CONFERENCE THEME: INNOVATIVE ASSET MANAGEMENT

Full Paper

IMPROVING THE MANAGEMENT OF BUILDING LIFE CYCLE: A DATA MINING APPROACH

Rabee M. Reffat, John S. Gero and Wei Peng

Key Centre of Design Computing and Cognition
University of Sydney, NSW 2006, Australia
Email: {rabee or John@arch.usyd.edu.au}

ABSTRACT

The construction industry has adapted information technology in its processes in terms of computer aided design and drafting, construction documentation and maintenance. The data generated within the construction industry has become increasingly overwhelming. Data mining is a sophisticated data search capability that uses classification algorithms to discover patterns and correlations within a large volume of data. This paper presents the selection and application of data mining techniques on maintenance data of buildings. The results of applying such techniques and potential benefits of utilising their results to identify useful patterns of knowledge and correlations to support decision making of improving the management of building life cycle are presented and discussed.

Keywords: Building Life Cycle, Data Mining, Management

1. INTRODUCTION

The construction industry has adapted information technology in its processes in terms of computer aided design and drafting, construction documentation and maintenance. 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 our ability to interpret and digest this data. This issue becomes critical when 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 the potential to improve the maintenance management of building assets. Traditional methods of data analysis such as spreadsheets and ad-hoc queries are not adequate since 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 management of the life cycle of buildings. Current information technology applied to facility maintenance utilises databases to keep track of information and notification of maintenance schedules. However, these databases are not well linked with interactive 3D models of buildings and are 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 the 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 a more economical life cycle of buildings. Furthermore, designers and maintenance managers will be better equipped to achieve higher performance by utilising appropriate techniques of information technology at their workplace.

This paper presents the selection and application of data mining techniques on maintenance data of buildings. The paper addresses potential benefits of applying such techniques to identify useful patterns of knowledge and correlations within the existing building maintenance data to support improving the management of future maintenance operations.

2. INCORPORATING DATA MINING INTO BUILDING LIFE CYCLE

Building systems should always be available to support building functions. The maintenance objective for a 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.

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. However, even with the best of such tools, low quality input information will produce inaccurate output of schedules and budget.

Data mining, as an extraction of implicit, previously unknown, and potentially useful information from data (Frawley et al, 1992) provides useful tools that help to explain how building systems that were once thought to be completely chaotic have predictable patterns (Peitgen et al, 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 building projects, and activities of future projects to avoid unanticipated consequences (Soibelman and Kim, 2002). Data mining presents a significant potential for addressing the problem of transforming knowledge implicit in data into explicit knowledge for decision makers.

2.1 A DATA MINING APPROACH FOR BUILDING LIFE CYCLE

The approach, presented in this paper, views the process of building design, maintenance, and replacement as a process that generates high volumes 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 can be much smaller than the costs of operating a building or other structure over the course of its life span. Data mining enables building owners, designers and facility managers to make important decisions about the building life cycle cost in advance, thereby significantly affecting and improving 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 1 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 present current approaches to building information management. The integration of data mining within the process of information flow provides the opportunity to increase the value of building data and to feedback and improve the processes of building design and maintenance.

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 1 includes:

- Components that frequently need maintenance and therefore need to be inspected carefully.
- Historical consequences of maintenance decisions that may inform future decisions.
- Components of buildings that significantly determine maintenance cost and therefore may inform future building designs, as well as refurbishment of the building in question.

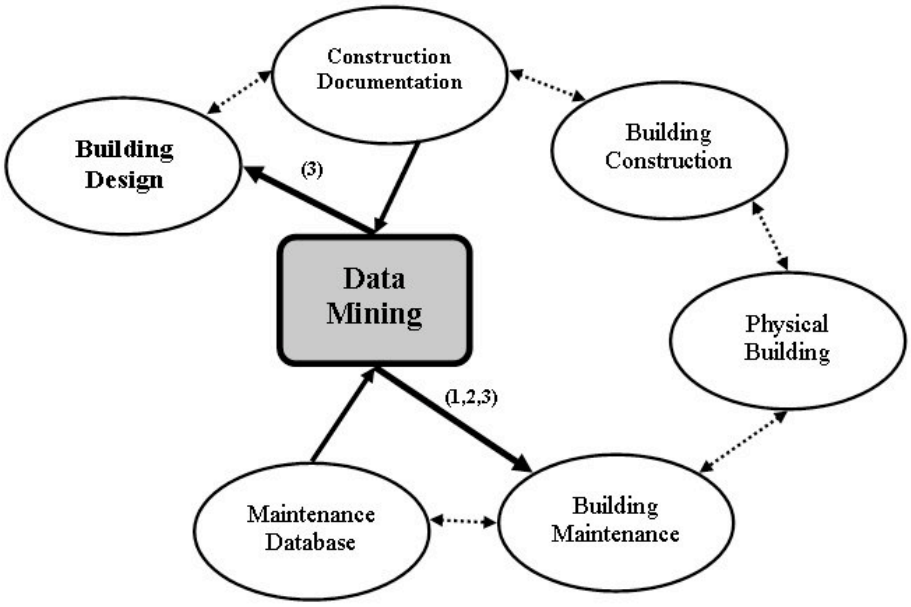


Figure 1. Integrating data mining within the life cycle of buildings.

It has been shown in the AEC (Architecture, Engineering and Construction) industry that major factors contributing to construction quality problems include inadequate information and poor communication (Burait et al, 1992; Arditi and Gunaydin, 1998) 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 has been developed and tested using an industry building maintenance database.

2.2 DATA MINING PROCESS

Data mining involves going from problem specification to the implementation of tools, and monitoring of the model. 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 2. The process focuses on finding interesting patterns that can be interpreted as useful knowledge and 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 not all of which 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 normally 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 not only contains 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. The information obtained is analysed.

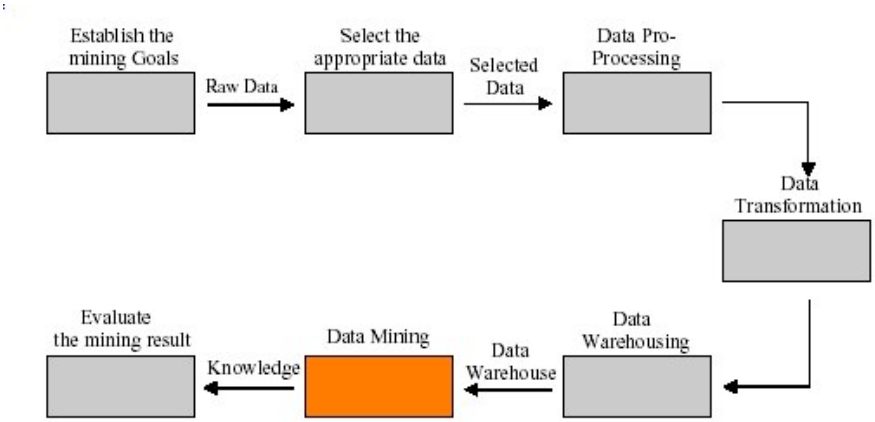


Figure 2. Stages of data mining process.

3. APPLYING DATA MINING TECHNIQUES ON INDUSTRY MAINTENANCE DATA

The maintenance data 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 are approximately 5,000 work orders recorded for Building 10 in the period from 1 January 2001 to 9 December 2002.

3.1 DATA MINING USING VISUAL ANALYSIS APPROACH (STACKED HISTOGRAM)

In a histogram, data is collected and sorted into categories. Histograms focus on the frequencies and distributions of one particular attribute, for example, the Priority description for the entire data set is illustrated in Figure 3. In order to find out correlations between various attributes, there is a need of an interactive visualisation rather than a static view of histogram.

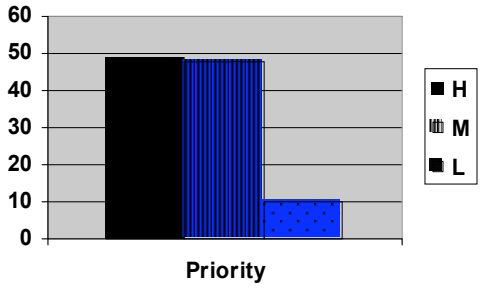


Figure 3. A standard histogram of the "Priority" attribute.

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 cross comparison of standard histogram by allowing different trends to be analysed using the same graph. Thus, the correlation between attribute "priority" and "cause-of-repair" can be visualised as shown in Figure 4. 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.

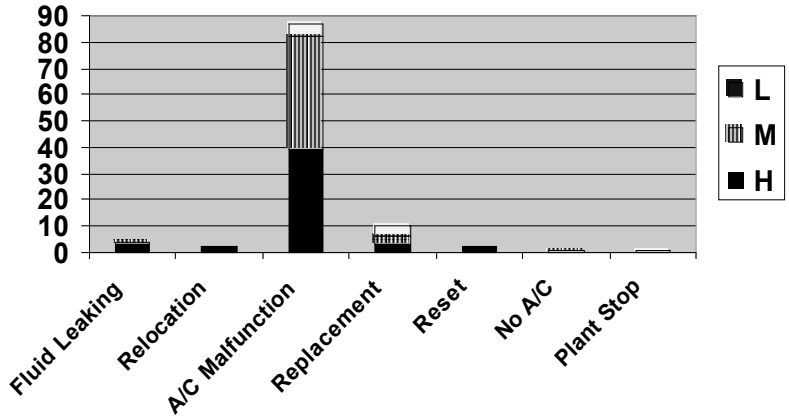
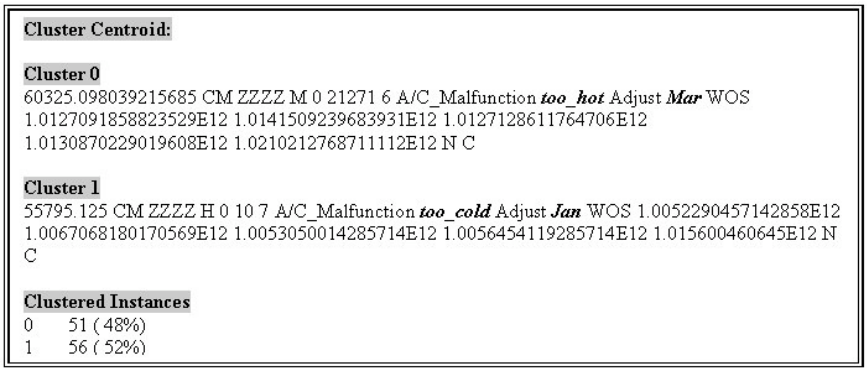


Figure 4. A stacked histogram of correlation between "Priority" and "Cause-of-repair".

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.

3.2 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. 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 5. This is potential knowledge which could be used to guide future maintenance and building management. 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.



Cluster Centroid:

Cluster 0
60325.098039215685 CM ZZZZ M 0 21271 6 A/C_Malfunction ***too_hot*** Adjust ***Mar*** WOS
1.0127091858823529E12 1.0141509239683931E12 1.012712861176470 6E12
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 5. A clustering result generated from applying the SimpleKmeans algorithm on the maintenance data of Building 10.

3.3 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. 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. 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.

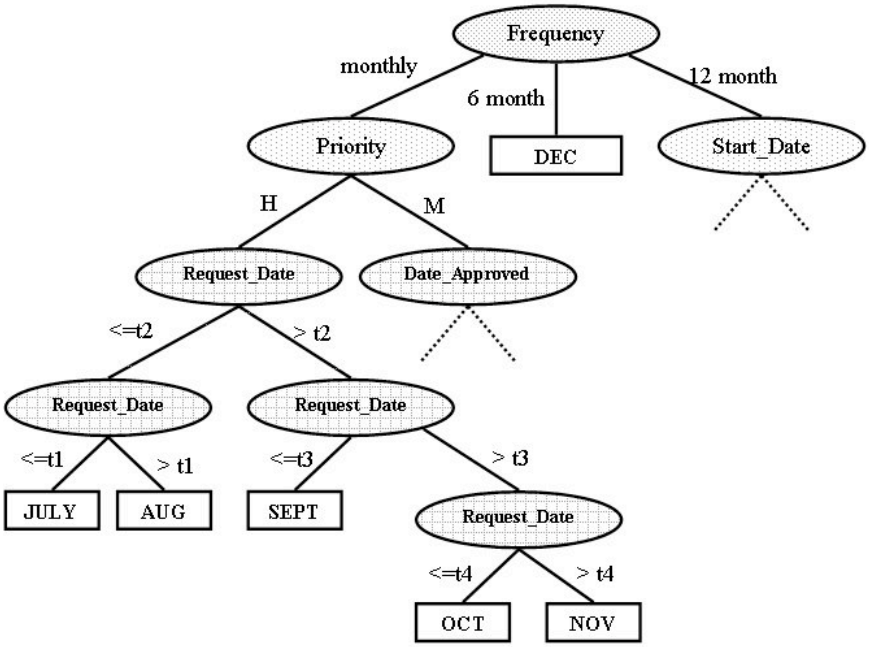


Figure 6. Part of decision tree generated from C4.5 on the "month" attribute (where t1<t2<t3<t4)

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.

3.4 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 and Kamber, 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 less meaningful rules, 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.

4. POTENTIAL BENEFITS OF UTILISING DATA MINING TECHNIQUES TO IMPROVE THE MANAGEMENT OF BUILDING LIFE CYCLE

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 potential benefits on improving the maintenance of existing buildings and the design of future facilities are shown in Tables 1 and 2.

Table 1. Applying data mining techniques on maintenance data of air handling units and potential benefits of improving the maintenance management during its life cycle.

Data Mining Technique	Acquired rules	Potential benefits on improving the management of building life cycle
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 2. Applying data mining techniques on maintenance data of thermostatic mixing valves and potential benefits of improving the maintenance management during its life cycle.

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.
Analysis + Decision Tree Algorithm	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.

5. 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, satisfy occupants and minimise 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 management of building life cycle.

The distribution of useful rules extracted from applying data mining techniques on two data sets of Building 10 is shown in Figure 7. 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.

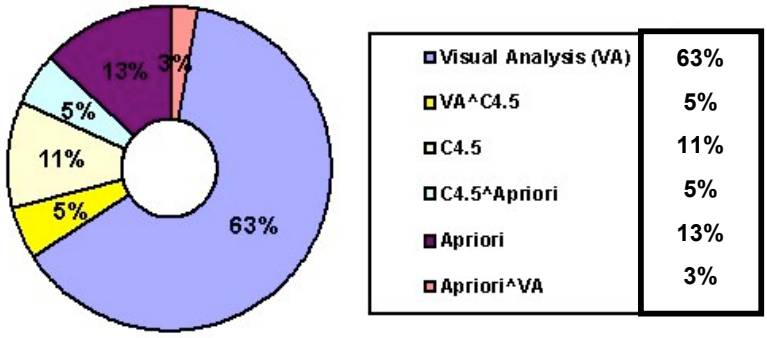


Figure 7. Percentages of rules extracted using various data mining techniques applied on Building 10.

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, 2003). Hence, applying data mining techniques assists 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 from 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 management of existing facilities and will guide the design of future facilities.

ACKNOWLEDGEMENT

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”. The industrial maintenance data is provided by Central Sydney Area Health Service, Royal Prince Alfred Hospital, NSW, Australia.

REFERENCES

Arditi, D., 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

Burait, J. L., Farrington, J. J., Bedbetter, W. B.: Causes of quality deviations in design and construction, Jo*urnal Of Construction Engineering And Management, ASCE*, **118**(1): 34-49.

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

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

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., Matheus, C.: 1992, Knowledge discovery in databases: An overview, *AI Magazine*, Vol. **13**: 57-70

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

Hui, S. C, Jha, G: 2000, Data mining for customer service support, I*nformation and Management*, **38**:1-13

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

Peitgen, H. O., Jurgens, H., 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 Mateo, Calif.

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