Using Data Mining Techniques for Improving Building Life Cycle

Report No. [2001-002-B report no. 3]

The research described in this report was carried out by:

Project Leader: Professor John S. Gero

Project Manager: Dr. Rabee M. Reffat

Team Members: Mr. Wei Peng
Mr Paksan Liew
Dr Julio Rosenblatt (on sick leave)

Research Program No: (2001-002-b)

Program Name: Sustainable Design

Project No.: 2001-002-B

Project Name: Life Cycle Modelling and Design Knowledge Development in Virtual Environment

Date: 1 October 2003

Distribution List

CRCCI
Authors

Disclaimer

The Client makes use of this Report or any information provided by CRC CI in relation to the Consultancy Services at its own risk. CRC CI will not be responsible for the results of any actions taken by the Client or third parties on the basis of the information in this Report or other information provided by CRC CI nor for any errors or omissions that may be contained in this Report. CRC CI expressly disclaims any liability or responsibility to any person in respect of any thing done or omitted to be done by any person in reliance on this Report or any information provided.

Please direct all enquiries to:

Chief Executive Officer
Cooperative Research Centre for Construction Innovation
9th Floor, L Block, QUT
2 George St
Brisbane Qld 4000
AUSTRALIA
T: 61 7 3864 1393
F: 61 7 3864 9151
E: enquiries@construction-innovation.info

Table of Contents

List of Figures

NOTE

This report includes a draft paper to be submitted to **The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining**, Calton Crest Hotel, Sydney, Australia, May 26-28, 2004.

http://www.deakin.edu.au/~pakdd04

Paper Title: **Using Data Mining Techniques for Improving Building Life Cycle**

Authors: **Rabee M. Reffat, John S. Gero and Wei Peng**

1. ABSTRACT

The construction industry has adapted the information technology in its processes in terms of computer aided design and drafting, construction documentation and maintenance. Hence, the data generated within the construction industry has become increasingly overwhelming. The growth of many business, government, and scientific databases has begun to far outpace human's ability to interpret and digest this data. This issue becomes critical with the high degree of complexity of work flow is taken into account in the decision making process during the lifetime of a building. Furthermore, past experience often plays an important role in building management. Therefore, applying data analytic techniques to efficiently deal with information at different stages of a building life cycle has great potentials in this regard. Data mining is a sophisticated data search capability that uses classification algorithms to discover patterns and correlations within a large amount of data.

This paper presents how and what data mining techniques can be applied on maintenance data of buildings. The paper illustrates the results and shows potential benefits of applying such techniques searching for useful patterns of knowledge and correlations within the existing building maintenance data to support the decision making on future maintenance operations

2. INTRODUCTION

The construction industry has adapted the information technology in its processes in terms of computer aided design and drafting, construction documentation and maintenance. Hence, the data generated within the construction industry has become increasingly overwhelming. The growth of many business, government, and scientific databases has begun to far outpace human's ability to interpret and digest this data. This issue becomes critical with the high degree of complexity of work flow is taken into account in the decision making process during the lifetime of a building. Therefore, applying data analytic techniques to efficiently deal with information at different stages of a building life cycle has great potentials in this regard.

However, traditional methods of data analysis such as spreadsheets and ad-hoc queries were not able to fit in because they can only create informative reports from data, but can not analyse the contents of these reports. Hence, there is a significant need for a new generation of techniques and tools with the ability to automatically assist humans in analysing a large amount of data to provide useful knowledge within the construction industry.

The increasing use of databases to store information about facilities, their use and maintenance provides a platform for the use of data mining techniques. Knowledge Discovery in Databases (KDD) and Data Mining (DM) are tools that allow identification of valid, useful, and previously unknown patterns so that building managers can analyse a large amount of project data. These technologies combine techniques from the areas of machine learning, artificial intelligence, pattern recognition, statistics, databases, and visualisation to automatically extract concepts, interrelationships, and patterns of knowledge of interest from large databases.

The work is this paper is motivated by several observations of the current situation in the building industry. The design of new buildings and facilities tends to focus on short-term cost and immediate needs of the building owner to meet a set of business and functional requirements. Current technologies such as Computer-Aided Design (CAD) have focussed on the needs of designers to develop designs without giving much attention to the life cycle of buildings including modelling the life cycle cost of buildings at the design and management stages to forecast and achieve the most economical life cycle. Another observation is that current information technology applied to facility maintenance utilises databases to keep track of information and notification of maintenance schedules. However, so far these databases are not well linked with interactive 3D models of buildings and mostly presented in tabular formats.

Applying data mining techniques to the records of existing facilities has the potential to improve the management and maintenance of existing facilities and future design of new facilities. This will lead to more efficient and effective facilities maintenance and management through better planning based on models developed from available maintenance data, resulting in more economical life cycle of buildings. Furthermore, designers and maintenance managers will be better equipped to achieve high performance by utilising appropriate techniques of information technology at their workplace.

This paper presents how and what data mining techniques can be applied on maintenance data of buildings. The paper introduces in the remainder sections the results and potential benefits of applying such techniques searching for useful patterns of knowledge and correlations within the existing building maintenance data to support the decision making on future maintenance operations.

3. BUILDING MAINTENANCE AND LIFE CYCLE MODELLING

A simple statement of the maintenance objective for a building is that building systems should always be available to support building functions. More precisely, the maintenance objective for the building is that the cost of any maintenance activity should be less than the expected marginal value of production enabled by the planned activity. To support this objective, it is essential to tackle the maintenance from multiple facets including interpretation of observed data, diagnosis of problems, planning repair and maintenance, and business evaluation of the value-added from different repair and maintenance options. More significantly is defining the "value" of the maintenance from both engineering and business perspectives.

Introduction life cycle modelling provides the opportunity to choose the most cost effective approach from a series of alternatives to achieve the least long term cost of ownership (Barringer, 1996). Life cycle cost modelling (LCM) contributes to competitiveness of the company by providing strategic planning on rehabilitation and enhanced information for decision making. LCM also helps facility managers in evaluating alternative equipment and process selection based on total costs rather than the initial purchase price. The multidimensional information of LCM is merged from hybrid project domains such as management, engineering and finance.

There are several life cycle models available for buildings as a whole and for their component systems. However, there is no specific model that has been considered as a standard model. Life cycle cost models form predictions based on several parameters, some of which include a degree of uncertainty, such as the reliability of a part. These inputs can range from the cost of installation to the cost associated with carrying spare parts in inventory (Siewiorek, 1982). Some of the inputs and their values that could be potentially identified include (Dhillon, 1989):

- mean time between failure,
- mean time to repair,
- average materials cost per repair,
- labour cost per corrective maintenance action, and
- labour cost per preventative maintenance action.

The values of these input variables, along with their probability distributions, can be predicted for each component by applying the appropriate Data Mining techniques. This allows for more accurate estimation of average life cycle cost to be determined. By accurately predicting failure rates and repair costs, it will be possible to compute the optimal schedule of preventative maintenance for each building asset. However, unavoidably, what can be predicted and the accuracy of those predictions depends of course on the availability and accuracy of the maintenance data that is available. Furthermore, current life cycle modelling systems fail to provide a seamless integration of hybrid information that provides user access to previously inaccessible knowledge. This paper focuses on selecting and applying the appropriate data mining techniques to improve knowledge acquisition and accessibility of Life Cycle Modelling of buildings.

4. INCORPORATING DATA MINING INTO LIFE CYCLE MODELLING

4.1 Data mining overview and motivations

Past experience often plays an important role in building management. “How often will this asset need repair?” or “How much time is this repair going to take?” are the types of questions that project managers face daily in their planning activities. Failure or success in developing good schedules, budgets and other project management tasks depends on the project manager's ability to obtain reliable information in order to be able to answer these types of questions. Other aspects of building management include improving available scheduling algorithms, estimating spreadsheets and other project management tools. A micro-scale level of research is important in providing the required tools for the project manager's tasks. However, even with the best of such tools, low quality input information will produce inaccurate output of schedules and budget. Thus, it is also important to have a broad approach of research at a macro-scale level.

These days, the architectural, engineering, construction (AEC) industry is witnessing a massive growth in generating and collecting construction and maintenance data. This provides an opportunity and a challenge to appropriately use this data to obtain useful knowledge. However, in most cases, the richness of data might be difficult to achieve due to inadequate analysis. Project managers do not have enough time to analyse the data since the complexity of the data analysis are beyond the capabilities of the relatively simple building maintenance systems commonly used (Soibelman, 2002). Furthermore, there has been no well defined automated mechanism to extract, pre-process and analyse building data and summarise the results to allow site managers to benefit from them.

Data Mining, as an extraction of implicit, previously unknown, and potentially useful information from data (Frawley, 1992), provides useful tools that help to explain how building systems that were once thought to be completely chaotic have predictable patterns (Peitgen, 1992). By applying data mining to identify novel patterns, project managers will be able to build knowledge models that may be used for the recurrent activities of on-going construction projects, and activities of future projects to avoid unanticipated consequences (Soibelman, 2002). Data Mining presents a significant potential for addressing the problem of transforming knowledge implicit in data into explicit knowledge for decision makers.

4.2 Data mining approach for building life cycle modelling

The approach, presented in this paper, is based on a comprehensive view of the building management problem. It views the process of building design, maintenance, and replacement as a process generating an enormous amount of information. While current practices address only parts of this information generation and management, this approach attempts to account for the life cycle flow of this information. The cost of designing and building structures are much smaller than the costs of operating a building or other structure over the course of its life span. Data mining enables a building owner to make important decisions about life cycle cost in advance, thereby significantly affecting and improve design decisions.

The rich set of building data generated or accumulated during the design and documentation phases of buildings remains relevant even after the building is constructed. Building data becomes richer as maintenance data is included and updated regularly. Architects, interiors designers, engineers, contractors, marketing and sales personnel, building managers and owners can extract useful information from the databases for building renovation, maintenance, and operation. Figure illustrates a proposed model of the information flow in building design and maintenance. The bold arrows depict the new functions provided in this model while the dashed arrows describe current approaches to building information management. The integration of data mining within the process of information flow provides the opportunity to make a goo use of building data to feedback and improve the processes of building design and maintenance.

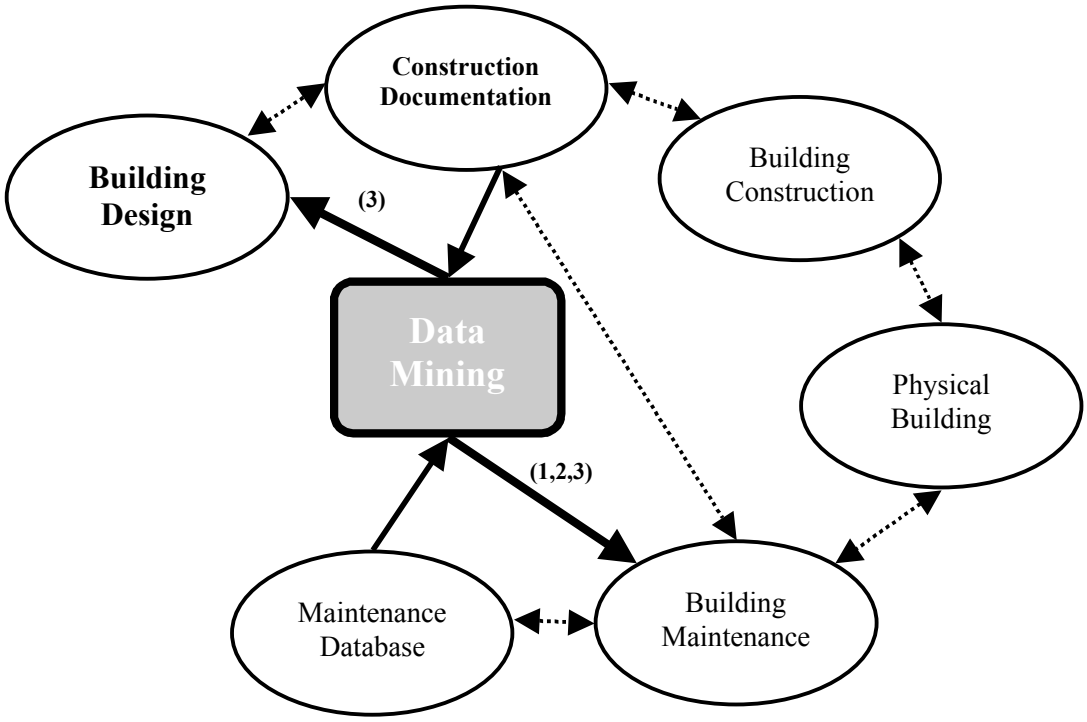

Figure 4.1. Integrating data mining within the life cycle of building information management.

Data mining techniques can be used effectively on data stored in a Building Maintenance System (BMS) by extracting useful knowledge that can be used for future management and design decision making. Knowledge that implicitly resides in BMS databases and corresponds to Figure 4.1 includes:

1. components that frequently need maintenance and therefore need to be inspected carefully
2. historical consequences of maintenance decisions that may inform future decisions
3. components of buildings that significantly determine maintenance cost and therefore may inform future building designs, as well as refurbishment of the building in question

It has been shown in the AEC industry that major factors contributing to construction quality problems include inadequate information and poor communication (Arditi et al., 1998; Burat et al., 1992). The detection of previously undiscovered patterns in BMS data can be used to determine factors such as cost effectiveness and expected failure rate of assorted building materials or equipment in varying environments and circumstances. These factors are important throughout the life cycle of a building, and such information could be used in the design, construction, refurbishment, and maintenance of a building, representing a substantial decrease in cost and increase in reliability. Such knowledge is significant for saving resources in construction projects. In order to examine the feasibility of the proposed approach, a prototype of the data mining system will be developed and tested with the building maintenance databases. Data mining tools can be used to identify the causes of problems such as cost overrun, quality control and assurance. Predictable patterns may be revealed from construction data that were previously thought to be chaotic.

4.3 Data Mining Process

Data mining requires many significant steps from problem specification to the implementation of tools, and monitoring of the model (Feelders et al. 1999). Successful data mining requires several collaborative expertises such as; subject area expertise, data expertise, and data analysis expertise. Data mining is an explorative process since new knowledge is discovered and new hypotheses can be formed. The data mining process for extracting hidden knowledge from large databases can be depicted as shown in Figure 4.2. The process focuses on finding interesting patterns that can be interpreted as useful knowledge and it consists of seven steps (Hui, and Jha, 2000).

- Establishing the mining goals. This involves the understanding of building maintenance process and its acquired database.

- Selection of data. This step identifies a subset of variables or data samples, on which mining can be performed. There are many tables in the database in which not all are suitable for mining since they are not sufficiently rich.
- Data pre-processing. This step aims to remove the noisy, erroneous and incomplete data. The presence of too many different categories of data makes visualisation of the displayed information very difficult. Hence, those categories with only a few records are eliminated. Moreover, all the records with missing values are deleted to avoid potential problems in visualisation. Since the proportion of such records is quite small, their deletion will have little effect on the results.
- Data transformation. The data stored in the various tables are required to be in a specified format. Sometimes, it is useful to transform the data into a new format in order to mine additional information.
- Data warehousing. Data warehousing is the process of visioning, planning, building, using, managing, maintaining and enhancing databases. The data suitable for mining are collected from various tables of customer service database and stored in WEKA's data warehouse. WEKA is a collection of machine learning algorithms for solving real-world data mining problems. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA does not only contain tools for data pre-processing, classification, regression, clustering, association rules, and visualisation, but is also suitable for developing new machine learning schemes.
- Data mining. WEKA is used to perform the data mining functions, including summarisation, association, classification, prediction and clustering.
- Evaluating the mining results. Different data mining functions have been exercised, providing data. The information obtained is next analysed.

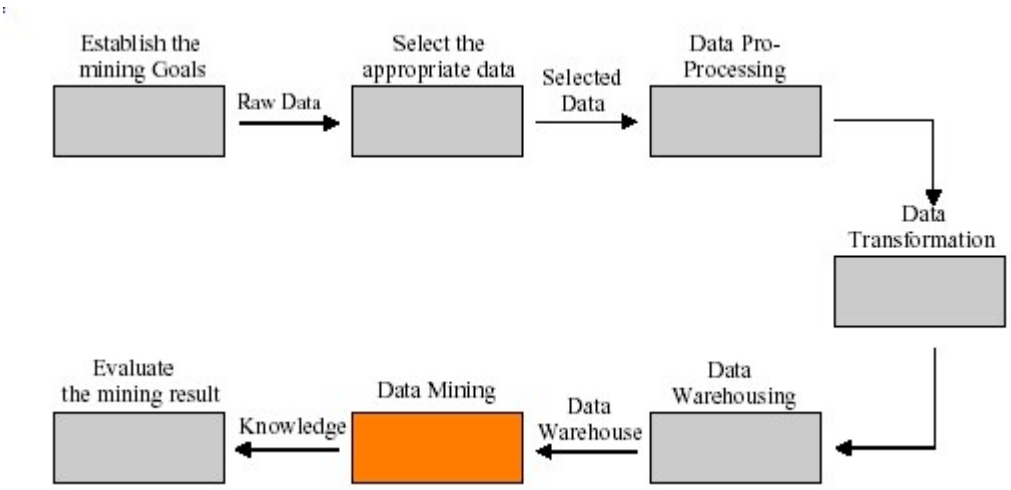

Figure 4.2. Stages of data mining process (Hui, and Jha, 2000).

5. MAINTENANCE DATA

The maintenance data handled in this paper is provided by the Engineering Division of the Central Sydney Area Health Service (CSAHS) for one of their hospital buildings. This building is a five-storey building and referred to here as Building 10. Maintenance data for the last two and a half years is available in SQL format and contains data that is highly detailed and structured. There is an approximately 5000 work orders recorded for Building 10 in the period from 1/1/2001 to 9/12/2002. An example of the available maintenance data of the Air Handling Units at Building 10 is shown in Figure 5.1. The detailed work order records kept since 2001 contain the following fields:

- Work Order Number. Work orders are prefixed by a code as follows:
  - P Preventative Maintenance (PM)
  - Z Corrective Maintenance (CM)
  - R Corrective Maintenance submitted electronically. Corrective Maintenance (Z and R) accounts for approximately 55% of all work orders.
- Description. For CM, this field contains a description of the fault to be repaired. For PM, it refers to one of a set of pre-defined scheduled upkeep tasks.
- Priority Code. Each work order is assigned a priority of Low (L), Medium (M), or High (H), or a number in the range of 0-9. Over half of the work orders in the Building 10 data are assigned a priority of M, with another 20-25% assigned priorities of 5 and 6 (presumably roughly equivalent to M), with the rest being assigned a priority of L or H. Less than 1% of the work orders were assigned to numerical values other than 5 or 6.
- Work Order Status Code. The current disposition of the work order, which is one of: C Completed 92.5%, O Outstanding 5.9% and X Cancelled 1.6%.
- Job Type Code. One of the following: AD Administration / Development, CM Corrective, Maintenance, MAC Major Capital Works, MIC Minor Capital Works, MW Minor Works, PM Planned Maintenance and SM Statutory Maintenance.
- Requested Date/Time. The date and hour at which the CM is requested or the PM is scheduled.
- Date Approved. The date that the work order was authorised.
- Start Date. The date (and sometimes the hour) on which the maintenance action commences. This is the same day as the request date over 99% of the time, but occasionally can be as much as 3 weeks later.
- Completion/Cancel Date/Time. This is the date and time at which the work was either completed or cancelled.
- Estimated Completion Date. An estimated completion date is entered for many work orders, presumably as the order is entered into the system.
- Asset Number. The work orders often indicate an asset number that refers to a specific building element or component. A description is provided for each asset, along with detailed location information in the form of a building code, room code, level code, and department code. Additional information is also offered when available, such as installation date, purchase value, make, model, and serial number. For each asset, there is also an indication of the asset category to which it belongs.
- Task Details. For Preventative Maintenance work orders, a description of the planned task is provided from a list of pre-defined descriptions. A short textual description is given, as well as an extended description. Further information is referenced using the task code.
- Task Code. A number is also provided that refers to the task list, where extensive details about the task is given in a highly structured format, including task frequency, job type code, job sub type code, and priority code.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Line No	Work Orde	Job Typ	Job Sub T	Priority	Cost C	Depar	Floor	Room	Asset No	Task Num	Description	Extra Text	Comments	Work Orde
2	51	P0043720	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
3	52	P0043721	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
4	53	P0043722	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
5	54	P0043723	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
6	55	P0043724	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
7	56	P0043725	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
8	232	P0046282	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
9	233	P0046283	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
10	234	P0046284	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
11	235	P0046285	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
12	236	P0046286	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
13	237	P0046287	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
14	429	P0048765	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
15	430	P0048766	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
16	431	P0048767	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
17	432	P0048768	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
18	433	P0048769	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
19	434	P0048770	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
20	632	P0051193	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
21	633	P0051194	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
22	634	P0051195	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
23	635	P0051196	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
24	636	P0051197	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
25	637	P0051198	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
26	799	P0053370	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
27	800	P0053371	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
28	801	P0053372	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
29	802	P0053373	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
30	803	P0053374	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
31	804	P0053375	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR003	FILTER REF	1. REPLACE	FILTER MEDI	PMS
32	923	P0054595	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
33	924	P0054596	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
34	925	P0054597	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
35	926	P0054598	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
36	927	P0054599	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
37	928	P0054600	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
38	1010	P0055780	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
39	1011	P0055781	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
40	1012	P0055782	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
41	1013	P0055783	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
42	1014	P0055784	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
43	1015	P0055785	PM	FILT	H	0	10	6	MPR2	AHU1001-0	FR001	FILTER REF	1. REPLACE	FILTER MEDI	PMS
44	1890	P0065164	PM	FILT	H	0	10	1	MPR1	AHU1000-0	FR002	FILTER REF	1. REPLACE	FILTER MEDI	PMS

Figure 5.1. An example of the available maintenance data for the Air Handling Units at Building 10, Royal Prince Alfred Hospital, Central Sydney Area Health Service.

6. APPLYING DATA MINING TECHNIQUES ON MAINTENANCE DATA

6.1 Data Mining Using Visual Analysis Approach (Stacked Histogram)

A histogram is defined as a bar graph that shows frequency data. In a histogram, data is collected and sorted into categories. Analysis using histograms is a powerful technique for looking at and processing large amount of data. Histograms focus on the frequencies and distributions of one particular attribute, for example, the priority description for the entire data set as illustrated in Figure 6.1which shows only that Priority attribute. In order to find out correlations between various attributes, there is a need of an interactive visualisation rather than a static view of histogram.

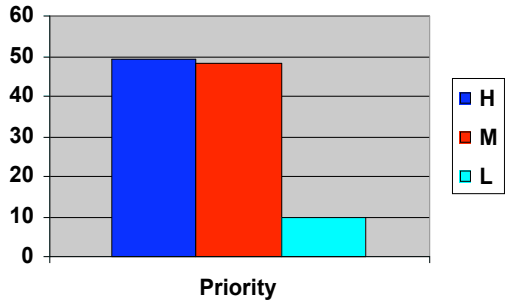

Figure 6.1 A standard histogram of the "priority" attribute

WEKA incorporates a stacked histogram which allows three judgments: (i) the trends on the total height of the columns, (ii) the proportion of each category within each column and (iii) the trends in the lowest category (Dix and Ellis, 1998). This interactive stacked histogram solves the problem of incapability of cross comparison of standard histogram by allowing different trends to be analysed using the same dynamic graph. Thus, the correlation between attribute "priority" and "cause-of-repair" can be visualised as shown in Figure .2. A rule can be learned from this interactive stacked histogram, that is about 94% of A/C (Air/Condition) malfunction belongs to high or medium priority jobs.

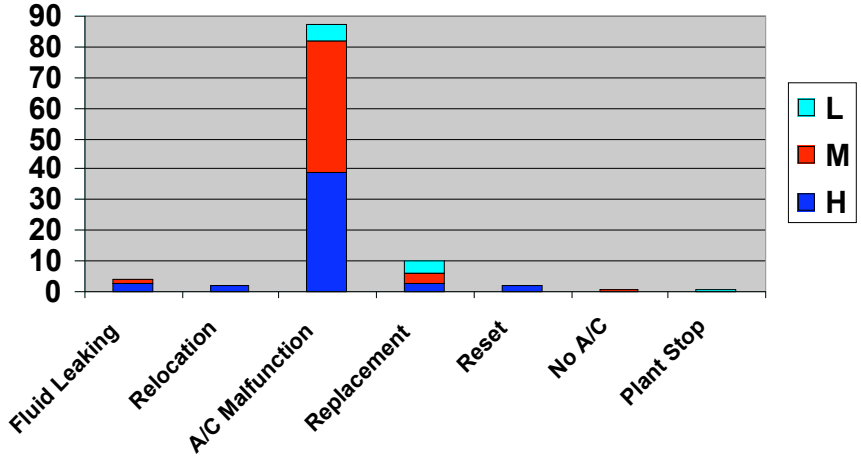

Figure 6.2 A stacked histogram of correlation between "priority" and "cause-of-repair".

A number of rules were generated by analysing the correlations between various attributes using WEKA's stacked histograms function. Stacked histogram is capable of handling both numeric and nominal attributes, but it does not perform well in analysing date and continuous numeric values.

6.2 Data Mining Using the Clustering Algorithm

Clustering techniques are applied when there is no class to be predicted but rather when the instances are to be divided into natural groups (Witten and Frank, 2000). Based on a number of trials with all available clustering algorithms provided by WEKA, it was found that the classic SimpleKmeans which employs the k-means algorithm generates meaningful clusters. The K-means algorithm forms clusters in numeric domains by partitioning instances into disjoint clusters. As an iterative distance-based clustering, k-means is very simple. It is performed by specifying in advance k clusters that are being sought. This k points are chosen at random as cluster centres. According to the ordinary Euclidean distance function of instance to the centres, each instance is assigned to a different cluster. Subsequent steps modify the partition to reduce the sum of distances of each case from the mean (centroid) of the cluster to which each case belongs. The modification consists of allocating each case to the nearest of the k means of the previous partition. This leads to a new partition for which the sum of distances is strictly smaller than before. The improvement step is repeated until the improvement is very small.

This clustering method is effective in generating similar collections that simplify the representation of data sets. Simplification plays a significant role especially with very large scale of data with multi-dimensional attributes. From a practical perspective, the clustering algorithm can help to identify the critical attributes in a multi-dimensional space. For example, applying this data mining technique of clustering using SimpleKmeans on available industrial data, the data set was found to be divided into two clusters centred at two major types of A/C malfunction: *too_hot* in March, and *too_cold* in January with an approximately even distribution around 50% as shown in Figure 6.3. This is a potential knowledge which could be used to guide future maintenance and building management.

Cluster Centroid:

Cluster 0
60325.098039215685 CM ZZZZ M 0 21271 6 A/C_Malfunction ***too_hot*** Adjust ***Mar*** WOS 1.0127091858823529E12 1.0141509239683931E12 1.0127128611764706E12 1.0130870229019608E12 1.0210212768711112E12 N C

Cluster 1
55795.125 CM ZZZZ H 0 10 7 A/C_Malfunction ***too_cold*** Adjust ***Jan*** WOS 1.0052290457142858E12 1.0067068180170569E12 1.0053050014285714E12 1.0056454119285714E12 1.015600460645E12 N C

Clustered Instances
0 51 (48%)
1 56 (52%)

Figure 6.3. A clustering result generated from applying the SimpleKmeans algorithm on the maintenance data of Building 10.

However, K-means have proven to be one of the most popular clustering algorithms because of its simplicity and effectiveness, it is time-consuming for relatively massive data sets due to the numbers of iterations involved in the process of clustering.

6.3 Data Mining Using the Classification Tree Algorithm

A decision tree is a tree-based knowledge representation methodology used to present classification rules. The leaf nodes present class labels. Various classification algorithms offered by WEKA have been applied and it was found that several algorithms were not able to deal with the maintenance data sets available due to some limitation in processing certain data types. For instance, some algorithms were not able to accommodate numeric values while others failed to accommodate nominal variables. The C4.5 algorithm (built on the top of ID3

proposed by Quinlan (1993)), was selected because of its ability to deal with numeric and nominal variables, and to handle missing values and pruning. The latter can be done by replacing a whole sub-tree by a leaf node if the expected error rate in the sub-tree of a rule obtained is greater than it in the single leaf.

The C4.5 algorithm generates a classification-decision tree for a given data-set by recursive partitioning of data. The construction of decision tree is described as “Divide and Conquer”. The search is for an attribute with best information gain at root node for splitting the tree into sub-tree which can be further partitioned recursively following the same rule. The splitting stops when there is no an information gain or whenever it reaches the leaf node. This process is sometimes called top-down induction of decision trees. Once the tree is constructed, rules can be generated by traversing each branch of the tree and collecting the conditions at each branch of the decision tree. For an instance, the rules associated with decision tree from applying C4.5 on the “month” attribute of “thermo static mixing valve” on the maintenance data of Building 10 is described in Figure 6.4. All monthly high priority maintenance works were carried out in the later part of the year, July to November. All 6 monthly maintenance works happen to be in December.

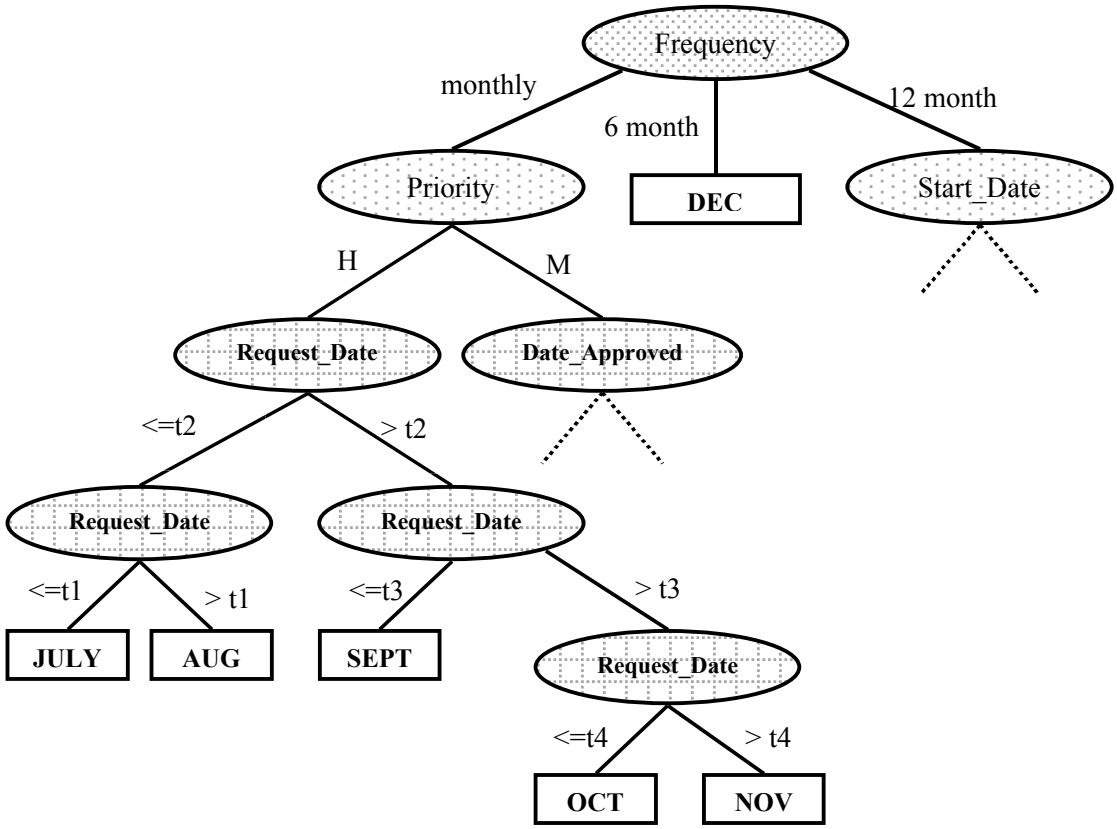

Figure 6.4. Part of decision tree generated from C4.5 on the “month” attribute (where t1<t2<t3<t4)

C4.5 is a powerful classifier that is robust to noise but its performance relies on the data sets on which it runs. For instance, C4.5 is applied to the maintenance data of air handling units of Building 10 in which some attributes have unchanged values, such as “job_subtype” with 105 “zzzz” out of 107 and “workorder_status” with 105 “C” out of 107. The performance of the classifier was not that effective and its results were less meaningful.

6.4 Data Mining using the Association Rule

The association rule technique involves finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories (Han, 2001). The association rule algorithm adopted in WEKA is “Apriori” which is developed by IBM’s Quest project team. Apriori finds all associations that satisfy a set of criteria with minimum support and minimum confidence. Support (also called coverage) refers to the number of instances predicted correctly. Confidence

(also called accuracy) is the proportion of the number of instances that a rule is correctly applied to them (Witten and Frank, 2000). Rules with high support are of interest and some rules are pruned out due to their low coverage. The basic idea of Apriori is to generate item sets that are combinations of attribute-value pairs with the minimum coverage. Apriori is efficient in searching the associations and correlations between attributes. However, to avoid having a great number of rules but less meaningful, there is a need to filter out all irrelevant attributes and find the groups of correlated attributes prior to applying the algorithm. WEKA provides an attribute evaluator in which some search methods such as the “BestFirst” can be used to sort out groups of correlated attributes. After applying this filtering process on the available maintenance data of the Air Handling Units at Building 10, the association rule algorithm “Apriori” was later applied and meaningful relational rules were obtained. Some of these rules include: “for floor 5, 6, 7, workOrder_Status was always completed”; “for all too_hot and too_cold descriptions, workOrder_Status was completed”; and “all works in floor 7 and in department 26462 belongs to A/C malfunction”. Unfortunately, “Apriori” deals only with nominal attribute values. Numeric, date, string data types are not applicable to most of the associative rule algorithms.

7. EVALUATING RESULTS OF APPLYING DATA MINING TECHNIQUES ON BUILDING MAINTENANCE DATA

Visual data analysis and data mining techniques were applied on two selected data sets: air handling units and thermostatic mixing valves at Building 10, Royal Prince Alfred Hospital, Central Sydney Area Health Service. The evaluation of the results obtained from mining the maintenance data of the above two asset types and their impact on improving the maintenance of existing buildings and the design of future facilities are shown in Tables 7.1 and 7.2.

Table 7.1 Evaluating results of applying data mining techniques on air handling units and their impact on improving the maintenance of existing buildings and the design of future facilities.

Data Mining Technique	Rules Obtained	Potential Impact on Facility Maintenance and Design
Visual Analysis	Approximately all "A/C malfunction" belongs to high and medium priority.	A/C malfunction" is of a major concern in guiding the allocation of maintenance resources.
	"A/C malfunction" is concentrated on the problems of: *too_hot* 32%, *too_cold* 28%, *not_working* 7.5%.	Temperature should be automatically adjusted and a provision of self-reporting faults equipments should be considered.
	The lowest levels of "A/C malfunction" took place in August followed by June and April while other months share similar high rate of "A/C malfunction".	Correlations between seasons and malfunction rates should be considered in guiding maintenance resource allocation.
	The maintenance work at 4, 5, 6, 7th floor constitutes most of the reports of A/C malfunctions, with 86% of A/C malfunction reported from these floors.	
	Approximately all the descriptions of *too_cold* or *too_hot* were associated with high or medium priority.	The appropriate temperature is of high priority from users' perspective.
Visual Analysis + Association Rule Algorithm	All 7th floor jobs were of high and medium priority and the cause of repairing was "A/C malfunction".	Investigate the possibility of poor design or maintenance of air conditioning function in 7th floor. A special attention in the design should be given to a specific floor due to its high demand of corrective or preventive maintenance or special design of A/C.
Visual Analysis	Higher percentage of user dissatisfaction in terms of work completed not meeting expectations is associated with maintenance work of high and medium priority.	Seeking feedback from users is important in order to improve the efficiency of building maintenance and achieving user satisfaction.
	Higher level of unhappiness related to completion not meeting expectation with a focus at *too_hot* and *too_cold* adjustment activities.	Paying attention to temperature adjustment in maintenance and design phrase may improve users' happiness.
	Cost centre 0 has the highest percentage of maintenance not meeting expectations (44%)	A special attention should be directed to certain places in the building wherein maintenance work is more likely to consume more time and effort than expected compared to normal places.

Decision Tree Algorithm –C4.5, Association Rule Algorithm	Department 26462 only reports A/C malfunction. (all 18 cases)	A special attention should be directed to certain places in the building wherein maintenance work is required more often.
	96% jobs for cost_centre = 0 is CM (corrective maintenance).	
Association Rule Algorithm	For floors 5, 6 and 7, the workOrder_Status was always completed.	Benefiting from successful maintenance practices including both equipments and labour is useful to achieve a high level of an overall maintenance performance.

Table 7.2 Evaluating results of applying data mining techniques on thermostatic mixing valves and their impact on improving the maintenance of existing buildings and the design of future facilities.

Data Mining Technique	Rules Obtained	Potential Impact on Facility Maintenance and Design
Visual Analysis	The percentage of high priority work constitutes of 55% of monthly work, 22% of 6mthly work and 24% of 12mthly work.	These percentages should direct the allocation of maintenance resources at the appropriate time of the year to achieve better planning and scheduling of maintenance work.
Visual Analysis + Decision Tree Algorithm (C4.5)	12mthly work occurred during the middle of the year – June-Sept, while all 6mthly occurred in December.	
Visual Analysis	All outstanding works took place in December	
	All monthly and 12mthly works were completed. Parts of 6mthly works (50%) were outstanding.	The 6mthly maintenance work should be thoroughly analysed to identify the actual reasons of incompletion.
	All high priority works were did not meet the expected completion date.	Maintenance job required with high priority should be investigated in relation to maintenance labour and the practicality of initial expected date of completion that has not been met.
	All works between August and December did not meet the completion data.	
	All medium priority works were completed on the expected completion data.	Analyse this phenomenon to help identifying the deficiency of not meeting the expected completion with high priority works.
Decision Tree Algorithm (C4.5)	All monthly high priority works are carried out in the later part of the year – July to November.	Distribution of priority of maintenance work is important in planning and scheduling maintenance work and resources.
Association Rule Algorithm	There is an incremental relationship between the work priority, the estimated time to complete the work and associated budget.	A better planning and scheduling will help to advance this pattern of relationship.

8. DISCUSSION

Facility managers and building owners are more concerned with highlighting areas of existing or potential maintenance problems in order to be able to improve the building performance, satisfying occupants and minimising turnover especially the operational cost of maintenance. Applying data mining techniques on the available industrial maintenance data has helped to discover useful rules that allowed locating some critical issues that have substantial impact on improving the building life cycle.

Data mining techniques assisted in identifying critical cost issues. For instance, discovering that corrective maintenance accounts for approximately 55% of all work orders implies a high level of unplanned maintenance that contributes to increasing the operational cost. The maintenance services required for the air conditioning system were related to thermal sensation complaints (too_hot 32%, too_cold 28%, not working 7.5%; total 67.5%). Federspiel (1998) found that thermal sensation complaints in buildings account for 75% of all environmental complaints from occupants and estimated that the labour cost associated with HVAC maintenance could be decreased by 20% by reducing the frequency of thermal sensation complaints that cost $2 billion annually in the U.S. (Martin et al, 2002).The “cost of discomfort” or “unsolicited complaints” is associated to increasing the operational cost of buildings due to the unexpected maintenance service (Federspiel et Al, 2003). Hence, applying data mining techniques greatly helps facility and building managers to identify the crucial maintenance issues and directs the improvement of strategic planning to add value to the life cycle of buildings.

Other benefits include constructing predictive plans based on correlations obtained from applying data mining techniques on the maintenance data sets of buildings. For instance, considering the role of potential correlations between seasons and malfunction rates in guiding the allocation of maintenance resources. Also, investigating any abnormal phenomenon discovered fro the maintenance data set such as “all outstanding works took place in December”. An investigation is required to study the relationship between the cause of increasing the outstanding maintenance jobs taking place in December the Christmas holiday or any the other causes. Appropriately addressing this problem will lead to better activities to improve the maintenance performance of existing facilities and will guide the design of future facilities.

The distribution of useful rules extracted from applying data mining techniques on two data sets of Building 10 is shown in Figure 9.1. The outcomes of applying data mining on industrial maintenance data can be enriched by including cost-related information and complete description of the task carried out on site and cause of repair.

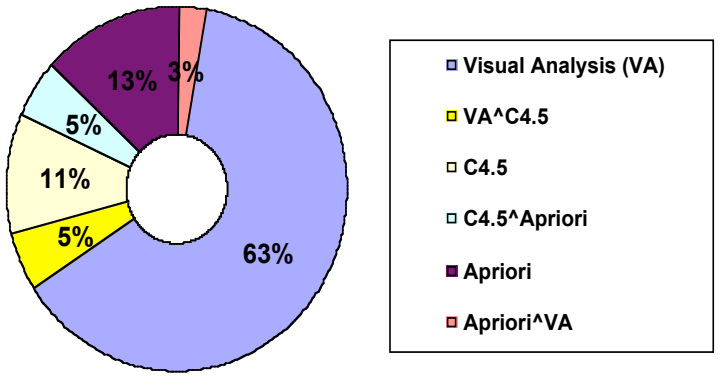

Figure 8.1 Percentages of rules extracted using various data mining techniques applied on Building 10.

9. ACKNOWLEDGEMENTS

The research presented in this paper is funded by CRC-Construction Innovation, Project No.: 2001-002-B “Life Cycle Modelling and Design Knowledge Development in Virtual Environment” and the University of Sydney. The industrial maintenance data is provided by Central Sydney Area Health Service, Royal Prince Alfred Hospital, NSW, Australia.

10. REFERENCES

Arditi, D. and Gunaydin, M. H.: 1998, Factors that affect process quality in the life cycle of building projects, *Journal of Construction Engineering and Management*, ASCE, **124** (3): 194-203.

Barringer, H. P. and Weber, D. P.: 1996, Life cycle cost tutorial, *Fifth International Conference on Process Plant Reliability*, Gulf Publishing Company, Houston, TX.

Burait, J. L., Farrington, J. J. and Bedbetter, W. B.: 1992, Causes Of Quality Deviations In Design And Construction, *Journal Of Construction Engineering And Management*, ASCE, **118**(1), 34-49.

Dhillon, B. S.: 1989, *Life Cycle Costing: Techniques, Models, and Applications*, Gordon and Breach, New York.

Dix, A. and Ellis, G.: 1998, Starting Simple - adding value to static visualisation through simple interaction, *in* T. Catarci, T., Costabile, M., Santucci, G. and Tarantino, L. (eds.), *Proceedings of Advanced Visual Interfaces AVI98*, ACM Pres, L'Aquila, Italy, pp. 124-134.

Federspiel, C., Martin, R. and Yan H.: 2003, Thermal comfort models and complaint frequencies, *CBE Summary Report*.

Federspiel, C.C.: 1998, Statistical analysis of unsolicited thermal sensation complaints in commercial buildings, *ASHRAE Transactions*, **104**(1): 912-923.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C.: 1992, Knowledge discovery in databases: An overview, *AI Magazine,* **13**: 57-70.

Han, J. and Kamber, M.: 2001, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco.

Hui, SC, Jha, G: 2000, Data Mining for Customer service support, *Information and Management*, **38**: 1-13

Martin, R. M., Federspiel, C. and Auslander, D.: 2002, Responding to thermal sensation complaints in buildings, *ASHRAE Transactions*, **112** (1): 407-412.

Peitgen, H. O., Jurgens, H., and Saupe, D.: 1992, *Chaos and Fractals: New Frontiers of Science*, Springer-Verlag, New York.

Quinlan, R: 1993, *C 4.5 Programs for Machine Learning*, Morgan Kaufmann, San Francisco.

Siewiorek, D. P. and Swarz, Robert S.: 1982, *The Theory and Practice of Reliable System Design*, Digital Press, Bedford, MA.

Soibelman, L. and Kim, H.: 2002, Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, **16** (1): 39-48.

Witten, I. and Frank, E.: 2000, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufman, San Diego.