JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT
2008
14(2): 79–84

UTILIZING DATA MINING TO DISCOVER KNOWLEDGE IN CONSTRUCTION ENTERPRISE PERFORMANCE RECORDS

Jen-Rong Lee[1]**, Sung-Lin Hsueh**[2]**, Hung-Ping Tseng**[3]

[1]Dept of Construction Engineering, National Kaohsiung First University of Science and Technology (NKFUST), University Road, Yuanchau, Kaohsiung 824, Taiwan
[2]Dept of Arts and Crafts, Tung Fang Institute of Technology, No. 110. Tung Fang Rd. Hu-Nei Shang Kaohsiung Hsien, Taiwan, R.O.C
[3]Dept of Construction Engineering, National Kaohsiung First University of Science and Technology (NKFUST), University Road, Yuanchau, Kaohsiung 824, Taiwan
E-mail: [1]jrlee@ccms.nkfust.edu.tw; [2]hsueh.sl@msa.hinet.net; [3]ncctseng@yahoo.com.tw

Received 13 July 2007, accepted 10 April 2008

Abstract. Although Data Mining (DM) has been applied extensively in information systems and identified as a crucial tool for automatic data analysis and enterprise knowledge inference, still the practicability of DM has been little explored in the construction industry in particular. This study is conducted using data from actual practice in the customer service department of the target enterprise. Starting with data preparation, decision tree analysis and domain knowledge inference, practical verification is performed and previously unknown knowledge is discovered. Two practical barriers to enterprise information mining are found: the separation of information among various information systems and lack of key information attributes. Hence, there is a significant limitation on practical information data mining to generate new information, and only with a proper information integration one can benefit from the potential practical application of DM.

Keywords: data mining; information system; construction industry; knowledge of enterprise; service and maintenance.

1. Introduction

Data mining (DM) is one of the core methodologies of knowledge discovery in databases (KDD). Most DM research has focused on developing algorithms for more accurate models or for faster ones. Hence, DM studies have ignored not only data preparation, domain knowledge and conceptual models but also the importance of data integration for mining in the knowledge discovery process (Chaudhuri *et al.* 2001). Likewise, publications of practical applications of DM in industry are comparatively rare. In the meantime, the construction industry in Taiwan is confronting low-profit situations and severe competition. Raising profits and competitiveness of organizations has become a significant issue that executives and entrepreneurs should face. Thus, this paper demonstrates a successful DM application to encourage the application of knowledge management in the construction industry to overcome its present unavoidable difficulties.

2. Information characteristics of the construction industry

The construction industry is characterized by keen competition and scattered trait (Cheng *et al.* 2001) and it is a project-oriented industry (Voordijk *et al.* 2000) in which designers, builders and suppliers vary on different projects (O'Brien, Li 2001). In addition, it is a producing process, which provides a completed product, service and satisfaction (Torbica, Stroh 2001). The management of construction is multi-dimensional and complex in every aspect: labour, capital, technique and service. Construction knowledge is mainly stored in disconnected segments namely: proprietors, supply chains, organizations, employees, and customers. Therefore the knowledge related to construction is complicated to manage. Due to the rapid development of information technology (IT), which allows efficient handling of masses of information, the construction industry can now solve management problems caused by the plethora of information. The use of IT can be positive. By comprehending the knowledge processes within the company and recognising the hurdles in the way of change, a company should be able to identify and exploit its own unique sources of competitive advantage (Jordan, Jones 1997).

3. DM overview

Data mining, one powerful tool for knowledge discovery in databases, is a process of seeking and inspecting data in order to find complicated, but potentially useful, embedded information (Berry, Linoff 1997). It is also a process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures from large amounts of data stored in a database, data warehouses, or other information repositories

ISSN 1392–3730 print / ISSN 1822–3605 online
DOI: 10.3846/1392-3730.2008.14.2

(Hui, Jha 2000; Mao-Lin Chua, Ju-Hung Lan 2005). Data mining is able to automatically analyse the information in a database and attempts to interpret irrational knowledge so as to achieve the goal of creating new knowledge. It applies recursive iteration to classify data into groups based on the characteristics of data. Accordingly, there are different approaches utilized in DM for different data types of information such as decision tree, clustering, and other sophisticated stochastic algorithms.

4. DM analysis

Algorithms of all kinds are being developed, at present, within the field of DM research, and the results are categorised into the following groups:

1. Classification Rules (Anwar *et al.* 1992): using the accessible information to establish the behaviour pattern of each parameter in a database so as to categorize them into different classifications such as: decision tree.
2. Association Rules (Agrawal *et al.* 1993; Houtsma, Swami 1995): focusing on two items (or more) of information and searching for connections between them.
3. Sequence Rules (Agrawal, Srikant 1995): aiming at one pair (or more) to find the sequence rules in between.
4. Similar Time Series: applying them allows us to discover similar events happening within a certain period of time and to link them with certain connecting features, which can be additional information for advertisements or commercial promotions and thus extend promotions. Besides, it can be employed for medical research to discover the connection between symptom and illness in order to improve medical quality.
5. Clustering Rules: these rules have been long applied in statistics for dealing with numerical information. In data mining, some interesting problems are involved when handling non-numerical information.

Most of the above algorithms are designated for specific data types or for a specific analysis. Some of them can only be applied to continuous numerical data analysis. The decision tree approach is one among them that is commonly used and is capable of analysing all types of data. Thus this study employs decision tree analysis in its DM process.

5. Use of decision tree to conduct data mining

Since the data types of attributes in the target system are various, this study employs the decision tree (DT) algorithm aiming to reveal underlying management problems in order to improve business performance. The following are reasons for and advantages in applying decision tree algorithm:

1. The DT algorithm can manage both continuous and discrete information. Especially when the key attributes are not clearly identified in the target system, the algorithm is needed to have the competence to handle variations among all original data attributes. DT algorithm therefore meets the needs of this study.
2. DT algorithm generates and demonstrates easily comprehended rules, helping the researcher to contrast the analysis results with the domain knowledge and to discover possible causes of poor quality in building construction.
3. The DT algorithm identifies the level of significance of differentiation among independent variables, guiding people to distinguish the most important variations in the classification process. Thus, from its analysis reports, people can comprehend the key variations reflected by the information and further, discover the causes of quality deficits through the integration of variation features and domain knowledge.

In addition to the previously stated advantages, DT analysis provides a better efficiency than other algorithms. It identifies the most different variables, then separates data into branches, until significant differences cannot be found in a single branch. The calculation steps in DT algorithm are comparatively simple, thus efficiency can be easily recognised when processing a large amount of data using DT algorithm. Consequently, more processes can be applied and more conclusions may be achieved.

6. Prototype experiments and analysis

The source data employed in this study was collected from the service and maintenance department of the target company. That company is a listed construction enterprise in Taiwan with approximate capital of USD 175 million. The total received maintenance request cases in year 2000 are taken as the target data for DM analysis. Since external environmental factors, weather especially, have great impact on the requiring tendency of some after-sale service items, taking one complete year as the research duration, may significantly reduce influence due to weather-related effects.

6.1. Data collection and selection

This study takes historical maintenance data from the company's residential building project records to serve as the target data for analysis. These data were generated and maintained based on the maintenance and repair work done by the service and maintenance department of the studied entity during 1994–97. The total number of after-sale service requests during these four years was 7790 cases and those cases are categorised into 35 service items in the database, as shown in Fig. 1. Among these data, the major problems caused by building construction defects are leakage- and crack-related problems. Consequently, this study focuses on the exploration of knowledge associated with the causes of cracks in concrete members.

In order to reduce environmental influence and achieve all needed information, only service request cases from the year 2000 and only for those buildings constructed by the company are extracted as target data in this research. Major statistics of the extracted request cases in service and maintenance department in 2000 are as follows:

1. Total cases of requests for service: 1857.
2. Total formally logged cases: 864.

3. Logged request cases for buildings built by the company: 589.
4. Number of buildings: 28.
5. Years of buildings completion: 1992–2000.

Although services were requested, some of them were not caused by building-related defects. The number of analysed data entities thus reduces from 1 857 to 864. Afterwards leaving out those services requests from buildings which were not built by the company, 589 data entities are gathered and analysed.

6.2. Data mining target

The target of the studied DM is information on after-sale service. As a result, it is tending to focus on the construction quality deficits hoping to discover the hidden causes of problems existing in planning, design, procurement, and construction processes via data mining so as to elevate the construction quality, reduce unnecessary expenses, increase customers' satisfaction and further to improve competitiveness. The conclusion of the primary data analysis of the information concerning 2000 indicates that the maintenance items mainly concentrated on leakage, which apparently relates tightly to the external environment. Consequently, the study aims at leakage as the main target of the data mining and discusses the factors influencing leakage.

With the intention to pinpoint the major causes of building construction defects and reduce maintenance costs for the enterprise, the revelation of novel knowledge linked to concrete cracking and leakage during the design and construction stages is crucial. The preliminary information analysed in this study is achieved from the service and maintenance database. There are separate data tables contained in that database system, including Project, Building, Unit, Request, Customer, Employee, and Service Provider data tables. Among these tables, the Request data table records the most needed information for this research. Major fields in the Request data table are: request number (primary key), project name, unit number, request time, transference time, address, phone number, employee, assigned service provider, service item, process status, customer name, accomplish time, customer cost, company cost, total cost, profit etc.

Although precious knowledge can be attained from the analysis of data held in the original database by applying statistical techniques and other research approaches such as DM, the contribution may well be limited due to insufficient data associations. To enlarge the accomplishments of data analysis, the establishment of connections with additional information can be vital. In order to discover the causes of leakage and cracking in a building, analysing only the service and maintenance data is not enough. The relationships between the service request data and the design and construction data should be established to generate inferences between the maintenance request and the design as well as the construction process. Thus additional data are collected from the building design and construction stages.

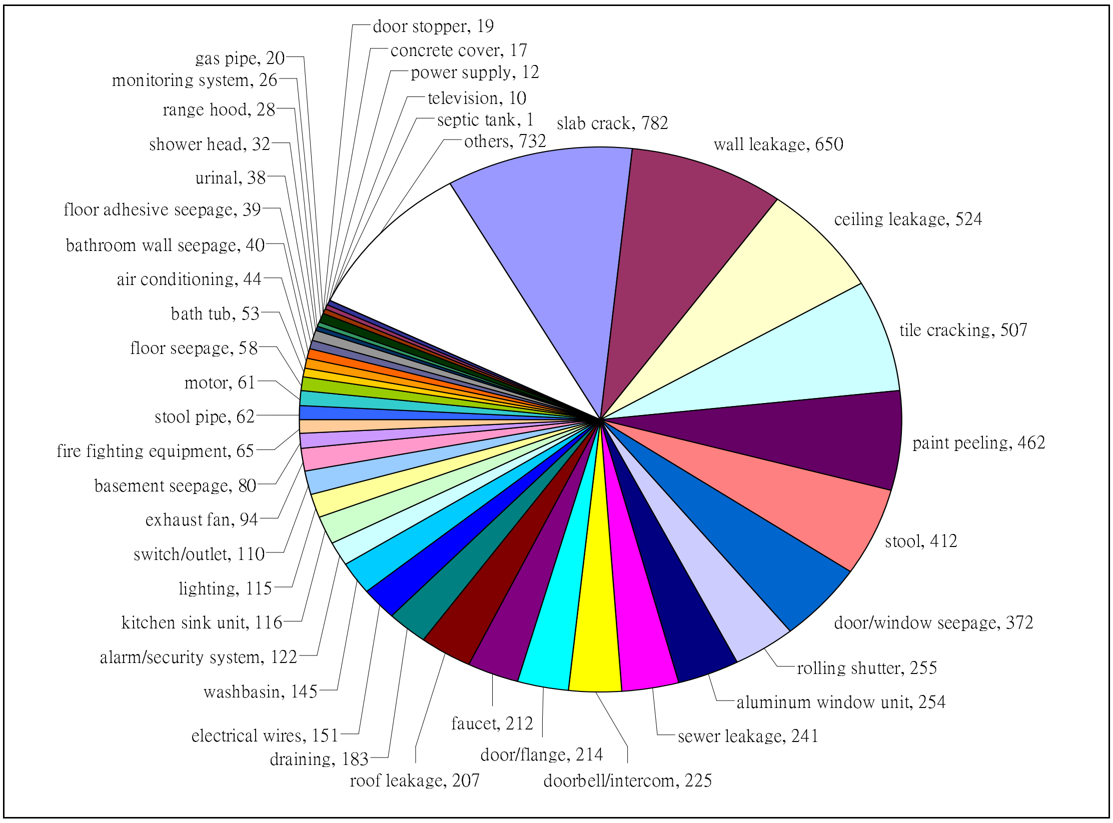


Fig. 1. Pie chart of maintenance works requested during 1994–1997

6.3. Data mining process

A complete DM process includes 5 steps: 1) data selection, 2) data preparation and cleansing, 3) data reduction and coding; 4) algorithm selection; and 5) mining and reporting. During the process, not only the initially proposed target of the analysis, but also some neglected phenomena may be clarified. Although the process flow is linear, every step of this process can be reversed, modified or conducted by other analysis methods if needed. DM is a very flexible tool and can be modified by the needs of researchers.

Data preparation is one of the bottlenecks in DM due to data flaws and missing data. Key columns of business information are mostly accurate such as sums of payment and payment locations. On the other hand, the majority of data fields, which are always taken as unnecessary information by employees, are commonly neglected, mistaken or missing. Although these mistaken, missing or incomplete data items can be reproduced, predicted or inferred through the application of statistical or fuzzy logic methods (Glymour *et al.* 1996; Cios *et al.* 1998; Rubin 1998), patient and careful data preparation is still vital.

In this research, the service and maintenance information system was originally designed to record the payment, cost and profit information for the company, the information that the shareholders care about the most. For this reason, DM has also been applied to predict the future profits and the real estate price indices (Yang Jianhui, Cai Hetting 2007). However, a significant portion of the engineering details is unclear in the target database. Therefore, in addition to the correction and clarification of the service and maintenance information, a huge amount of extra work was undertaken to recollect those missing technical data by checking all the work sheets written on site. Furthermore, some of the building design and construction information was not designated to be input into the database. Data items were collected from the design drawings and site records to carry out the following data mining process.

Following the data preparation steps, the refined data are analysed applying the Decision Tree method. Figs. 2 and 3 demonstrate the DT analysis results based on different target attributes. Fig. 2 displays the analysis results based on average concrete strength and it shows that, among buildings completed after June 15, 1998, 78.72 % of the leakage occurred on concrete members with average strengths of over 251 kg/cm^2. Fig. 3 is the result of the DT analysis, which takes the project manager as the target attribute and shows that 55.1 % of the leakage cases, in buildings completed after April of 1999, were supervised by the same project manager. Therefore, the analysis results suggest that the concrete strength and supervisor are probably important factors for building construction quality especially for leakage-related defects.

6.4. Interpretation of data mining result

The following are inferences drawn from DM results complying with the domain knowledge applied in construction industry after-sale service:

1. High-strength concrete, theoretically, can improve the water-tight performance of concrete, yet on the contrary; it is the main cause of leakage. The survey conclusion about the cause are:

(1) The leakage was concentrated on higher floors (58 % of leaks are on floors 20–26, with approx. elevation of 60–76 m). First, this is because high-floor concrete construction needs strong pumping pressure sending concrete to high elevations, so it can easily cause material segregation. And second, for convenience of workers, the management of the pumping process was often ignored, thus it was quite easy to have too much water in the concrete and this involves shrinkage caused by excessive water.

(2) Dismantling time. In general, the side form is dismantled on the second day, when the hydrogenation of concrete is not yet completed and, besides the severe wind on the high-floor levels together with direct exposure to the sun without shelter, let the water content of the wall evaporate rapidly, and thus cracking becomes inevitable.

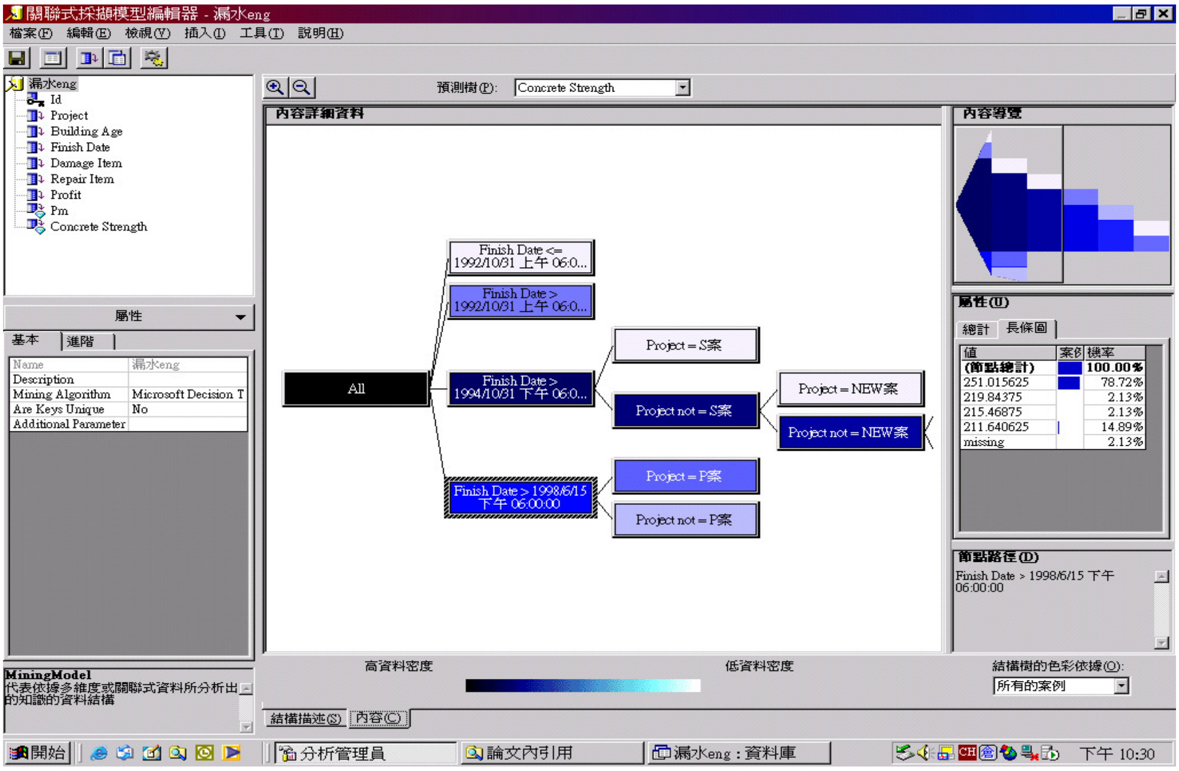
(3) Difficulty of high-strength concrete construction: there is a large consumption of many materials in making high-strength concrete, such as cement and slag. This leads to a high-density of concrete which will cause many disadvantages in construction. As a result, the consumption of water will increase in order to keep the concrete workable and this will cause an increase of shrinkage.

2. There are higher leakage rate and numbers on the construction sites supervised by the specific project manager than on other similar projects performed in the same period of time, and the reflected phenomenon shows the following implication:
   Insufficient training: high-floor construction needs different construction awareness from ordinary floor construction. Therefore it requires an intensive training on materials, process allocations, and material transportation (vertical and horizontal); otherwise the construction will be defective.

Although the key factors of the leakage causes are finally identified in this study, DM analysis is a quantitative analysis application performed primarily on massive data collections and its analysis results offer only an indication of potential causes of problems. The exploration of actual causes cannot be derived, unless the investigators are provided current knowledge, domain experts' experience and the analysis results.

7. Conclusions

An Enterprise Information System (EIS) consists of several individual information systems. These systems may operate individually without frequent interaction, but the relationships among data entities in different systems may generate inferences and reveal important knowledge. In this study, the target EIS is divided into several segments, namely: procurement management, construction management, service, and maintenance management. This study aimed initially to interpret the data stored in the service and maintenance management system and then explore a way to reduce maintenance costs and increase customer satisfaction. Subsequently, applicable results cannot be found merely from the analysis of maintenance records; relevant data are collected and included to increase the complexity of inferences and in the expectation they will generate valuable outcomes. Eventually, conclusions are attained based on the DM analysis. The concrete cracking resulting in leakage occurred 78.72 % of the time in the concrete with a design strength greater than 251 kg/cm^2, and 55.10 % of leakage cases occur on building projects inspected by the same site manager.



預測樹(P): Concrete Strength

Name	漏水Keng
Description	
Mining Algorithm	Microsoft Decision T
Are Keys Unique	No
Additional Parameter	

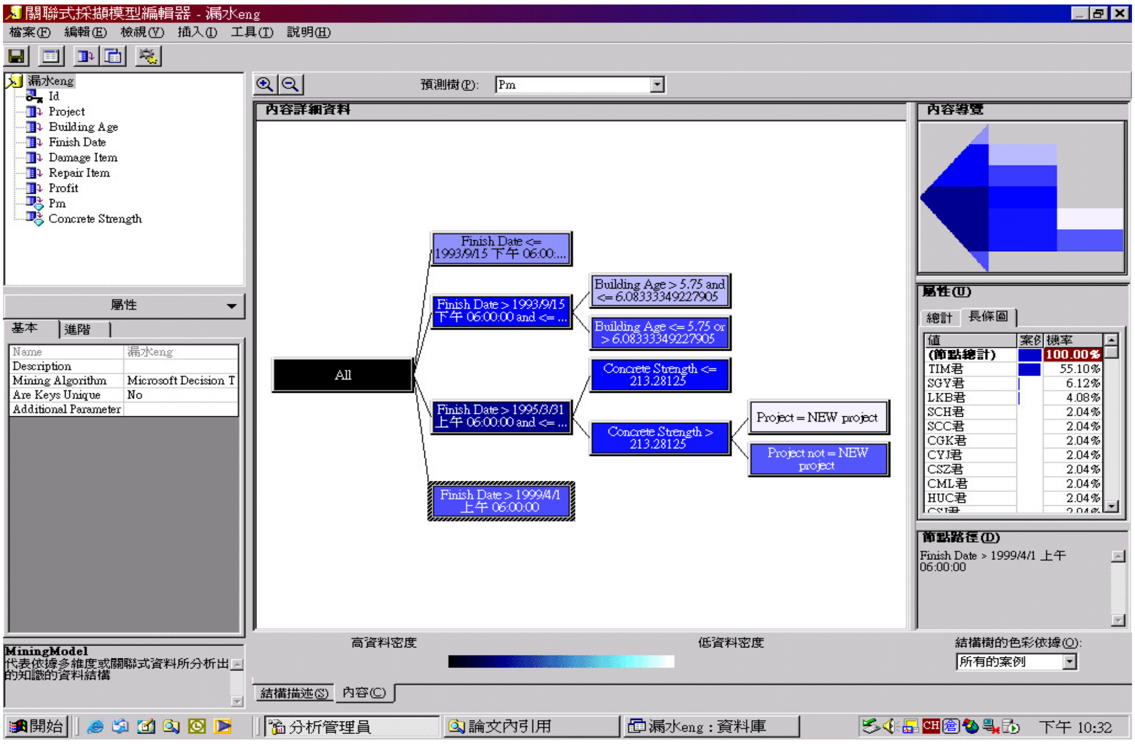
值	案例	機率
(節點總計)		**100.00%**
251.015625		78.72%
219.84375		2.13%
215.46875		2.13%
211.640625		14.89%
missing		2.13%

節點路徑(D): Finish Date > 1998/6/15 下午 06:00:00

結構樹的色彩依據(O): 所有的案例

MiningModel
代表依據多維度或關聯式資料所分析出的知識的資料結構

Fig. 2. Usage of average concrete strength for decision tree analysis


關聯式採擷模型編輯器 - 漏水Keng
檔案(F) 編輯(E) 檢視(V) 插入(I) 工具(T) 說明(H)
漏水Keng
Id
Project
Building Age
Finish Date
Damage Item
Repair Item
Profit
Pm
Concrete Strength
屬性
基本
進階
內容詳細資料
內容導覽
All
Finish Date <= 1993/9/15 下午 06:00:...
Finish Date > 1993/9/15 下午 06:00:00 and <= ...
Finish Date > 1995/3/31 上午 06:00:00 and <= ...
Finish Date > 1999/4/1 上午 06:00:00
Building Age > 5.75 and <= 6.08333349227905
Building Age <= 5.75 or > 6.08333349227905
Concrete Strength <= 213.28125
Concrete Strength > 213.28125
Project = NEW project
Project not = NEW project
屬性(U)
總計
長條圖
節點路徑(D)
高資料密度
低資料密度
結構描述(S)
內容(C)
開始
分析管理員
論文內引用
漏水Keng : 資料庫
下午 10:32

預測樹(P): Pm

Name	漏水Keng
Description	
Mining Algorithm	Microsoft Decision T
Are Keys Unique	No
Additional Parameter	

值	案例	機率
(節點總計)		**100.00%**
TIM君		55.10%
SGY君		6.12%
LKB君		4.08%
SCH君		2.04%
SCC君		2.04%
CGK君		2.04%
CYJ君		2.04%
CSZ君		2.04%
CML君		2.04%
HUC君		2.04%
CSJ君		2.04%

節點路徑(D): Finish Date > 1999/4/1 上午 06:00:00

結構樹的色彩依據(O): 所有的案例

MiningModel
代表依據多維度或關聯式資料所分析出的知識的資料結構

Fig. 3. Usage of project manager for decision tree analysis

Although the DM analysis method used in this research is fundamental, the research result proves that practical data can be analysed by DM without preliminary heuristic inferences. This indicates that the knowledge creation and findings can be instigated by people without an expertise in the specific domain of knowledge and achieve useful results. On the other hand, this research concludes that the completeness of data relationships among data entities in a large information system is one of the key factors in the successful knowledge management. Insufficient linkage between data tables may not influence the operation of the information system, but it may lead to poor results in knowledge extraction.

References

Agrawal, R.; Imielinski, T.; Swami, A. 1993. Mining association rules between sets of items in large databases, in *Proceedings of ACM SIGMOD*, May 1993, 207–216.

Agrawal, R.; Srikant, R. 1995. Mining sequential patterns, in *Proceedings of the 11th International Conference on Data Engineering*, Taipei, Taiwan, March 1995, 3–14.

Anwar, T. M.; Beck, H. W.; Navathe, S. B. 1992. Knowledge mining by imprecise querying: a classification-based approach, in *Proceedings of the International Conference on Data Engineering,* 622–630.

Berry, M. J. A.; Linoff, G. 1997. *Data mining techniques: for marketing sales, and customer support*. New York: Wiley Computer Publication.

Chaudhuri, S.; Dayal, U.; Ganti, V. 2001. Database technology for decision support systems, *Computer, IEEE* (December): 48–55.

Cheng, Eddie W. L.; Li, Heng; Love, Peter E. D.; Irani, Zahir. 2001. An e-business model to support supply chain activities in construction, *Logistics Information Management* 14(1/2): 68–77.

Glymour,C.; Madigan, D.; Pregibon, D.; and Smyth, P. 1996. Statistical inference and data mining, *Communication of The ACM* 39(11 November): 35–41.

Houtsma, M.; Swami, A. 1995. Set-oriented mining of association rules in relational databases, in *Int'l Conference on Data Engineering*, Taipei, Taiwan, March 1995, 25–33.

Hui, S. C.; Jha, G. 2000. Data mining for customer service support, *Information and Management* 38(1): 1–13.

Yang Jianhui; Cai Heting. 2007. Study of data mining in China's building industry, in *2007 IEEE International Conference on Control and Automation*, Guangzhou, China, 2735–9.

Jordan, J.; Jones, P. 1997. Assessing your company's knowledge management style, *Long Range Planning* 30(3): 392–398.

Krzysztof, J.; Cios, Pedrycz, W.; Świniarski, R. W. 1998. *Data Mining Methods for Knowledge Discovery*. Kluwer Academic, USA.

Lopes, J.; Ruddock, L.; Ribeiro, F. L. 2002. Investment in construction and economic growth in developing countries, *Building Research and Information* 30(3): 152-–159.

Mao-Lin Chiua; Ju-Hung Lan. 2005. Information and INformation Information mining for supporting collaborative design, *Automation in Construction* 14: 197–205.

O'Brien, D.; Li, C. 2001. A quantitative analysis of relationships between product types and supply chain strategies, *Int. J. Production Economics* 73: 29–39.

Rubin, S. H. 1998. A fuzzy approach towards inferential data mining, *Computer and Industrial Engineering*, *Elsevier Science Ltd*, 35(1–2): 267–270.

Torbica, Z. M.; Stroh, R. C. 2001. Customer satisfaction in home building, *Journal of Construction Engineering and Management* 127(1): 82–86.

Voordijk, H.; Haan, J.; Joosten, G. J. 2000. Changing governance of supply chains in the building industry: a multiple case study, *European Journal of Purchasing & Supply Management* 6: 217–225.

UTILIZUOTAS DUOMENŲ KAUPIMAS

J.-R. Lee, S.-L. Hsueh, H.-P. Tseng

Santrauka

Nors duomenų ryšys buvo plačiai taikomas informacinėse sistemose ir apibrėžiamas kaip pagrindinė priemonė duomenims automatiškai analizuoti ir įmonės žinioms pateikti, tačiau praktinis jo taikymas statybos pramonėje buvo mažai nagrinėtas. Šis tyrimas atliktas naudojantis praktine informacija, surinkta Tikslinių įmonės vartotojų paslaugų departamente. Pradedant duomenų rengimu, sprendimo medžio analize ir įmonės žinių pateikimu, yra atliekamas praktinis tikrinimas ir identifikuojamos prieš tai nenustatytos žinios. Atrasti du praktiniai duomenų ryšio kliuviniai: skirtingų informacijos sistemų informacijos atskyrimas ir pagrindinių informacijos atributų trūkumas. Tai yra didelis praktinės informacijos duomenų ryšio apribojimas, trukdantis naujai informacijai kurti. Tik tinkamai integruojant informaciją, galima pasiekti naudos iš praktinio duomenų ryšio taikymo.

Reikšminiai žodžiai: data mining, informacinės sistemos, statybos pramonė, įmonės žinios, paslaugos ir priežiūra.

Jen-Rong LEE. Assistant Professor. Jen-Rong Lee earned his Master and PhD degrees in civil engineering from the Ohio State University, USA. He holds Bachelor degrees in both civil engineering and financial law. Currently an Assistant Professor of Construction Project Management in the Dept of Construction Engineering, National Kaohsiung First University of Science and Technology (NKFUST).

Sung-Lin HSUEH. Assistant Professor. Sung-Lin Hsueh earned his PhD degree in the Dept of Architecture, National Taiwan University of Science and Technology in 2006. Currently an Assistant Professor of Project Management in the Dept of Interior Design, Tung Fang Institute of Technology. Concurrently, he is the Managing Director of SIN-YA International Engineering Consultants Inc. (Taiwan) engaging in developing real estate on the Chinese market.

Hung-Ping TSENG. MSc in construction management from the National Kaohsiung First University of Science and Technology in 2002. Chief of Quantity Control of Nishimatsu Construction Co.