biggest advantage when applied to real-life problems. The elimination of input parameters makes it possible to run the algorithm against any dataset, with this flexibility allowing it to be built into an equipment information management system to work as a universal detector of outliers.

- The algorithm can handle a dataset containing multiple clusters of arbitrary shapes and heterogeneous densities. Though the clusters do appear in a real-world dataset, such as the equipment dataset in Figure 6-3, the clusters in the real-life dataset are not well-formed as in the synthetic dataset. The similar objects in the dataset are naturally grouped together with unknown hyper-shapes; therefore, it is important for an outlier mining algorithm, such as the RB-outlier, to be capable of handling such cases.

- The RB-outlier mining algorithm provides better results than the current algorithms because it accounts for both the "local" and "global" features (neighborhood and community) of a dataset.

Ranking the data records based on their degree of inconsistency in the dataset provides the user with the top listed records of most interest for further analysis. From the perspective of equipment management, the equipment usage records in the top list, and the equipment maintenance and repair cost records in the top list all deserve special attention from the equipment management team. Obviously, this interesting list of exceptional records cannot be generated based on simple ordering by a single attribute or a set of attributes.

Although seemingly inconsistent records in a dataset can be flagged as outliers, not all the outliers in the top list are truly problematic. Some records are identified as outliers merely because they are isolated without similar records in the dataset. The real outliers can only be identified after further investigation into the records in the top list.

The integration of the algorithm into the current equipment information management system makes it possible to perform the outlier analysis on large amounts of equipment operation and management data over a live connection. Running the algorithm against various derived datasets automatically generates short listed records from each dataset in the order of level of interests. This is expected to renovate the current practices of diagnostic analysis in construction equipment management.

CONCLUSIONS AND FUTURE RESEARCH

This chapter presented the decision support system for construction equipment management that used the non-parametric RB-outlier mining algorithm. The nonparametric outlier mining algorithm demonstrated ease of use, high flexibility, and satisfactory performance as compared with currently popular outlier mining algorithms. The idea behind the notion is resolution change, similar to the non-parametric clustering algorithm TURN*. Through implementation and tests on both synthetic datasets and a construction equipment dataset, it was found that the RB-outlier factor can measure the degree of outlying of an object based on its isolation relative to its neighborhood and community. The RB-outlier mining algorithm is capable of identifying top listed outliers from a dataset containing clusters of arbitrary shapes and heterogeneous densities. The application of the outlier mining algorithm in equipment management can assist the detection of anomalies by ranking the records based on their degree of ''inconsistency'', subsequently saving significant amounts of time and effort spent on the diagnostic analysis of equipment management.

REFERENCES

AEM (2004). “2004-2005 outlook for construction equipment business.” Association of Equipment Manufacturers, Milwaukee, WI.

Breunig, M.,Kriegel, H., Ng, R., and Sander, J. (2000). “LOF: identifying density-based local outliers.” *Proc., ACM SIGMOD 2000 International Conference on Management of Data*, Dallas, TX.

Foss A., and Zaiane, Z. (2002). “A Parameterless method for efficiently discovering clusters of arbitrary shape in large datasets.” *Proc., International Conference on Data Mining*, IEEE, Maebashi City, Japan

Gnanadesikan, R., and Kettenring, J. R. (1972). “Robust estimates, residuals, and outlier detection with multi-response data.” *Biometrics Journal,* International Biometric Society, 28, 81-124.

Fan, H., Kim, H., & AbouRizk, S. (2005). “A non-parametric outlier mining algorithm for detecting anomalies in construction equipment database.” *Proc., 6th Construction Specialty Conference* . CSCE, Toronto, ON.

Fan, H., Zaiane, O. R., Foss, A., & Wu, J. (2006). “A nonparametric outlier detection for effectively discovering top-N outliers from engineering data.” *Proc., 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 557–566.

Hawkins, D. (1980). *Identification of outliers.* Chapman and Hall, London, U.K.

Knorr, E., and Ng, R. (1998). "Algorithms for mining distance-based outliers in large datasets." *Proc., 24th International Conference on Very Large Databases*, New York, NY.

Patak, Z. (1990). *Robust principal component analysis via project pursuit.* M. Sc. Thesis, University of British Columbia, BC.

Pena, D., and Prieto, F. (2001). "Multivariate outlier detection and robust covariance matrix estimation." *Technometrics*, American Statistical Association and the American Society for Quality, 43(3).

Raz, O., Buchheit, R., Shaw, M., Koopman, P., and Faloutsos, C. (2004). "Detecting semantic anomalies in truck weigh-in-motion traffic data using data mining." *J. Comput. Civ. Eng.,* ASCE, 18(4), 291~300.

Stewart, L. (2000). "Giants replace machines to control costs." *Construction Equipment: Lincolnwood*, 102(3), 62.

Tang, J.,Chen, Z., Fu A., and Cheung, D. (2002). "Enhancing effectiveness of outlier detections for low density patterns." *Proc., 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Taipei, Taiwan, 535 – 548.

CHAPTER 7. CONCLUSIONS AND FUTURE RESEARCH

SUMMARY OF THE THESIS

Construction equipment management plays an important role in heavy construction projects due to its increasing reliance on modern equipment. Heavy construction contractors face fiercer competition in the market and have to make the correct and timely decisions on construction equipment management at both the operational level and corporate level to maintain their competitive edge. Some recent technological innovations have simplified the practice of construction equipment management in various aspects. While the fundamental mechanical design and operations of heavy construction equipment remain unchanged, advanced technologies in sensing techniques, computer control, and management have been widely adopted in heavy equipment to control and monitor equipment field operations. While the traditional principles in decision making in construction equipment management remain the same, the application of communication and information technologies have made it a reality that equipment data can be potentially transmitted to anywhere, at anytime, and used to generate displays and reports to support daily equipment operations and decision making. In the meantime, some new challenges have emerged with these changes. For instance, contractors are capable of collecting an entire suite of equipment data, but cannot interpret the data to obtain needed information and knowledge. Current equipment management solutions fail to provide analytical tools for decision makers to retrieve as-needed information in a flexible and interactive approach and fail to enable the exploitation of the accumulated equipment data for knowledge generation and fact-based decision support.

In conjunction with a large heavy equipment contractor in Alberta, Canada, this research addresses the decision support problem in equipment management from the following four perspectives: (1) how to employ cutting edge information technologies to realize intelligent construction equipment management based on the current information infrastructure of the heavy construction contractors; (2) how to design and implement an "equipment data warehouse" to support flexible information retrieval for construction equipment management; (3) how to design and integrate intelligent data mining modules in the current information system in order to automate the processes of knowledge generation, validation, and utilization; (4) design and implement a novel outlier mining algorithm to address the specific needs of outlier detection from equipment data.

A Framework for Intelligent Construction Equipment Management

A system framework for intelligent construction equipment management was proposed in this research to enhance the decision support functions in the current equipment information management system. After identifying the problems with the current system design in decision support and analyzing the needs of the contractor for an integrated platform for decision support, the research proposed a three-layer architecture for intelligent system design and implementation, with each layer addressing different aspects of the intelligent system. The first is the Data layer, in addition to the relational databases, an equipment data warehouse was proposed to integrate disparate equipment data into a centralized data source. Although both the current equipment data and the data in the separately built equipment data warehouse can be used as the data sources in the data layer of the system, the data in the data warehouse is more suitable for information retrieval and knowledge discovery due to the multidimensional data structure, and high data quality. The second layer is the Application layer; the knowledge repository in the application layer includes both explicit knowledge and implicit knowledge. In addition to the explicit knowledge, which is known to the decision makers a prior and can be programmed into the system for intelligent support, some implicit knowledge can be uncovered and included in the knowledge base by using the integrated data mining process. The third layer of the system is the presentation layer in which the current information system was enriched by improved man-machine interaction for information retrieval and knowledge exploitation through visual interactive data analysis, event notification, predictive analysis, knowledge visualization and the explanation of computer inference. The proposed intelligent framework was exemplified using a real life decision support application in equipment management (i.e. automated evaluation of work orders on equipment repair).

Data Warehousing for Construction Equipment Management

A concept of an "equipment data warehouse" is proposed in this research and implemented in a prototype system to illustrate the methodology, benefits, opportunities, and challenges in building an equipment data warehouse in order to support construction equipment management. First of all, the research addressed the problem of the high level design of a useful, multi-subject data warehouse using the bus architecture proposed by Kimball (2002). The data warehouse design based on bus architecture ensures conformed and shared dimensions in the data warehouse. Secondly, the data warehousing process including the identification of data sources; data Extraction, Transformation/Translation and Loading (ETL); as well as data presentation are explained for the case of equipment management in Standard General, Inc. Equipment data at the obsolete Microsoft Access database, and the data at the current Microsoft SQL Server database are integrated into an equipment data warehouse after an ETL process, and then the equipment data warehouse is presented to the decision makers through a web-based interface which delivers the data analysis functions of drill-down, roll-up, slide and dice, and pivoting in interactive visual formats. Finally, the research analyzed the reason why an equipment data warehousing system is a better solution than the current equipment operation system for decision support in equipment management. The other benefits of implementing an equipment data warehouse, including explorative data analysis using Online Analytical Processing, and mining warehouse data for knowledge discovery are also presented in this thesis.

Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model

Accurate assessment of the market value of heavy construction equipment facilitates equipment life cycle cost analysis and equipment repair/replacement decisions.

Current approaches of either rule-of-thumb or statistical regression are error-prone or difficult to apply in an equipment information management system. This research proposed a data mining based method to automate the residual value prediction process. Based on the real life data from a construction equipment auctioneer, predictive data mining models are built up using the AutoRegressive Tree (ART) data mining algorithm proposed by Meek et al. (2002). The equipment category of wheel loaders is selected to exemplify the training and validation of a data mining model using 10-fold cross-validation tests and box plots. Comparison of the prediction results from the ART model and these from the Artificial Neural Network (ANN) model and the Multivariate Linear Regression (MLR) model proves that the ART model is more accurate and efficient in prediction and is also easy to interpret. After presenting and analyzing the entire data mining process from data selection and preprocessing, to model training, validation, and evaluation, the research explored how to integrate the data mining into the current equipment information management system for automated decision support. The thesis concluded that a data mining enabled equipment information management system provides unprecedented benefits for decision support in the following perspectives: the automated model generation from data, the visualization of data mining model for human interpretation and analysis, the informed process of data mining-based inference, as well as the real-time model updating to reflect recent changes.

A Nonparametric Outlier Mining Algorithm for Detecting Anomalies from Equipment Data

In the outlier detection problem domain, neither traditional statistical methods nor current outlier mining algorithms address the specific features of equipment data, such as a lack of prior knowledge on the dataset, or a dataset containing multiple clusters of arbitrary shapes. A novel non-parametric outlier mining algorithm is proposed in this research to address these challenges and detect top-N outliers from large amounts of equipment data. First of all, based on the concept of "resolution change", a definition of a Resolution-based Outlier Factor (ROF) is proposed to measure the degree of outlying giving consideration to both local and global features of the dataset. Secondly, based on the definition of the ROF, a nonparametric RB-outlier mining algorithm was proposed and implemented using C++. The RB-outlier mining algorithm was proven to be effective and efficient after a large number of experimental tests conducted on the computer using both synthetic datasets and real-world datasets. Finally, the proposed RB-outlier mining algorithm was applied in a construction equipment dataset to detect top-N outliers to identify equipment of unusual behavior. For the purpose of comparison, the other two well-known algorithms DB-outlier (Knorr and Ng 1998) and LOF-outlier (Breunig et al 2000) are applied to the same problem, and compared with the proposed RB-outlier mining algorithm. Based on the experimental results and analysis, the research concluded that the proposed RB-outlier mining algorithm is more suitable for mining top-N outliers from equipment data due to its unique features, such as no need for input parameters, being capable of handling datasets of wide variety, and ranking the mining results more

accurately in the top-N list. The proposed outlier mining algorithms can also be applied in other engineering disciplines for anomaly detection or data preprocessing.

CONCLUSIONS

This research problem was formulated to address the information needs of the contractor by leveraging the large amounts of operational data in construction equipment management. As a collaborating research project between the industry and academia, this research contributed to the body of knowledge in computerized construction equipment management, and to the best practices of decision support in construction equipment management. The proposed system architecture, methodologies, solutions, and algorithms can be applied to other management tasks of project construction or other engineering disciplines as well.

This research is intended to pave the way to transfer data warehousing and data mining technologies to construction industry. While the overall framework and methodologies are justified in this research, there are weaknesses and limitations that needs special attention in similar research works in the future:

- Data warehousing and data mining are only demonstrated in the prototype system. Because of time constraints and limited resources, the research results are not transferred to a fully functional system; this is considered to a weakness of this research;

- Selection of data mining algorithms is addressed in this research but not to a full extent: as a relative new science of data mining, there are large number of data mining algorithms available for each category of data mining task, there is no "best" algorithms for each task because each algorithm has its own strengths and limitations. More extensive evaluation of different data mining algorithms is necessary in mining construction data;

- Data quality issue in data mining is not fully addressed in this research. The data quality in data mining cannot be emphasized enough in mining construction data, this issue is only partially addressed in chapter 5 because the case used in this research happens to have a better quality. In typical data mining applications, data quality may become mission critical and needs significant amounts of efforts.

RECOMMENDATIONS FOR FUTURE RESEARCH

The following topics are recommended for future research:

Time Series Analysis for Financial Planning and Inventory Control

Time series data is a sequence of observations taken at equal intervals of time. The purpose of collecting and modeling time series data is to interpret the change patterns in data and to forecast the future values. Time series analysis is a commonly encountered problem within construction equipment management. For example, the fleet repair and maintenance costs at different levels or at different time intervals are an important series of data for budgetary analysis and planning. Also, the inventory of fluids and parts over time are time series data which are used for optimizing inventory levels and reducing inventory costs.

Time series analysis is an independent research subject in statistical science. Time series data are analyzed traditionally by being broken down into four components (Han and Kamber 2006): (1) Trend movement, which is the general direction in which a time-series is moving over a long interval of time; (2) Cyclic variations, being the long-term oscillations of the time series about trend; (3) Seasonal variations, which are the seasonal movements of the time series; (4) Irregular variations, which are variations of the time series due to random shocks. Both the simple approach, of decomposing the time series into components, and the popular AutoRegressive Integrated Moving Average (ARIMA) method (Bowerman and O'Connell 1993) , also called Box-Jenkins method, are iterative trial-and-error processes which involve subjective personal judgments. No research and development in applying time series analysis techniques in construction equipment management have been reported due to the subjective nature of the methodology and the need for substantial statistical backgrounds. An interview with the collaborating contractors indicate that the tasks of time series analysis in equipment management are usually performed by using spreadsheets for data summary and then forecasting based on personal experience and judgments.

Time series analysis using data mining techniques makes it easy to model, analyze, and forecast time series data due to the ease of use, satisfactory flexibility, and its accuracy. For example, the Microsoft Time Series (MTS) algorithm in Analysis Services of SQL Server 2005 (Microsoft 2007) is a time series mining algorithm which has greatly simplified time series analysis and forecasting. The MTS algorithm is based on the traditional assumption of AutoRegression in the time series forecasting model, which assumes the current value of a time series depends on its previous N observed values (Box et al. 2005), modeled by the following Yule-Walker equation:

$$X_t = f(X_{t-1}, X_{t-2}, \ldots\ldots, X_{t-n}) + \varepsilon_t \tag{1}$$

Where X_t — Current observation

X_{t-i} — Previous n observations, i=1,2,…,N

ε_t— Current observation

In order to train a data series model from data, the MTS algorithm first converts a time series data into cases with the current observation as a dependent variable and previous n observations as predictor variables; it then applies the AutoRegressive Tree (Meek et al 2002) algorithm to establish the logical and mathematical relationship between the current observation value and its previous n observations. To account for the seasonality features of a time series, the data mining algorithm uses Fast Fourier Transform (FFT) to search all the seasonality features automatically or based on the known seasonality parameter from user input.

The time series data mining model built using MTS algorithm has many features such as visualization of the time series model to represent the Yule-Walker equation, automated process, and visualization of results. In addition, this modeling technique can be used to analyze multiple time series in order to identify their relations of changes over time. A preliminary test on mining time series data was conducted in this research through modeling and forecasting two concurrent time series: monthly fleet maintenance costs and fleet repair costs. Figure 7-1 shows the forecasting results of both time series for the next three months.

Characterization Analysis for Decision Support in Equipment Management

Characterization of data using data mining techniques summarizes and describes a specific group of data to help the decision makers identify the interesting common features of the target data group for decision support. For examples, the decision makers in equipment management may be interested in characterizing the top 10% heavy construction equipment in the fleet with a low rate of underutilization, or comparing the features of this group of equipment with a contrasting group of the top 10% with a high rate of utilization, with the objective of taking measures to improve the rate of utilization based on the identified features. Other application includes the characterization of equipment with a low rate of return in investment, highly underestimated work orders, etc.

Many data mining algorithms can be used for data characterization, such as a decision tree family of algorithms, association algorithms, and the Naive Bayes algorithm. Compared with human interpretation, the data mining-based approach is much more efficient and accurate.

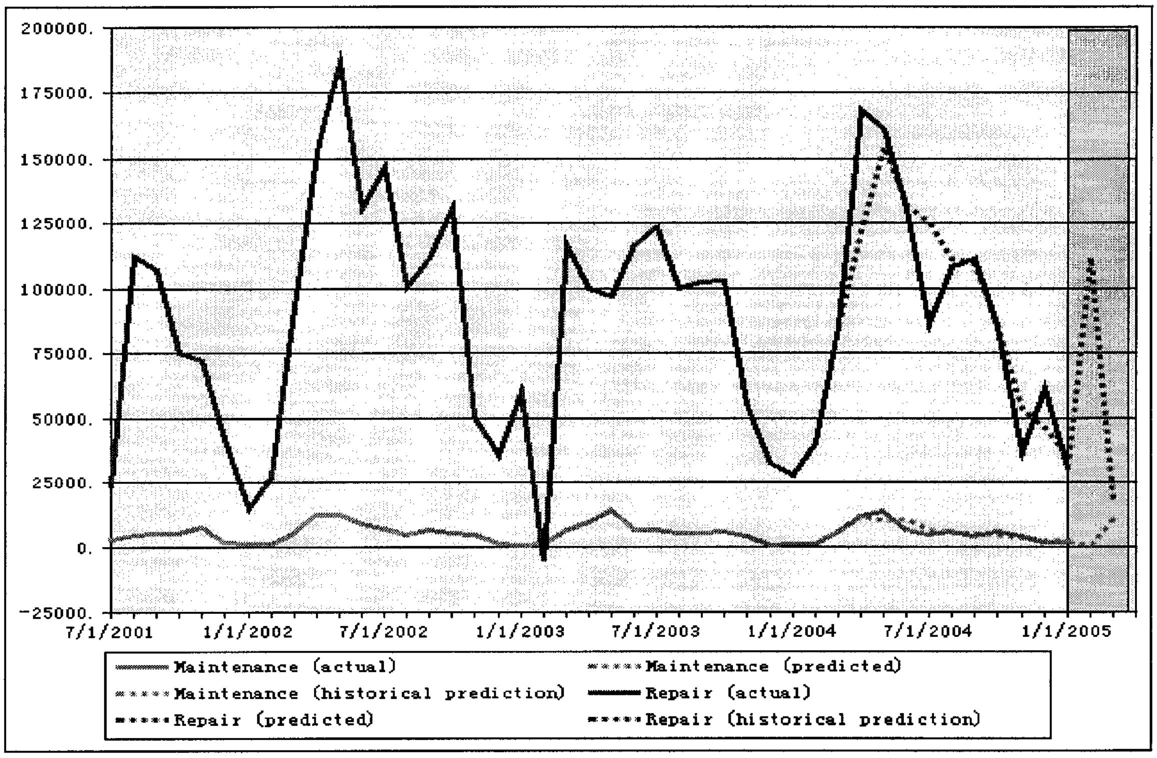


Figure 7- 1. Forecasting Monthly Fleet Maintenance Costs and Fleet Repair Costs Using Time Series Data Mining Model

Mining Simulation Data for Optimization of Equipment Resources

Discrete event simulation provides a powerful experimental tool for the design and analysis of construction operations in the computer. Under controlled experiments, large amounts of simulation data can be generated to describe the system performance under various scenarios. While human interpretation of simulation data is limited to a specific set of parameters, inputs and outputs, analysis of simulation data using data mining techniques helps decision makers gain insight into the system behavior in a wide variety of aspects. The following categories of simulation data can be collected and analyzed using data mining techniques: (1) simulation model descriptive data including model topology, model parameters, and system inputs; (2) simulation results including resource utilization, resource status, and simulation model outputs, (3) trace file which includes time-stamped data tracking of the operational details of the simulation system.

Construction equipment is one of the most important resources in most simulation projects. As a decision support tool in construction management, discrete event simulation can be used to optimize the combination of equipment resources to maximize productivity or minimize costs. Besides the manual approach of optimization (e.g. performing a sensitivity analysis on the simulation results by changing input parameters in simulation model), There is also simulation research work dedicated to the optimization of equipment resources in an automatic or semi-automatic approach, for example, AbouRizk and Shi (1994) proposed an automation system to seek an optimized combination of equipment resources based on the iterative simulation-feedback. In another example, Shi (1998) designed a Neural Network-based system in combination with a simulation modeling tool to minimize unit production cost and predict the system performance under a given crew of combined equipment and labor.

A hybrid approach of stochastic simulation and data mining can be used to tackle the same domain problem. After iteratively running the simulation model by enumerating different resource allocation schemes, all the simulation input and output data are collected and represented in a dataset. Using this dataset, predictive data mining models are generated and validated using data mining algorithms such as the AutoRegressive Tree (Meek et al. 2002) model, or the Classification and Regression Tree (Breiman 1984) model. The derived data mining models can be used for the following decision support tasks:

- Ranking the system inputs as influential factors as per their degree of impact to the simulation results;
- Deciding which possible scenarios lead to a high production output;
- Comparing the different work methods, i.e. influence of different model topologies in productivity, progress, and costs.

The hybrid method for simulation and analysis can be built into an automation system with features of a visualized simulation model, visualized data mining model, and interactive analysis. This method is expected to facilitate the design of construction operations by partially automating the parameter selection process of the simulation model and explicitly representing all the rules leading to different performance results. Therefore, the user can identify the best strategy of equipment resource allocation with minimal efforts.

FINAL REMARKS

As a pilot project, this research proposed the concepts of "intelligent construction equipment management", and "equipment data warehouse", which can be implemented among heavy construction contractors as well as other equipment-owning organizations. The recommended topics for future research cannot cover all the aspects of this emerging research area, as large amounts of research work is still needed in order to facilitate the information retrieval or automate the knowledge discovery in support of equipment management and replace the experience-based decision making tasks with automated processes as much as possible. However, the application framework, principles, best practice, and algorithm proposed in this research are equally applicable to other areas of construction management for intelligent data analysis and decision support.

REFERENCES

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (1994). *Time series analysis: forecasting and control,* 3rd Ed., Prentice Hall, Upper Saddle River, NJ.

Bowerman, B. L., and O'Connell R. T. (1993). *Forecasting and time series: an applied approach,* 3rd Ed., Thomson Learning, Woodbridge, CT.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*, Wadsworth International Group, Belmont, CA.

Breunig, M.,Kriegel, H., Ng, R., and Sander, J. (2000). "LOF: Identifying density-based local outliers." *Proc., ACM SIGMOD 2000 International Conference on Management of Data*, Dalles, TX.

Han, J. and Kamber M. (2006). *Data mining concepts and techniques*, 2nd Ed., Elsevier, Holland

Kimball, R. and Ross, M. (2002). *The data warehouse toolkit: the complete guide to dimensional modeling*, 2nd Ed., John Wiley & Sons, Hoboken, NJ.

Knorr, E., and Ng, R. (1998). "Algorithms for mining distance-based outliers in large datasets." Proc., *24th International Conference on Very Large Databases*, New York, NY.

Meek, C., Chickering, D.M. and Heckerman, D. (2002). "Autoregressive tree models for time-series analysis." *Proc., 2nd SIAM International Conference on Data Mining*, Arlington, VA.

Microsoft (2007). "SQL Server 2005 books online: Microsoft Time Series Algorithm." Microsoft Corp., <http://msdn2.microsoft.com/en-us/library/ms174923.aspx> (May 15, 2007).