S13-2

Factors Clustering Approach to Parametric Cost Estimates And OLAP Driver

JaeHo, Cho[1] and BoSik, Son[2] and JaeYoul, Chun[3]

[1] PhD Candidate, School of Architecture Engineering, Dankook University, Yongin, Korea
[2] Professor, School of Architecture Engineering, Namseoul University, Chonan, Korea
[3] Professor, School of Architecture Engineering, Dankook University, Yongin, Korea
Correspond to cjhace@naver.com

ABSTRACT: The role of cost modeller is to facilitate the design process by systematic application of cost factors so as to maintain a sensible and economic relationship between cost, quantity, utility and appearance which thus helps in achieving the client's requirements within an agreed budget. There are a number of research on cost estimates in the early design stage based on the improvement of accuracy or impact factors. It is common knowledge that cost estimates are undertaken progressively throughout the design stage and make use of the information that is available at each phase, through the related research up to now. In addition, Cost estimates in the early design stage shall analyze the information under the various kinds of precondition before reaching the more developed design because a design can be modified and changed in all process depending on clients' requirements. Parametric cost estimating models have been adopted to support decision making in a changeable environment, in the early design stage. These models are using a similar instance or a pattern of historical case to be constituted in project information, geographic design features, relevant data to quantity or cost, etc. OLAP technique analyzes a subject data by multi-dimensional points of view; it supports query, analysis, comparison of required information by diverse queries. OLAP's data structure matches well with multiview-analysis framework. Accordingly, this study implements multi-dimensional information system for case based quantity data related to design information that is utilizing OLAP's technology, and then analyzes impact factors of quantity by the design criteria or parameter of the same meaning. On the basis of given factors examined above, this study will generate the rules on quantity measure and produce resemblance class using clustering of data mining. These sorts of knowledge-base consist of a set of classified data as group patterns, of which will be appropriate stand on the parametric cost estimating method.

Keywords: Data Mining, OLAP, Parametric Cost Estimates, Quantity, Unsupervised Clustering

1. Introduction

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.(Jiawei Han,2001) In particular, the industry fields using a quantity of data shall find out the method to automatically analyze, classify and summarize data, to automatically discover and characterize trends in it, and to automatically flag anomalies. Data mining is the natural results of evolution in the IT fields. Database industry developed the functional evolution path including data collection and database creation, data management(data storage, retrieval, and transaction processing of data), and data analysis and understanding(involving data warehouse and data mining). In sum words, we can say that data mining is the extraction of knowledge from data.

Data mining can be applied for cost estimates in AEC industry. Historical data will be the most critical sources for cost estimating method. The major objectives of cost estimates are to support the decision of clients by cost planning as well as to improve accuracy, in the traditional points of view, by technical estimating using information from given design. For such cost planning, it is required to identify influencing factors related to design, to recognize the data structure associated with those factors, and finally to use the patterns in them. Useful information extracted through a pattern analysis can be used as the knowledge for cost planning and expanded for the final decision making and business strategy. For data mining taking the initiative in such application fields, data warehouse implementation including data filtering and integration will become the important part of preprocessing stage. In addition, the data warehouse provides the OLAP(On-Line Analytical Processing) tool enabling the interactive analysis on the multi-dimensional data in various levels and it is offering effective data mining. With integration of various kinds of data mining functions including classification, prediction, correlation or clustering into OLAP, interactive knowledge mining on diverse abstraction levels is more enhanced, on accelerating pace in information technology.

In the same manner, it will be significant preprocess that is recognition of historical data structure for the data warehouse platform in the application field as cost estimates. For enabling a cost modeler to apply such data in their own application system, the data mining technique shall be adopted to identify useful knowledge and patterns related to cost information, which varies on diverse kinds of design requirements. Due to difference of cost estimating requirement by design level, this study aims to suggest a knowledge-based model for quantity prediction in the aspects of parametric estimates, and it can be utilized on each its demand.

The knowledge base mentioned above is implemented on resemblance class using factors clustering. These impact factors are deduced by OLAP technology which is operational driver support to analysis at the multi-dimensional views.

2. Problem Issues

Cost estimating process description should include the concepts of tasks, resource requirements, quantities and cost.(G.Tesfagaber, 2002) Cost estimate can be assigned by evaluating how much material and resource each item requires. A rule-of-thumb quantity take-off in the Preliminary Cost Estimates can be extrapolate from historical data on similar design information refer to drawings or BOQ, and finally calculates the total cost by multiplying the unit price. This approach may be completely different as the case may be in a cost estimating model. In other words, when using knowledge-based computer aided automatic design, structural quantity is automatically measured after making drawings based on the relationship of objects and artificial relation rules. (Structure design research has been actively examining this issue.) Subsequent cost estimating process, starting with the automatic measurement, applies unit price using cost knowledge. The unit price is universally applicable to all case with the same rule. The following figure 1 shows the general process of cost estimating concept.

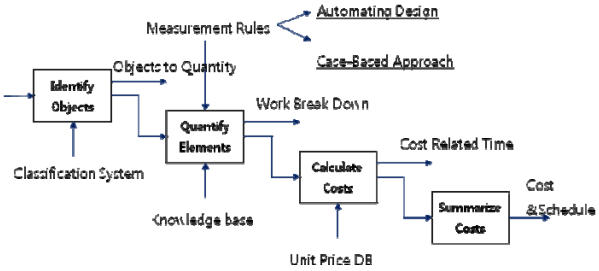

Figure 1. General Process of Cost Estimating Concept

Representative cost estimating approaches using the process above will be method of quantity measure by automation design or statistical case-based database. This study examines each approach relating to research and consider what to answer more effective approach in Chapter 5, Better Way of this paper.

The cost estimating method which is the main theme of this study focuses on the statistical case-based application for quantity prediction, and this approach is our present concern. As the problem issue, the argument here describes what kind of attribute dimensions can be established for information structure related to quantity prediction or measure. In this aspect, the attribute dimension of data can be considered as the impact factor for quantity prediction. The following figure 2 describes a 3D dispersion on quantities case with the 4D attribute.

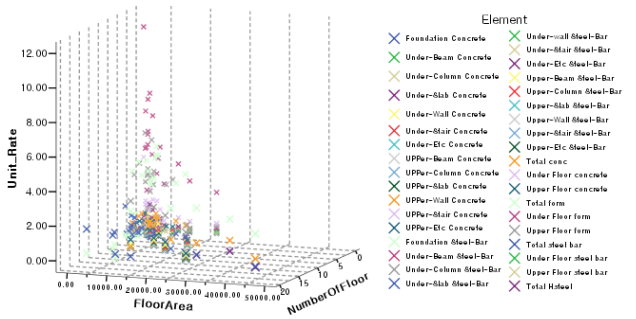

Figure 2. 3D dispersion of Quantities Case(704 points) (Concrete, Steel, form by Element & Space)

Dispersion of the instance shows the quantities case of concrete, steel, and form on the multi-dimension.(gross floor area, number of floor, and unit rate level) The data dimension on the facility level(by building element and geographic space type) is described on 3D space.

Cost estimates involve different requirements depending on to what extent the design information gets concrete form. Generally cost estimating process in early design stage is defined as follows: feasibility, preliminary appraisal, and approximate estimate. It is described above that referring to the design information available in each design stage facilitate cost estimates within acquired information level. According to the related research, information factor accounts for about 14% of the estimate accuracy.(Garold D. Oberlender, 2001)

As we recognize from Figure 2 above, the class can't be clearly recognized by the attribute of data if we only have simply 4D-dimention. It is required to analyze by more diverse points of view or divide into more detail level for data classification and clustering (e.g., related to the data dimension: plan design shape, facility type, facility service, space attributes, structure type, special characteristics, and features, etc.)

Since the impact factors get a concrete form by design stage, cost estimates need the integrated, unified, and pre-structured database that can be used by stage on cost estimating requirement level. Accordingly, this study will design the database structure for impact factors in consideration of the following topics, on the basis of quantities case data.

1. What kinds of impact factors relate to the design information in terms of building quantity type or value?
2. Is it possible to reflect impact factors by design requirements that are extracted from each cost estimating and design process?
3. How can the best resemblance case be inquired?
4. Is it possible to do data abstraction, hierarchical structure, classification, and clustering by impact factor?
5. Is it possible to consider the dynamic or static factors, which is changed as time series?
6. Are any specific patterns found in changeable value by impact factor or among impact factors?

The multi-dimensional database on impact factors is implemented in consideration of the above topics. This mapping out between the considerations and the destination database as the preliminary process for data warehouse enables us to employ OLAP technology that is very effective for multi-dimensional information analysis approach.

3. Research Scope and Methodology

Cost estimates are divided into 3 levels in the early design stage as stated above. (e.g., Feasibility, Preliminary Appraisal, Approximate Estimate) The cost estimating model in this study is based on parametric cost estimating approach. A parametric cost estimating model can support various kinds of cost planning in accordance with impact factors related to design information as the parameter. The estimating approach in this study is a prediction for quantity values using the rules produced by clustering with impact factors and resemblance classes.

This study performs the process to identify the useful knowledge and the information patterns related to historical case through the data warehouse on OLAP and Unsupervised Clustering in data mining. The OLAP technique would be given a chance to analyze impact factors by interactive query using a data cube. Unsupervised Clustering suggests a classification for categorical attributes that is used impact factors. Using the Unsupervised Clustering of ESX, a commercial program, instances are clustered into resemblance classes and the average and distribution of each class will be calculated. Furthermore, the class rule models can be built.

The scope of data analysis is based on the existing historical case such as unit ratio of concrete, form, steel bar and H-beam quantity in the building structure and the ratio of those constituent items used as the dependable value.

Impact factor analysis is limited to the solely design items associated in quantity data. External factors that are the project information having impact on quantity value, such as contract type, cost index, location index and quality level, and etc. are excluded from the scope of this study.

In the Chapter 5, estimating approach is identified by the techniques available at present state through the current research and it analyzes the direction for the progresses of this study. Moreover, we will examine how to apply the parametric cost estimating model into 3D/4D technique on BIM field for which the researchers have been enthusiastically investigating at present.

The purpose of this study is how to implement knowledge base applicable model in parametric cost estimates and to mining the corresponding patterns of the cost parameter. The further study will focus on the application system implementation of parametric cost estimating model and information delivery interface technique for interoperability on BIM design basis.

4. Relevant Research

The initial concept of pattern recognition was started from the arts field. The recent pattern recognition is

observed in both data mining and artificial intelligence. The Patten recognition includes data mining exploiting information hidden in data and sensing interface to external objects. It is the extensive research topic under progressive research with tremendous attention as a key technology of next generation.

Although the technical term called pattern recognition is not usually used in the research of cost estimates or the computer aided automatic design for structure in the ACE industry, we can easily recognize that a number of research have been conducted on the pattern recognition principle. For example, these researches have been continued in general pattern recognition using regression or probability, neural network or fuzzy logic, etc.

In particular, the recent trend in the automatic design field tries to establish the relationship by semantic analysis on object-based data. The representative example is ontology langue.(H.Kim, 2007) This approach above has also been using pattern recognition methods in the abstract aspect and connects with object-relational data mining (e.g., relational, transaction, object-oriented, object-relational and data warehouse mining system according to the classification of data mining, Data Mining, Jiawei Han, 2001).

Pattern recognition firstly selects an object, extract its attributes through preliminary analysis and selects diverse models for pattern analysis. The next stage of pattern recognition is learning stage, and then it makes a determination of a class or a category having the attributes.

Pattern recognition model is largely classified into statistical approach, neural network approach and structural(semantic) approach. (Pattern recognition, Han Hakyong) It is not easy to definitely identify these kinds. This study doesn't aim the clear classification on pattern recognition. However, the general classification of pattern recognition is required to determine the potential direction for cost estimating research. Establishing the research direction as explained above will help to find out the most suitable system for the researchers focusing on cost estimating methods.

It is meaningful to select the semantic pattern recognition if the distinct structural information related to a pattern is available in database design. On the contrary, statistical pattern recognition can be applied if such semantic structural information may not be acceptable. A number of practical issues applying pattern recognition are going to be in these two approaches. We noticed the research that had been conducted in every sense of pattern recognition in cost estimating application.

In case of the classification for the research in the quantity prediction process, there are statistical approach by a similar case and automatic design approach by quantity take-off. The cost estimating method by the statistical case approach, to which we are very familiar, is classified into the two types, one is based on a cost data directly and another is on a quantity case indirectly. The parametric cost estimating model using the statistical approach based on cost rate was one of the benchmark for this study.(Kan Phaobunjong, 2002)

A variety of techniques for the statistical approach includes regression analysis, correlation analysis, probability distribution and cluster analysis, etc.

In other case, automatic design can be represented with the structural information consist of regular grammar on objects. For example, the structural information in compliance with correlations of objects, constraints, and grammatical inferences among the object, make it possible to automatically draw the object design. However, this research field is under the initial stage, because it is very difficult to learn the structural rules.

Another application field related to cost estimating research is BIM(3D/4D) using object database. The cost estimates in the early design stage using BIM have the interface between the building product model and the cost data on the statistical technique-based or on the case-based unit modules relatively simple ways. This objective estimating approach is required to study further in-depth due to the limit of support the detail in the early design stage.

The last cost estimating approach uses neural network as one of the pattern recognition fields. The pattern recognition using neural network is a kind of approach mixing statistic and semantic. The neural network theory is also applied for the data classification and prediction in data mining.

The rest of applications is these sorts of theories, such as Fuzzy Sets, Generic Algorithm, CBR, Discriminant Analysis, Principal Component Analysis, and OLAP, etc. in preliminary cost estimating model. The OLAP in cost estimates was found in construction stage.(Hao Howard Nie,2007) These approaches are also as data mining to combine several models, basically use statistical approach.

Table 1 shows the general application fields to cost estimates in the early design stage. Thus, in the aspects of pattern recognition along with data mining theory, these research can be classified into automatic design using semantic model, parametric cost estimating using statistical model, and the case base using data mining. And Data mining may be classified by type of database, knowledge or application subject to mining.

5. Better Way

In consideration of future development on the information technology, computer aided design will be more promising in all aspects of accuracy or performance in cost estimates. However, the research on this field doesn't achieve remarkable result, as still being in the primitive phase. In terms of applicability and efficiency of cost estimates, researches on existing statistical approach and data mining field will be continually extended with the advantage of accommodation in user requirements, simple user interface in the estimating process, and free of time consuming.

In particular, CBR and Neural Network in the data mining will be expected to more study in the cost estimating model. CBR and Neural Network are sub-fields of data mining which used to perform data classification and prediction.

For the recent cost estimating techniques in BIM-based in the early design stage, the statistical approach is applied for it to calculate the cost. (e.g., Its system use fundamentally RSmeans' square foot costs as the database.)

The future cost estimating technology will be conducted by connecting object database consist of abstractive building information. In other words, the cost estimates using BIM in early design stage can make prediction for quantity through the reverse engineering approach. Its conceptual idea can be inferred from the thesis of Min-Yan Cheng, 2001.(Concept of GIS-Based Quantity Takeoffs Algorithm) We can't definitely say what kind of method is the best among cost estimating models. However, when we deduce through the technical status up to now, OLAP technique in the data mining includes the classification and prediction function of CBR or Neural Network. This makes it possible to be an elaborate information analysis.

Thus, OLAP technique can be used as the operative engine to query the best resemblance case within the abstractive quantity information. And then, it will calculate the cost by connecting the information to BIM model and unit prices in the early design stage. However, what has to be solved in this system is scarcity about the similar data on OLAP engine.

Therefore, we need to find a way out of it with some alternative solution in this study. The following figure 3 is conceptual mapping for the parametric cost estimates using OLAP technique as the knowledge extract driver and linking IFC information in BIM design field. The knowledge-based model is accumulated through factors evaluated to pattern using data mining. Such knowledge-base supports various kinds of cost planning interface to BIM having the geographic information, which is needed to employ the parametric impact factors in the cost estimating system.

Table 1. Previous Research (The broad lines on the table correspond to approaches for a better way)

Kinds of Pattern Recognition (With Data Mining)	1. Semantic Approach ∈ Data Mining	2. Statistical Approach ∈ Historical data Based	3.Neural Network Approach ∈ Data Mining
Cost Estimating Principle	- Computer Aided Design - Conceptual Structure Design	- Parametric Cost Estimates (Based on Cost Data) - Interface to BIM (3D/4D, Early Design Stage)	- Mix Statistic and Semantic - Case Based(Based on Cost Data) :Classification, Prediction,…,etc,
Operation Method, Driver or Engine	- Object-Oriented Relationships - Ontology Model - Design Rule, Semantic Object, UML - Knowledge Base (Member Properties) - Object-Oriented Data + Case Based Reasoning	-Cost/Usable Floor Area -Quantity /m2 of Floor Area :Unit Rates for Constituents (Concrete, Steel, Formwork) -Regression -(Semilog, Multiple) Regression - Discriminant Analysis - Principal Component Analysis	- Fuzzy Sets - Neural Network - CBR, + Generic Algorithm, etc.

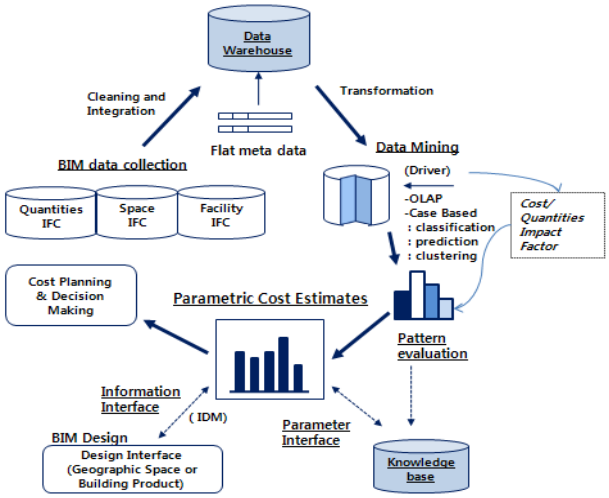

Figure 3. Better Way of Cost Estimating Model

6. OLAP(On-line Analysis Processing) State

OLAP was firstly used by E.F. Codd in 1993. It is the concept against On-Line Transaction Processing(OLTP). OLAP has been securing its position as the critical factor for data approach strategies in data warehouse. OLAP is defined as the "the process that the final clients directly approach to the multi-dimensional information, analyze information in interactive ways and use for decision making.(Jo Jaehui/Bak Seongjin, 1996)

OLTP system focuses on 'What' in the recording system, whereas OLAP system focused on 'Why' in terms of application of collected data for decision making. This study firstly aims to identify what kinds of impact factors inducing a variable of quantity with OLAP technique. The quantity case to be queried includes design information related to impact factors and it can be analyzed and compared in various ways with a hierarchical structure by the dimensions.

7. OLAP Design

Data warehouse and OLAP are based on the multi-dimensional data model, which considers it as data cube. This chapter suggests how data cube of quantity case design to n-dimension.

Data cube is defined by dimensions and facts. The dimension is considered just like time scale to record a certain object and each dimension has a hierarchy table describing the detail of attributes.

The following figure 4 shows the 4D data cuboid for quantity data. The cuboid term is used in the same way as data cube. The cuboid lattice is extracted from the group of given dimension. Each cuboid describes the summary of other stage, that is, data by 'Group By'(summary of dimensions as other subsets).

The figure 4 shows the cuboid prototype for 4-dimentional information that across on facility, shape, feature, and composition dimension.

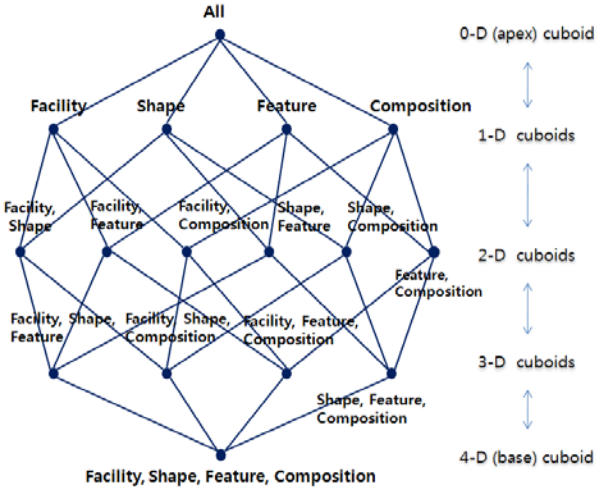

Figure 4. Concept of Cuboid Dimension

8. Operation of Multi-Dimensional Data Model

Data in multi-dimensional model comprises of several dimensions, and each dimension includes various abstraction levels defined by concept hierarchy. Such the composition provides the clients with the flexibility to see data in various points of view. Typical multi-dimensional operations include Roll-up, Drill-down, Slice, Dice, Pivot and Drill-across operations.

OLAP calculates sums, averages, ratios and variances on the cross points with all hierarchical dimensions. It can be implemented on the basis of query model adopted the StarNet schema in this study. (see Figure 5)

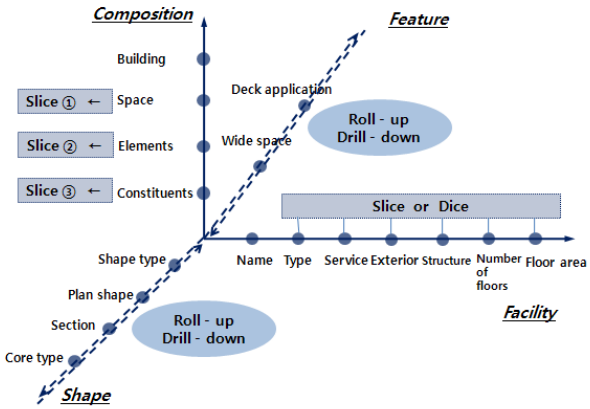

Figure 5. A Starnet Query Model for Querying Muti-dimensional Database

Starnet model comprises of radial lines stretching from a center point, and each line shows a concept layer for a dimension. Each abstraction level in the hierarchical structure is called footprint. Footprint means the segmentation unit used in OLAP operation including Drill-down and Roll-up.

Roll-up and Drill-down: Roll-up operation performs an aggregate on data cube using the reduction of dimension on the concept layer in a certain dimension and Drill-down performs the opposite operation to Roll-up. This study applied Roll-up and Drill-down on the shape and the feature dimensions.

Slice and Dice: Slice operation selects one dimension in a given cube and so it produces a sub-cube. Dice operation defines a sub-cube by selecting more than two dimensions.

The composition dimension in this study performs Dice in each dimension. Facility dimension selectively performs Slice and Dice along with other two dimensions to order.

9. Fact Constellation

The most common Data Schema for a data warehouse is basically a multi-dimensional model. Such a model can exist as shape of a Star Schema, a Snowflake Schema or a Galaxy Schema. Star Schema, the most common modelling paradigm, is made up of a big center table(fact table) containing a quantity of data without redundancy, and a relatively small dimension table set for each dimension. Complicated applications may require several fact tables sharing dimension tables. This kind of Schema is considered as group of the stars so that it is called a Galaxy Schema or a Fact Constellation.

The following figure 6. describes the Fact Constellation Schema suggested in this study. It shows two fact tables, the Space Fact Table and the Elements Fact Table. Each fact table illustrates unit value on concrete, form and steel quantity. It additionally includes the ratios of constituent items in each quantity, too.

The Fact Constellation Schema is actually used for modeling a number of correlated subjects. In other words, the information required by each design stage for

cost estimates may include total quantity of facility(or building), quantity by geographic space, and quantity by elements. Accordingly, The Fact Constellation can be defined by interrelated subjects in the composition dimension(by facilities, geographic spaces, and elements).

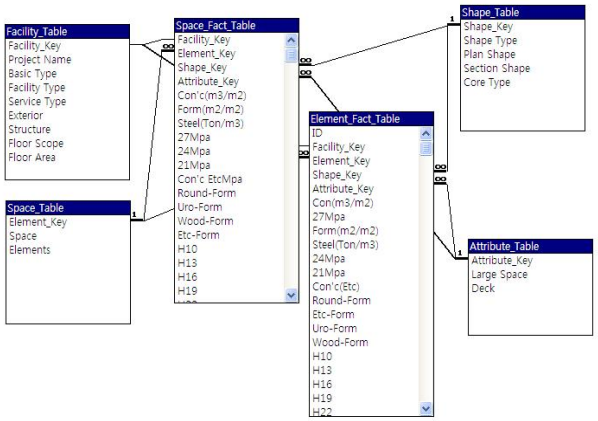

Figure 6. Fact Constellation for Quantity Database

10. Clustering Approach in Parametric Cost Estimate

It is required to logically calculating process for cost estimates after the query process in the cube dimension It is called the OLAP join in this study. Parametric cost estimates can be performed by multiple query from cubes in each dimension. The total cost is summed up by linking the relevant unit cost. For example, for the quantity of concrete and its constituents on OLAP cube, the cost is calculated by linking the external database of the relevant unit price. (e.g., 21mpa, 24mpa, 27mpa, etc-mpa) The OLAP join process in parametric cost estimating model adopts the concept, where the quantity cube is used as pivot tables, thereby join into external database for applying unit prices to corresponding order. The query by in each dimension could be performed using OLAP engine. However, there is an issue to be solved in the structural features of OLAP system due to the scarcity of data queried in the multi-dimensional aspects.

While the issue identified above can be solved by securing the sufficient quantity of data, the limit related to collection of the data still exists.

Accordingly, the factors putting impact on quantity will be analyzed using OLAP operation tools, and thus quantity prediction rules can be produced based on the resemblance by impact factors. In other words, it is required to find out the alternative approach which can be utilized to parametric cost estimates based on the representative information, where the unique or some scarcity data are normalized to a resemblance class, through enhancing data mining(e.g., clustering approach).

11. OLAP System Implementation

This study aims to implement the system enabling clients to perform query on various kinds of information and trace the impact factors using the OLAP driver.

The figure 7. adopted 'OlapCube Writer of Adersoft', the commercial software as the OLAP cube. Operation of the dependant data stored as the fact value basically applies the average calculation. For example, the average of ratio by unit of concrete, form and steel-bar are applied, along with each elements comprised of a work.

The database contains 20 government buildings quantity case recently in completion. All the quantity information and the determinants by facilities, geographic spaces, and elements are stored in the Fact Table.

(The number of total fact data is 280.)

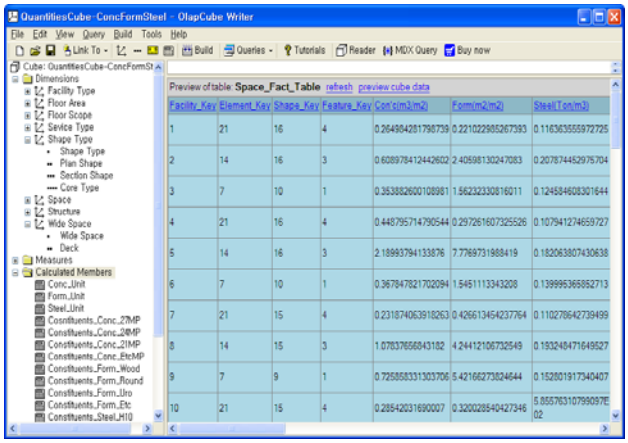

Figure 7. System Implementation by OlapCube Writer

12. OLAP Report

The various kinds of cases can be queried using the implemented OLAP driver. The figure 8. shows the data analysis screen by facility type depending on the floor area for unit quantity(concrete' ratio values in upper stories). Furthermore, the unit ratios of the forms and steels can be queried. Query by the elements is also enabled. Such all sorts of reports are not demonstrated in this study because of the limited space.

However, the preliminary analysis is allowed for data mining to be discussed in the Chapter 14. For deducing the meaningful knowledge through data mining, where the OLAP can make an analysis of the determinants for each dimension, further need to have the preliminary test. These determinants are evaluated as the impact factors if they have the valid different value within the attributes of the facts. And after, they have to do clustering on resemblance class for factors, such as facility type, floor area, plan shape, etc., which is containing a steel or a concrete ratio in upper stories. The figure 9. below shows the pie chart by determinant of each dimension for RC structure(steel's ratio in concrete) on upper stories. The following results are deduced from the chart above.

1. The steel' ratio in concrete by facility type is relatively high for the general government building and low for the government agency office. The ward office building and the post office have similar ratio values.

2. For floor area, the wider the floor area was, the higher the steel' ratio was, in general.

3. On the whole, the high stories(6~11 stories) had more higher than the low stories(1~5 stories).

4. The steel/concrete ratio may be relatively different by plane shape.

5. The steel/concrete ratio by core type didn't show a significant difference.

6. In the case of feature attributes, the facilities with wide spaces had relatively lower steel/concrete ratio.

The OLAP analysis results can be acceptable when it has statistically significant only through the analysis on basis a quantity of data. However, this study will be mean to adopt as just a practice model rather than a statistically significant test due to the limit of data collection.

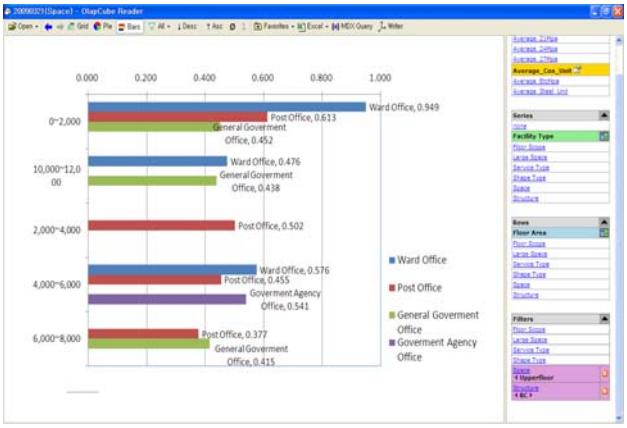

Figure 8. Comparison of Concrete Unit Quantity (by the Floor Area on Upper Stories)

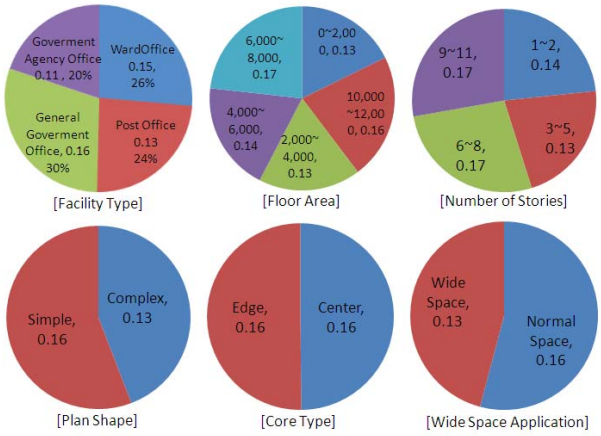

Figure 9. Comparison & Analysis of Steel Unit Ratio (by Factor: Facility, Floor Area, Number of Floor,… etc.)

13. Data Mining: Clustering

We performed the preliminary analysis on steel/concrete ratio in the Chapter 12. The 5 items were deduced as the factors having impact on steel/concrete quantity (facility type, floor area, number of stories, existence of wide space, plane shape).

Accordingly, this chapter will analyze the pattern of impact factors through the data mining. In other words, we will identify by what kinds of patterns for steel/concrete ratio can make a cluster.

This study adopted the ESX approach applying the Unsupervised Clustering of the categorical attributes to the model. The ESX is an exemplar-based data mining tool that builds a hierarchy concept to generalize a data. The ESX can also perform a Supervised Learning as case-based data mining tools. For the unsupervised clustering, ESX incorporates a globally optimizing evaluation function that encourages a best instance clustering. The primary data structure used by ESX is a three-level concept hierarchy. (Root-level, Concept-level, Instance-level), The concept-level nodes store summary statistics about the attribute values found within their respective instance-level children.

Class resemblance scores are stored within the root node and each concept level node. These provide a measure of overall similarity for the exemplars making up individual concept classes. Unlike the K-means algorithm, it is not required to make a determination about the total number of clusters to be formed. (Detail about the class resemblance computation are not suggested in this chapter. Refer to Data Mining, Richard J, Roiger, 2003)

To this end, iData Analyzer(iDA), the commercial software, is applied. In addition, iDA can create the production rules using RuleMaker. This study will produce the 6 similar classes through Unsupervised Clustering using data of 27 facilities(steel/concrete ratio on upper stories). The rule patterns of each class will be analyzed using RuleMaker. The following procedures are 5 basic steps for Unsupervised Clustering:

1. Data Input
2. Data Mining
3. Interpretation of summary
4. Interpretation on each cluster result.
5. Visualization of rules defining each cluster.

It is required to enter instance similarity value for data mining on 27 cases. This study applied the number 55 as shown in the figure 10, as the instance similarity value by consequential approach enabling to get 6 clusters. (The value for instance similarity encourages or discourages the creation of new clusters. A value closer to 100 encourages the formation of new clusters. A value closer 0 favors new instances to enter existing classes.)

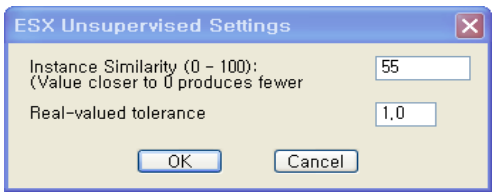

Figure 10. ESX Unsupervised Settings

The option for Rule Generator is set to 50 to generate rules and all other parameters keep the default values of the system. (Its screen is omitted because of limited space) The following table 2 is the summary on Class Resemblance Statistics.

Table 2. Class Resemblance Statistics

	class1	class2	class3	class4	class5	class6	Domain
Res. Score	0.68	0.77	0.66	0.81	0.7	0.83	0.49
No. of Inst.	7	3	3	6	6	2	27
Cluster Quality(%)	0.39	0.58	0.36	0.65	0.43	0.70	-

The 6 classes were generated and the resemblance scores and the cluster quality were analyzed. Higher cluster quality means higher class resemblance scores as compared to domain resemblance. (The domain resemblance indicates the overall similarity of all instances in a data set.)

It is important that resemblance score in a class is generally higher than domain resemblance score. The following results are domain statistics on numerical attributes. This summary provides the useful information on class average and standard deviation score. (see Table 3) Table 3. also shows in sequence the categorical attribute values that are most commonly generated among classes. The categorical attribute values help to recognize what categorical attributes is the best to identify each class.

Table 3. Domain Statistics & Commonly Categorical Attribute Values

DOMAIN STATISTICS FOR NUMERICAL ATTRIBUTES						
Class	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**
Steel (Ton/con)	0.148	0.156	0.170	0.165	0.111	0.160
(sd)	0.009	0.0116	0.010	0.010	0.017	0.014
MOST COMMONLY OCCURRING CATEGORICAL ATTRIBUTE VALUES						
class	**class1**	**class2**	**class3**	**class4**	**class5**	**class6**
Facility Type	***Ward office building***	***General Office***	General Office	General Office	***Post Office***	Post Office
Number of Floor	***Low***	***High***	Low	Low	***Low***	Low
Floor Area	***"4,000~ 6,000"***	***"10,000~ 12,000"***	"4,000~ 6,000"	"2,000~ 4,000"	***"0~ 2,000"***	"6,000~ 8,000"
Plan Shape	***Complex***	***Simple***	Complex	Simple	***Complex***	Complex
Large Space	***Yes***	***Yes***	Yes	Yes	***Yes***	No

The final process is about the results of rule production (see Figure 11). This chapter only examines the well learned Class 1, Class 2, and Class 5 above. We can identify the interesting rules from the preconditions of produced rules.

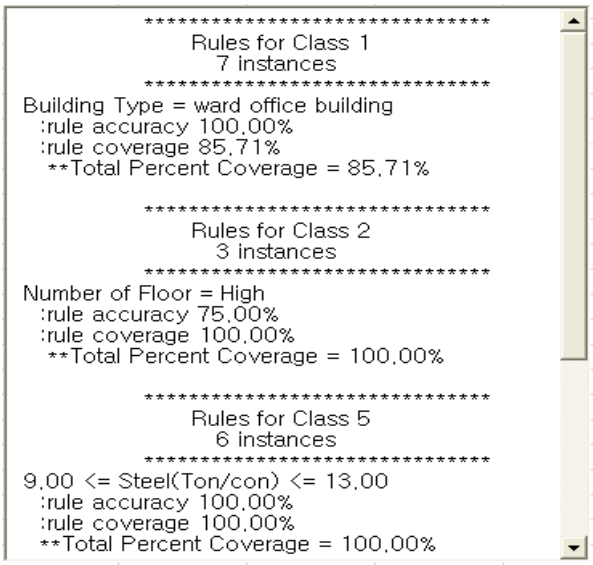

Figure 11. Rule Production Value of Each Class

The rule accuracy of Class 1 means 100% accuracy within the limit of instant 85.71% of Class 1. The rule coverage indicates the application scope of instance of relevant class. Let's look at the example. When the facility type is a ward office building, we can interpret that the rule coverage of Class 1 is 100% accurate within the scope of instance 85.71%. In addition, for Class 2, if the facility has lots of stories, 75% of all applications are accurate.

In other words, those rules will make wrong classification about 25% in all instances of the historical case. For a small post office in Class 5, the steel/concrete ratio per a unit is 0.09~0.13(ton/con) in accordance with the rule accuracy 100%.(see Table 3) All interpretations on the results above, the more a historical data have, the more useful rules can be produced.

This study excluded the feasibility of results and statistically significant interpretation because of the limit of legitimate collection of historical data. The clustering test of this study can be applicable as the knowledge based model supporting quantity prediction method by impact factors in the parametric cost estimating model.

14. Conclusions

More effective methods for cost estimating model have been continuously examined. There are a variety of research methods from a computer aided design of quantity take-off approach to a data mining of quantity prediction approach. Cost estimating model shall be able to cope with various kinds of environment since requirements for cost estimates gets more concrete forms depending on the design process and impact factors are also changed in the same way.

On the subject of design parameter by the factors analysis and clustering, this study tried to find out a solution against the limit in the OLAP system on the basis of quantity database. The OLAP driver was adopted as the preparatory step for data mining and performed as the preliminary analysis for various kinds of impact factors.

In conclusion, this study suggested the applicability in parametric cost estimating model by producing the useful knowledge and rules related to the quantity prediction by the Unsupervised Clustering, that it is using categorical attributes analyzed as the impact factors.

A further direction of this study will be to provide a additional evidence and a practical application system for the parametric cost estimating model.

ACKNOWLEGEMENT : This research was supported by a grant (06 CIT A03) from Construction Infra Technology Program funded by The Ministry of Land, Transport and Maritime Affairs.

REFERENCES

[1] Christian Stoy, "Driver for Cost Estimating in Early Design", ASCE/JANUARY 2008.
[2] Rodrigo Mora "Intergrating Conceptual Structural Design with Early Architecture", Information Technology, 2002.
[3] Saeed Karshenas "A Case-Based Reasoning Approach to Construction Cost Estimating", Information Technology, 2002.
[4] Hao Howard Nie "OLAP-Intergrated Project Cost Control and Manpower Analsis", Journal of Computing in Civil Engineering, May,June, 2007.
[5] Irtishad Ahmad "Data Warehousing in Construction Organization",ASCE.
[6] H, Kim, "Building Ontology to Support Reasoning In Early Design" Computing In Engineering, 2007
[7] Kan Phaobunjong, "Parametric Cost Estimating Model for Conceptual Cost Estimating of Building Construction Projects", Doctor of Philosophy, The University of Texas, May, 2002.
[8] Sevgi Zeynep Dogan "Using Decision Trees for Determining Attribute Weights in a Case-Based Model of Early Cost Prediction" , Journal of Construction Engineering and Management, February,2008.
[9] Claude Bedard, "Automating Building Design Process with KBES", Journal of Computing in Civil Engineering, Vol. 4, No. 2, April, 1990.
[10] Exchange Cost Model –Preliminary Appraisal(ER), http://idm, buildingsmart.no/confluence/display/IDM/Excang+Cost+Model+Preliminary
[11] D.K.H.Chua, "Process-Parameter-Interface Model for Design Management", Journal of Construction Engineering and Management, December,2003.
[12] AECbytes, http://www.aecbytes.com/review/2008/DProfiler_pr.html.
[13] Gregor Vilkner, "Integrated Process in Structural Enginneering", Structures Congress 2007.
[14] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001.
[15] Richard J. Roiger, Michael W. Geatz, "Data Mining A TUTORIAL-BASED PRIMER", Addison-Wesley,2003.
[16] G. Tesfagaber, "Semantic Process Modelling for Application Integration in AEC", Information Technology, 2002.
[17] Garold D. Oberlender, "Prediting Accuracy of Early Cost Estimaties Based on Estimate Quality", Journal of Construction Engineering and Management, June,2001.
[18] Min-Yuan Cheng, "GIS-Based Cost Estimates Integrating With Material Layout Planning", Journal of Construction Engineering and Management, August,2001.
[19] J.P. Marques de Sa., Pattern recognition : concepts, methods, and applications, Berlin ; New York: Springer, c2001.