# INTEGRATE DATA MINING WITH DECISION SUPPORT SYSTEMS TO SUPPORT ENGINEERING DECISION MAKING

**Jianfeng Guo[1], Jianping Zhang[2], and Qiping Shen[3]**

## ABSTRACT

Traditional engineering decision support systems (EDSS) cannot be well applied in the practice of construction management due to their inherent limitations in knowledge acquisition. In this paper, a new approach named DMDS is proposed, which integrates data mining technologies with decision support systems. DMDS aims to automatically extract knowledge from a large amount of historic data generated in engineering practice and then reason with the extracted knowledge to support engineering decision makings in construction. A prototype system named DM-EDSS has also been designed and developed to demonstrate the feasibility and effectiveness of the DMDS model. The findings demonstrate that at least the following benefits can be achieved: (1) providing a new approach for EDSS to automatically acquire knowledge from a large mount of engineering data; (2) enabling insights to be gained on factors that have impacts on construction problems; and (3) providing a uniform decision support process for different decision-making problems.

## KEY WRODS

integration model, data mining, decision support system, engineering decision making, reasoning mechanism

## INTRODUCTION

Due to the complexity and uniqueness of construction projects, most engineering decisions are experience-based, and may be affected by many factors such as decision makers' knowledge backgrounds, experience, emotion and characters, etc. In order to improve the efficiency and effectiveness of engineering decision makings in construction, many engineering decision support systems (EDSS) that try to reuse field experts' knowledge and experience to solve new engineering problems have been designed and developed. However, only few of them have been widely employed in the engineering practice because of their inherent limitations in knowledge acquisition. In traditional EDSS, the knowledge (e.g., rules or models) needed by their reasoning mechanisms have to be manually summarized from

[1] Ph.D. Candidate, Civil Engineering Department, Tsinghua University, Beijing, China. Phone +86-10-62778987, FAX +86-10-62784975, guojianfeng98@mails.tsinghua.edu.cn
[2] Professor, Civil Engineering Department, Tsinghua University, Beijing, China. Phone +86-10-62784975, FAX +86-10-62784975, zhangjp@mail.tsinghua.edu.cn
[3] Professor, Building and Real Estate Department, the Hong Kong Polytechnic University, Hong Kong, China. Phone +852-27665817, FAX +852-27645131, bsqpshen@inet.polyu.edu.hk

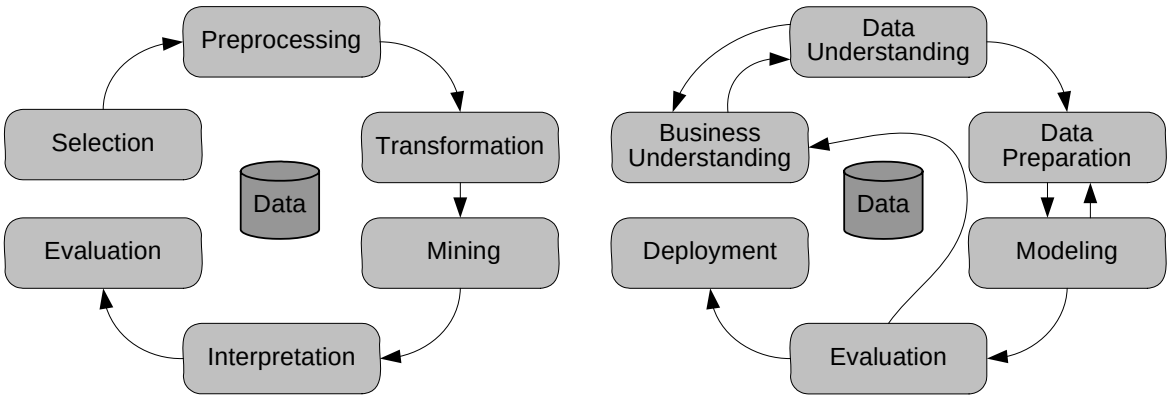
expert experience and field knowledge, transformed into appropriate forms and imported into their knowledge bases.

In order to overcome such inherent limitations in the traditional EDSS and to enhance their abilities and applicability, a new approach named DMDS that integrates data mining technologies with decision support systems is proposed in this paper. In this approach, knowledge is automatically extracted from a large amount of engineering data by employing data mining technologies, structured and transformed into appropriate forms, and finally retrieved and called by a hybrid reasoning engine in order to provide decision supports for engineering decision makings in construction. A prototype system named DM-EDSS has also been designed and developed to demonstrate the feasibility and effectiveness of the DMDS model. In the following sections of this paper, the basic concepts of data mining are introduced first, and then the detail of DMDS model and the implementation of DM-EDSS are presented in sequence. Finally, some potential benefits that can be brought by DMDS are summarized in the last section.

## BASIC CONCEPTS OF DATA MINING

Data mining, which is also referred as “knowledge discovery in database (KDD)”, was first coined to describe a new technique that could automatically extract knowledge from data in database (Piatetsky-Shapiro 1994). After more than ten years’ development, it has evolved into a science that integrates with databases, artificial intelligence, machine learning, neural network, statistics, pattern recognition, knowledge base system, knowledge acquisition, information retrieve, high performance computing and data visualization (Han and Kamber 2001). Piatetsky-Shapiro and Frawley (1991) regarded data mining as a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constrains, regularities) from data in database, and Grossman (1998) defined data mining as the semi-automatic discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data. Briefly, data mining aims to find knowledge from a large amount of historical data.

The process of data mining is an interactive, semi-automated process that begins with raw data and ends with outputs such as insights, rules and predictive models. In order to provide a common framework for carrying out data mining projects which is independent of both the industry sector and the technology used, many process models have been proposed to describe and standardize this process. For example, Fayyad et al. (1996) proposed a process model that consists of six phases, and the Data Mining Group proposed another process model named CRISP-DM (the Cross Industry Standard Process for Data Mining), as illustrated in Figure 1.




(a) Fayyad Process Model

(b) CRISP-DM Process Model

Figure 1: Different Process Models for Data Mining

Major progress has been made in the data mining community and many data mining techniques have been developed in the past ten years. According to the types of knowledge that can be mined, data mining techniques can be classified into the following categories (Han and Kamber 2001):

- Concept Description: present the general characteristics or a summarized high-level view over a set of user-specified data in database (Chen et al. 1996);
- Association Analysis: discover the association rules sharing attribute-value conditions that occur frequently together in a given set of data (Rao 2003);
- Classification and Prediction: find the common properties among a set of data and classify them into different classes based on their values in certain attributes (Chen et al 1996);
- Clustering Analysis: group a set of data (without a predefined class attribute) based on the conceptual clustering principle: maximizing the intraclass similarity and minimizing the interclass similarity (Chen et al 1996);
- Outlier Analysis: discover the outliers that don't comply with the typical patterns among a set of data.

Data mining have been successfully used for a variety of purposes in many industries such as banking, finance, insurance, retailing and business in order to reduce costs, enhance research and increase sales. Many commercial products such as DBMiner, SPSS, ESX and Nuggets have also been developed to facilitate the applications of different data mining technologies. In the construction industry, some researches have also been conducted to investigate the approaches and technologies that can be used to facilitate the applications of data mining techniques in construction (e.g., Buchheit et al. 2000, Leu et al. 2001, Morbitzer et al. 2003, Zhang et al. 2004). However, most of these researches only focus on some specific aspects in the lifecycle of construction process, and it's still a lack of a common framework that can supervise the general applications of data mining in the construction industry.

## DMDS INTEGRATION MODEL

Figure 2 illustrates the main process of the DMDS integration model. As shown, the DMDS model is composed of four phases, i.e., Problem Definition, Data Preparation, Modeling, and Reasoning. It is a repetitive process, which means that the previous phases may be continuously modified and redone until the following phase can meet user's requirements. In Figure 2, the solid arrows indicate the critical operation sequence of all phases, while the dash arrows refer to the possible modification operations.

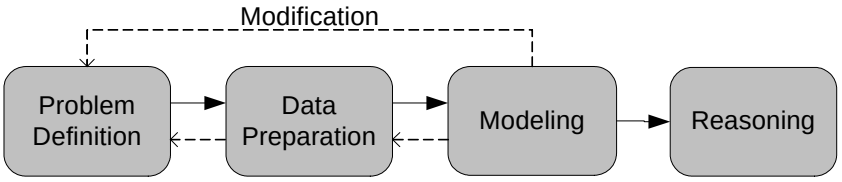



Figure 2: DMDS Integration Model for Construction Decision Making

### PROBLEM DEFINITION

Construction problems are usually very complex and may be affected by many factors. For example, the construction duration of a construction project may be affected by factors such as project scale, project complexity, cost requirement, material supply, construction level, weather condition and so on. It's unpractical to consider all possible factors for DMDS to describe a construction problem. In DMDS, construction problem is defined by a set of properties that correspond to such factors that have significant impacts to the under-study problem and can be easily observed and measured. Every property is further characterized by five attributes, as listed in Table 1.

Table 1: Attributes Used to Characterize a Property in DMDS

| Name | Description | Possible Value |
|---|---|---|
| Property Name | A unique name for indicating this property | |
| Usage Mode | Indicate how this property can be used | *Description; Input; Output* |
| Value Type | The type of value that can be assigned to this property | *Numerical; Enumerative* |
| Value Scope | Specify the boundaries for numerical property, or list all legal values for enumerative property | |
| Description | Additional information for user to understand the meaning of this property | |

All properties are organized into a definition vector $D_P$. According to the usage mode of every property, $D_P$ can be divided into three sub-vectors as following:

$$
\begin{aligned}
&D_P = [U_P, C_P, S_P] \\
&U_P = [u] = [u_1, u_2, \cdots, u_{n_u}] \quad (n_u = 0,1,2,\cdots) \\
&C_P = [c] = [c_1, c_2, \cdots, c_{n_c}] \quad (n_c = 1,2,3,\cdots) \\
&S_P = [s] = [s_1, s_2, \cdots, s_{n_s}] \quad (n_s = 1,2,3,\cdots)
\end{aligned} \tag{1}
$$

In Formula (1), $U_P$ denotes the description vector in which the usage mode of every element is *Description*, $C_P$ denotes the input vector in which the usage mode of every element is *Input*, and $S_P$ denotes the output vector in which the usage mode of every element is *Output*. Among them, only $C_P$ and $S_P$ are useful for DMDS when modeling and reasoning.

### DATA PREPARATION

The task of this phase is to prepare sufficient data for modeling. In general, the required data scatter in many different data sources such as databases, data files and electronic sheets. It is a difficult and time-consuming job to compile all these data into an integrated form. In order to overcome these difficulties, some researchers proposed using data warehouse to organize and store these data (e.g., Chau et al. 2002, Zhang et al. 2004). A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process (Inmon 1996). Data in a data warehouse are usually distilled from different transaction-oriented databases and organized by different subjects.

The DMDS model adopts a data warehouse that is designed for construction management by Ma (2004). This data warehouse contains 28 decision subjects that can be classify into eight categories, as listed in Table 2. Data corresponding with these subjects are organized in star schemas similar with Figure 3.

Table 2: Decision Subjects for Construction Management

| Categories | Subjects |
|---|---|
| Material Mgt. | Material Storage; WBS Material Usage; Project Material Usage; Material Check; |
| Equip. Mgt. | Equip. Purchase; Equip. Check; Equip. Usage; Equip. Consuming; Equip. Reparement; Equip. Alteration; |
| H.R. Mgt. | WBS H.R. Usage; Project H.R. Usage; Wage Mgt.; Bonus Mgt.; Training Mgt.; |
| Schedule Mgt. | Project Schedule; WBS Schedule; |
| Quality Mgt. | Quality Inspection; Project Quality Accident; WBS Quality Accident; |
| Safty Mgt. | Project Safty Accident; WBS Safty Accident; Equip. for Safty; Safty Training; |
| Cost Mgt. | Project Cost; WBS Cost; |
| Technology Mgt. | Technology Preparation; Drawings Mgt.; |

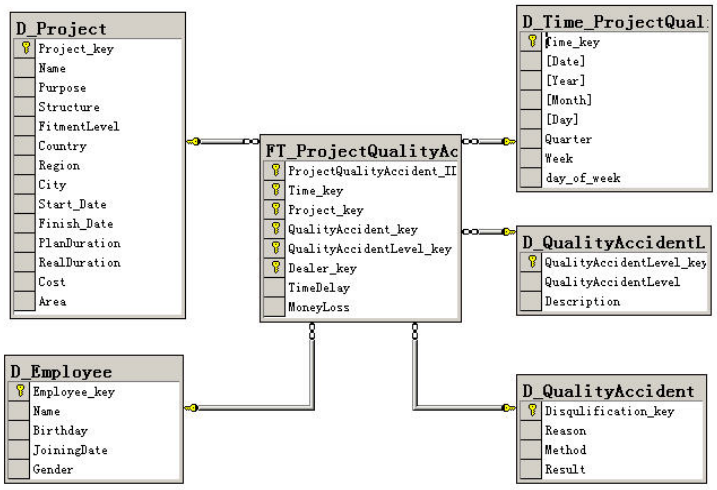



Figure 3: Star Schema for "Project Quality Accident" Subject

Data in data warehouse should be further organized into CSV (Comma-Separated Value) format before modeling. CSV format is a text-based data format that uses comma as the separator of different values. It's supported by many software packages such as Microsoft Excel and Microsoft Access, etc. DMDS model uses CSV format as the basic data format when modeling and reasoning, which facilitates the data exchange and sharing between DMDS and other software packages.

## MODELING

The task of this phase is to build predictive models. According to the difference in reasoning mechanisms, predictive models can be classified into different types, such as association model, clustering model, naive Bayesian model, neural network, regression model, rule set model, sequence model, support vector machine model and tree model, etc. DMDS adopts Predictive Model Markup Language (PMML), a XML-based specification proposed by Data Mining Group, to describe the logical structures of all these predictive models.

Many existing algorithms can be employed to build predictive models. However, most of these algorithms have their special requirements on data. For example, algorithms for neural network require all input and output are numerical values between 0 and 1, while ID3 algorithm for tree model requires all input and output are enumerative values. In order to meet their special requirements of different algorithms, DMDS adopts a three-step modeling process as illustrated in Figure 4. In step (1), original data (usually a CSV file prepared by the previous phase) is transformed into different forms (processed data) according to the requirements of different algorithms. In step (2), modeling algorithms build predictive models (rough model) on processed data. In step (3), rough models are transformed into final models in order to correspond with original data.

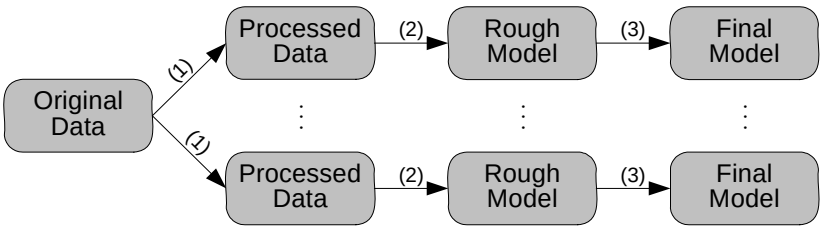



Figure 4: Modeling Process in DMDS

Final models still need to be further tested and validated in order to determine their accuracies. Accuracy of predictive model is an important attribute that will be used to calculate the reliability of solutions in reasoning phase. Only those models that are of satisfied accuracies can be accepted and stored for further use.

**REASONING**

The task of this phase is to provide decision support to users by reasoning with predictive models. DMDS adopts a hybrid reasoning engine that is composed of three independent reasoning engines (i.e., rule-based reasoning engine, model-based reasoning engine and case-based reasoning engine) and an integration engine, as illustrated in Figure 5. When a new problem is inputted, every reasoning engine selects appropriate knowledge (rules, models or cases) from knowledge base, respectively, and generates a rough solution for current problem based on the selected knowledge. All rough solutions generated by three reasoning engines are compiled into a final solution by the integration engine.

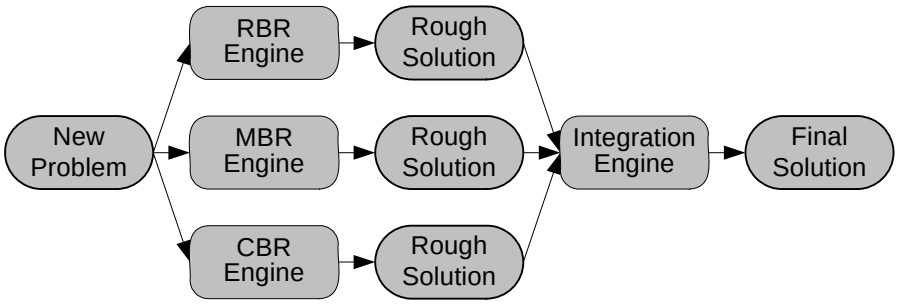



Figure 5: Hybrid Reasoning Engine in DMDS

A difficult but crucial question in this phase is how to integrate several rough solutions into a single one. In DMDS, every rough solution gets a weight calculated by the accuracies of reasoning knowledge. The final solution is calculated by the following two principles:

- If the predictive value is a numerical value, the weighted average of all rough solutions is selected as the final solution.
- If the predictive value is an enumeratic value, the value with the biggest weight is selected as the final solution.

## IMPLEMENTATION OF DM-EDSS

In order to validate the feasibility of the proposed DMDS integration model, a prototype system named DM-EDSS is designed and developed in this research. DM-EDSS aims to provide decision support to construction management. It predefines a set of decision subjects that are frequently met in engineering practice, and provides several tools to facilitate the processes of modeling and reasoning.

DM-EDSS adopts a Client-Server structure (see Figure 6) to enables several users to simultaneously operate on this system through LAN or Internet. On server module, DM-EDSS maintains three repositories: (1) data warehouse that contains all predefined decision subjects and corresponding datasets; (2) knowledge base that contains all predictive models generated by DMDS and other knowledge imported from outside; and (3) transaction database that contains some transaction data such as users' information and login information, etc. The main task for server module of DM-EDSS is to accept queries from different clients and feedback query results to corresponding clients.

The client module of DM-EDSS provides a friendly user interface (see Figure 7) for users to access the data stored in the server. Every client also contains a local database that stores some local information such as local settings, chat messages and operation log, etc. Through client module, users can obtain decision supports by reasoning with existing knowledge, generate their own predictive models and share them with other users, or create new decision subjects for their special demands. DM-EDSS also provide an online communication toolkit to facilitate the communication among different users.

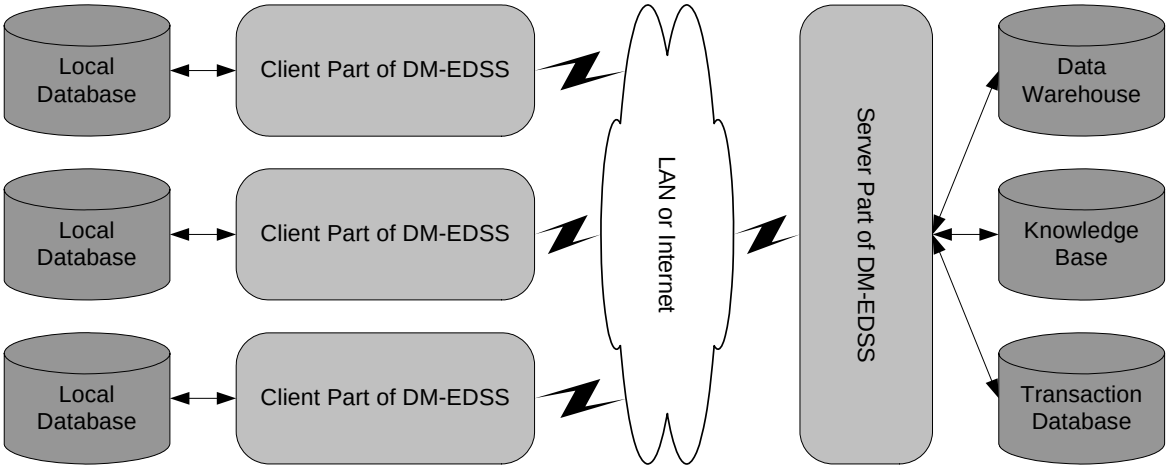



Figure 6: System Architecture of DM-EDSS

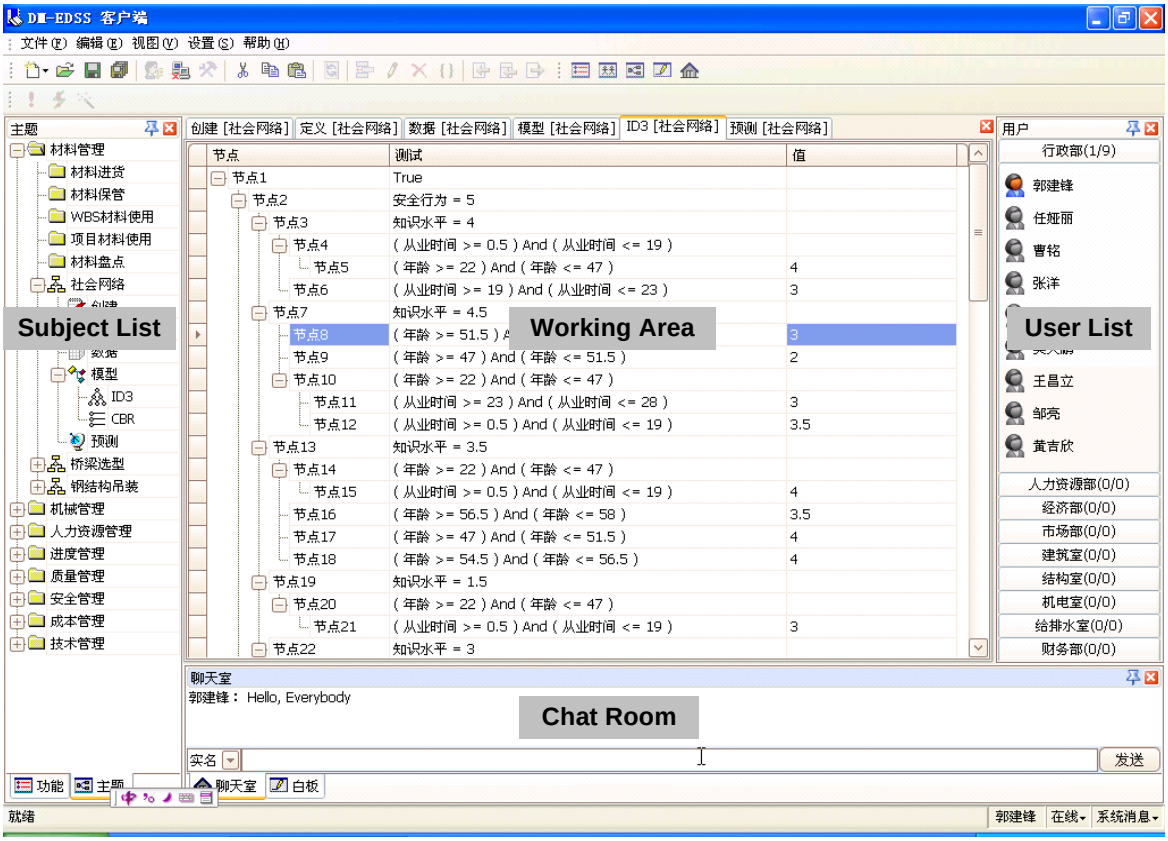



Figure 7: Main Interface of DM-EDSS

## CONCLUSIONS

This paper presents a new approach to facilitate the engineering decision makings in construction. The proposed approach aims to provide decision supports for engineering decision makings in construction by directly reasoning on a large amount of historic data. Many potential benefits can be achieved by this approach:

- Breaking through the inherent bottleneck of traditional EDSS in knowledge acquisition so that their functions and applicabilities can be greatly improved.
- Revivifying the large amout of engineering data accumulated by past projects and enabling insights to be gained in the potential deep knowledge hidden behind these data.
- Providing a uniform process to support different engineering decision makings in construction.

Data mining is still a fairly unfamiliar thing for the construction industry. The efforts in this research are expected to contribute to further development of corresponding techniques and technologies for applying data mining in the engineering practice.

**ACKNOWLEDGMENTS**

This research is supported by the Natural Science Foundation of China (Approved No. 50478015) and the Center for Information Technology in Construction between Tsinghua University and the Hong Kong Polytechnic University.

**REFERENCES**

Buchheit, R.B., Garrett, J.H., Lee, S.R. and Brahme, R. (2000). "A knowledge discovery framework for civil infrastructure: A case study of the intelligent workplace." *Engineering with Computers*, 16(3-4) 264-274.

Chau, K.W., Cao, Y., Anson, M. and Zhang, J.P. (2002). "Application of data warehouse and Decision Support System in construction management." *Automation in Construction*, 12 213-224.

Chen, M.-S., Han, J. and Yu, P.S. (1996). "Data mining: an overview from a database perspective." *IEEE Transactions on Knowledge and Data Engineering*, 8(6) 866 pp.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases." *AI Maganize*, 17(3) 37-54.

Grossman, R. (1998). "Supporting the Data Mining Process with Next Generation Data Mining Systems." *Enterprise System Journal*.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techiniques*, Morgan Kaufmann Publishers, San Francisco, USA.

Inmon, W. H. (1996). *Building the Data Warehouse* (2nd Edition), Johm Wiley & Sons, New York, USA.

Leu, S.-S., Chen, C.-N. and Chang, S.-L. (2001). "Data mining for tunnel support stability: neural network approach." *Automation in Construction*, 10(4) 429-441.

Ma, T.Y. (2002). *The Application of Data Warehouse and Data Mining in Construction Enterprises*. Master Diss. Civil Engrg. Dept., Tsinghua Univ., Beijing, P.R. China.

Morbitzer, C., Strachan, P. and Simpson, C. (2003). Application of data mining techniques for building simulation performance predictive analysis. The 8th International IBPSA Conference, Eindhoven, Netherlands, 911-918.

Piatetsky-Shapiro, G. (1994). "Knowledge discovery in databases: progress report." *The Knowledge Engineering Review*, 9(1) 57-60.

Piatetsky-Shapiro, G. and Frawley, W.J. (1991). *Knowledge discovery in databases*. MIT Press.

Rao, I.K.R. (2003). Data Mining and Clustering Techiniques. DRTC Workshop on Semantic Web.

Zhang, J.P., Ma, T.Y. and Shen, Q.P. (2004). Application of Data Warehouse and Data Mining in Construction Management. Proceedings of Xth International Conference on Computing in Civil and Building Engineering, Weimar, Germany, 66-77.