# DEVELOPMENT OF A DYNAMIC SENSOR DATA MODEL WITH CONTEXTS FOR DATA MINING FROM MONITORING OF INFRASTRUCTURES

Yoshihiro YOSHIDA*, Nobuyoshi YABUKI*
Osaka University*

***ABSTRACT*:** Sensor monitoring is more and more important for maintenance of infrastructures and for prevention of disasters. Micro Electro-Mechanical Systems (MEMS) technology enables the cost of sensors to decrease rapidly, and wireless sensor networks can reduce the cost of cables significantly. Thus, more and more sensors are expected to be installed in various places for monitoring in the future.

In order to find meaningful information and knowledge from a large amount of sensor data, data mining has attracted considerable attention. However, simple application of the data mining technique to sensor data may not be successful compared to our expectation from our experience. As sensors are installed for infrastructures, in order to discover meaningful knowledge, contextual information, which is the situation data of each sensor, would be necessary.

Currently, sensor data models such as SensorML and building product models such as Industry Foundation Classes (IFC) have been being developed without any interaction. In addition, input data for most data mining algorithms are formulated as tables, which are different from data implemented in accordance with product and sensor data models. Therefore, building and sensor data must be converted into a table format for data mining.

In this research, in order to discover meaningful information and knowledge from a large amount of sensor data for buildings, we developed a dynamic data model which employs building information as contextual data of sensors. Then, we applied experimental data to the data model for evaluation and validation. Finally, data mining was executed using data stored in a database.

As a result, it was verified that the developed data model can represent the contextual data of sensors with flexible table structure and can be useful for discovering new knowledge by data mining.

***KEYWORDS***: data mining, sensor monitoring of infrastructures

## 1. INTRODUCTION

Micro Electro-Mechanical Systems (MEMS) technology enables the cost of sensors to decrease rapidly, and wireless sensor networks can reduce the cost of cables significantly. Thus, more and more sensors are expected to be installed in various places for their monitoring in the future.

In order to find meaningful information and knowledge from a large amount of sensor data, data mining has attracted considerable attention (Wu and Clements-Croome, 2007). Judging from our experiences, however, simple application of the data mining technique to sensor data may not be so successful as expected. Because sensor data are mostly mere numerical values and the data is related to the place on the object that each sensor is installed.

Thus, in order to discover meaningful knowledge from a large amount of sensor data, we proposed to integrate product and sensor data models before (Yabuki and Yoshida, 2006, 2007). To obtain better performance, we propose to add contextual information of sensors to the integrated data in this research. Contextual information means situational data of each sensor including the place and relationship with the object on which the sensor is installed.

Since input data for most data mining algorithms are formulated as tables, relational database in which all data are stored as multiple tables seems to be suitable for data mining. However, sensor data are stored in various formats. Furthermore, it is difficult to make a data model which can be applied to all environments where sensors are installed because the sensor application situations are numerous and various. Hence, the sufficient flexibility is required for a data model for data mining.

In this research, first, a flexible data model was developed incorporating various contextual data as well as the product and sensor data. Then, the developed data model was verified. Finally, data mining was implemented using the data stored in the database.

## 2. CURRENT DATA MODELS

For representing information and data of products and structures, various standards and specifications have been proposed as product models. Among them, the international standards for products are ISO 10303, known as STEP (STandards for the Exchange of Product model data). For representing building information, Industry Foundation Classes (IFC) have been developed by International Alliance for Interoperability (IAI) of which name is being changed to building SMART. IFC is a standard data format for Building Information Modeling (BIM) and is expected to enable the interoperability among processes such as design, construction, and maintenance.

Well known sensor data models are the Sensor Web Enablement (SWE) standards including Sensor Model Language (SensorML), Observations & Measurements (O&M), etc., developed by OpenGIS Consortium. SensorML is a standard model for implementing sensor information in an XML format. The objective of SensorML is to describe specifications of sensors, data flow, etc. O&M is a standard model for implementation of sensor information in an XML format. O&M has information about measurement types, time period, etc. SensorML and O&M are expected to enable sharing and exchanging sensor information on the web smoothly.

## 3. PROPOSED DATA MODEL

### 3.1 A&A Method

All data is stored in the relational database management system (RDBMS) as multiple tables, each of which consists of columns and records. In RDBMS, tables are connected by primary-foreign key relationships. Initial data models are normalized to eliminate redundancy. Once the data model is fixed, no modification should be allowed. However, it is difficult to initially define all required tables because new objects and attributes may be added later. Although tables can be added to the existing database, re-input of all data is necessary and queries need to be modified. Thus, the current data modeling practice may not be suitable for a large amount of sensor and structure data.

Since sufficient flexibility is required for a data

model for data mining, Addition & Absorption (A&A) method was proposed in this research. Figure 1 shows an application example of the A&A method and Figure 2 shows the essential tables for A&A method. In general, tables and attributes are defined as schema of database. Then, data are stored in a table based on the defined schema. On the other hand, in the A&A method, tables and attributes are generated from tables “Element” and “Attribute”. Then, data are stored in table “Value” linking “ElementID” and “AttributeID” as foreign key.

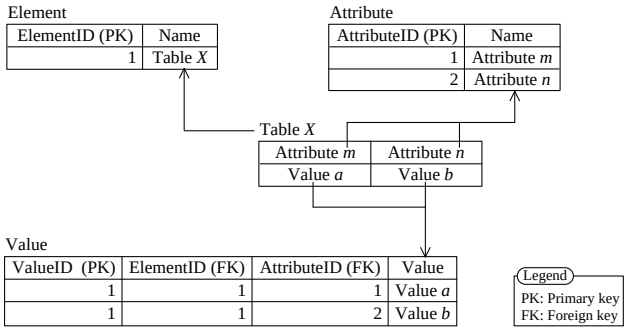



Figure 1. An example of A&A method

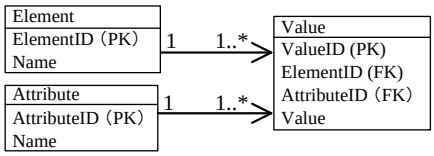



Figure 2. Essential tables for A&A method

### 3.2 Data model with contextual data—A&A data model

To discover hidden relationships from stored data in a database, environmental data where sensors are installed are necessary for data mining. In this paper, such environmental data of sensors are called “contextual data”. The reasons why contextual data are required are as follows: 1) Sensor data are usually represented as a set of mere numerical values and some data of sensor setting. But more data about sensor environment such as location, direction, the object that the sensor is attached, etc., is needed to clarify the meaning of sensor data. 2) For discovering hidden relationships among data or for improving the accuracy of prediction, restructuring and sophistication of data are required at each step of data mining. Contextual data are equivalent to factors of a target of knowledge discovery and are used for restructuring and sophistication.

IFC was examined for representing contextual data for sensors installed in a building. Figure 3 shows an analysis model used for examination of contextual data. The examination was executed

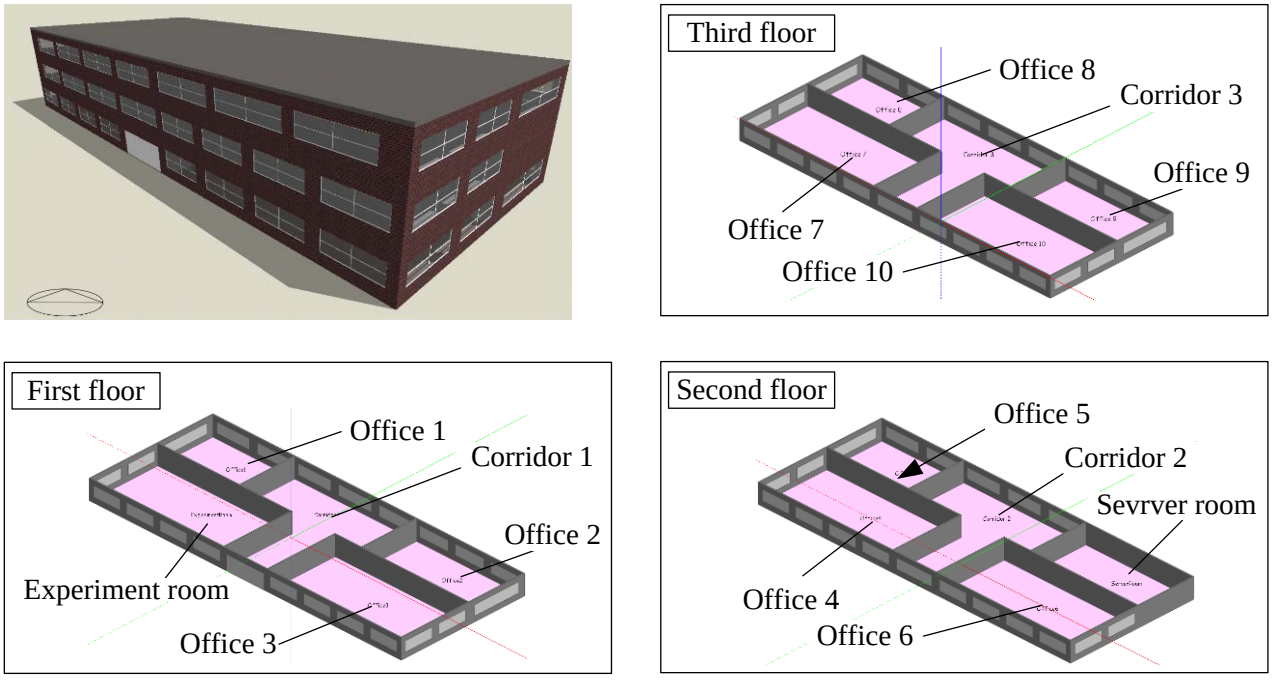



Figure 3. An analysis model of a building

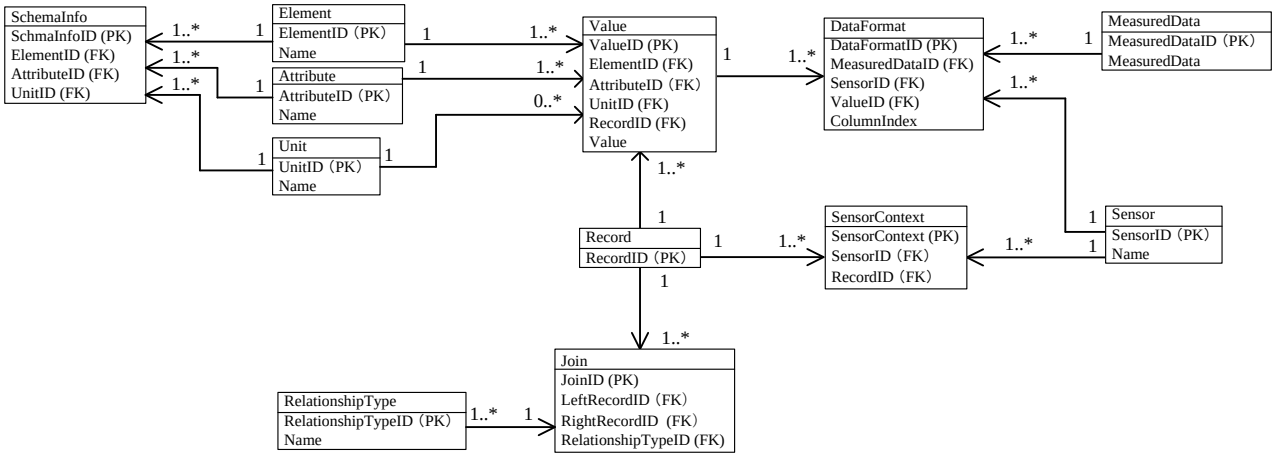



Figure 4. Integrated building and sensor data model with contextual data

making an instance based on IFC. As a result, the following points were found: 1) Contextual data is comprised of data about own context and relationship among contexts (e.g., relationship between a building and a room). 2) Relationships among contexts have several types such as “include”, “parallel”, etc.

Thus, a table which stores data about relationship data among contexts was required in a data model for data mining. Figure 4 shows the developed data model using A&A method. In this research the data model applied A&A method is called “A&A data model”. Figure 5 shows a part of data stored in a database based on the integrated data model.

### 3.3 Development of user interface for A&A data model

In A&A data model, when data are stored in a database, “ElementID” and “AttributeID” must be linked simultaneously. This requirement may increase input errors. Therefore, for improving practicability of A&A data model, user interface was developed using Microsoft Visual Basic.

For developing the user interface, required function was investigated comparing data storing process of A&A method and conventional way. As a result, the following functions are required for the user interface: 1) A function for schema definition. 2) A function for data input as table format. 3) A function for joining records in multiple tables. Figure 6 shows a part of the user interface.

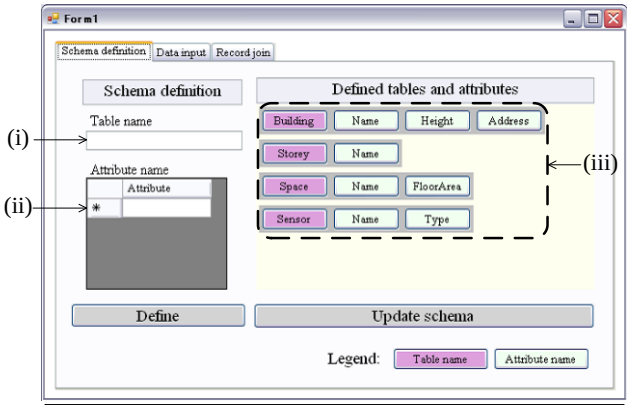
| Element | |
|---|---|
| ElementID (FK) | Name |
| 1 | Building |
| 2 | Story |
| 3 | Space |
| 4 | Sensor |

| Attribute | |
|---|---|
| AttributeID (FK) | Name |
| 1 | Name |
| 2 | Height |
| 3 | Address |
| 4 | FloorArea |
| 5 | Type |

| Unit | |
|---|---|
| UnitID (FK) | Name |
| 1 | Meter |
| 2 | Square meter |
| 3 | Celsius |
| 4 | Percent |

| Value | | | |
|---|---|---|---|
| ElementID (FK) | AttributeID (FK) | UnitID (FK) | Value |
| 1 | 1 | NULL | Test Building |
| 1 | 2 | 1 | 20 |
| 1 | 3 | NULL | Osaka, Japan |
| 2 | 1 | NULL | First floor |
| 2 | 1 | NULL | Second floor |
| 2 | 1 | NULL | Third floor |
| 3 | 1 | NULL | Office1 |
| 3 | 1 | NULL | Office2 |
| 3 | 1 | NULL | Office3 |
| 3 | 1 | NULL | Experiment room |
| 3 | 1 | NULL | Corridor1 |
| 4 | 1 | NULL | Sensor node 1 |
| 4 | 1 | NULL | Sensor node 2 |
| 4 | 1 | NULL | Sensor node 3 |
| 4 | 5 | 3 | Temperature |
| 4 | 5 | 4 | Humidity |

Figure 5. A part of data stored in a database based on the integrated data model




(i) Textbox for newly defined table
(ii) Multi-textbox for newly defined attribute
(iii) Existing tables and attributes in the database based on A&A data model

Figure 6. User interface developed for A&A data model

### 3.4 Verification of the data model

A sample scenario is applied to the A&A data model for verifying the data model. The scenario is as the following: A database had already stored contextual data of sensors installed in a building. Then, new contextual data of a bridge was required to be stored in the database.

Figure 7 shows a process of adding a new table using A&A method. For adding a new table, the word "Bridge" is inputted to the textbox in developed user interface. Then, this word is stored in attribute "Name" in table "Element" in the database. For adding attributes, the words "Name", "Address", "Span" are inputted to the multi-textbox in the user interface. Then, these words are stored in attribute "Name" in table "Attribute" in the database. For storing values in the new table "Bridge", data are inputted as table format using a function of the user interface. In the database, "ElementID" and "AttributeID" are linked as foreign keys and values are stored in attribute "Value" in table "Value" automatically. In this way, new contextual data was successfully added by the A&A method with no effect on existing data and tables in the database.

## 4. APPLICATION TO DATA MINING

### 4.1 Generating data for data mining

An environmental simulation software package, "Design Builder", was used to generate indoor environmental and energy consumption data of a three-story high building (Figure 3). Figure 8 shows the detailed information about simulation and the relationships with data sources for data mining. The data source for Building data was IFC and for indoor environmental and energy consumption data were SensorML and O&M. In this simulation, air conditioning was intentionally turned off in a room on the final day of simulation period due to an assumed damper trouble. Generated data and a part of scheduled data (i.e., occupancy, shading, and

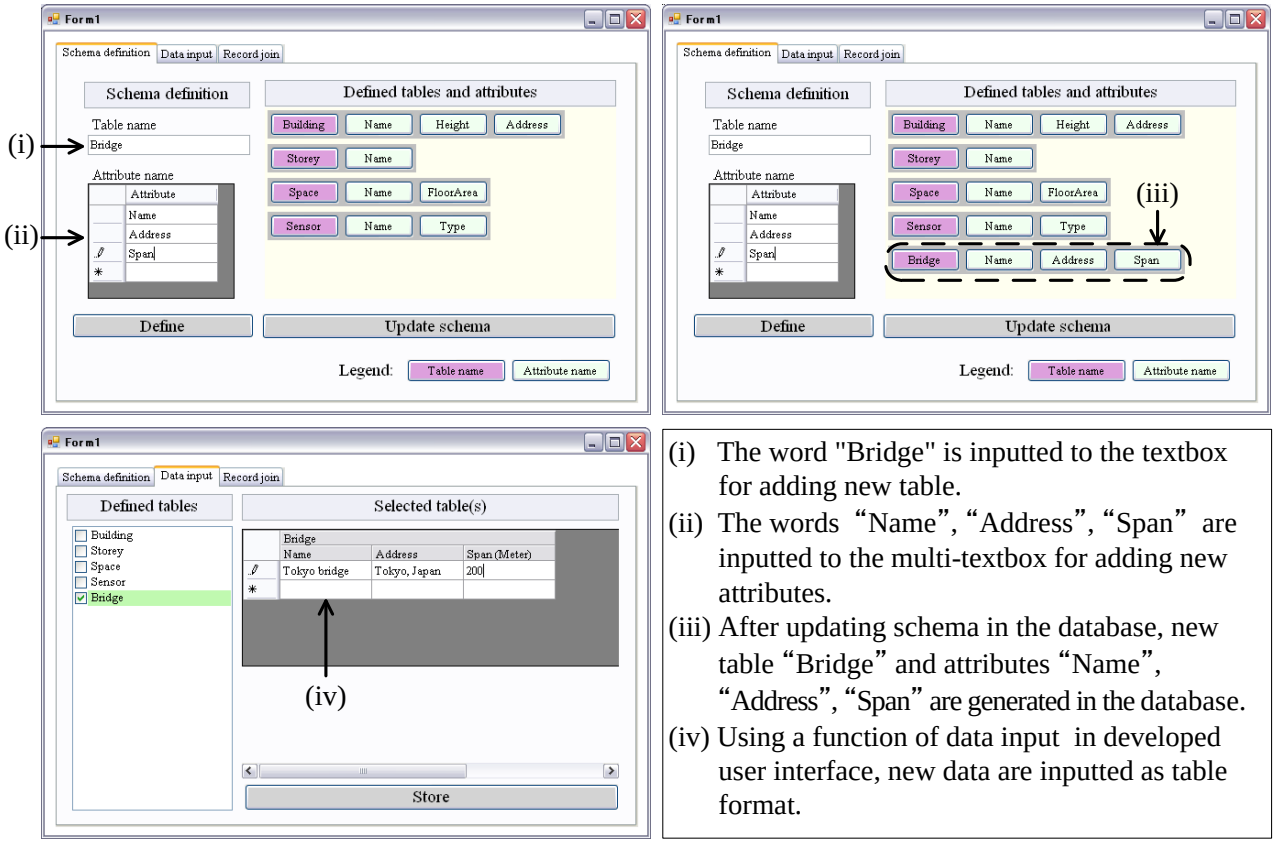



Figure 7. An example of addition process

door) were assumed to be measured by sensors installed in the building. After performing the simulation, sensor and contextual data were stored in the A&A database. The objective of data mining was understanding of inner environment of offices in the building.

### 4.2 Data mining results

Classification is required to generate rules or patterns by data mining. Sensor data has to be converted to categorical values in pre-processing step. Clustering algorithm was used for classification in this research. Two types for clustering procedures i.e., hierarchical and non-hierarchical, exist. In this research, non-hierarchical procedure was employed because the hierarchical procedure would make the understanding and evaluation of clustering complicated and difficult. The X-means algorithm (Pelleg and Moore, 2000) was employed as a non-hierarchical clustering procedure from the advantage of optimum value seeking process.

In pre-processing step, clustering was executed twice using temperature and humidity data in all offices. First clustering was performed for extracting temporal feature and for dividing time series data based on the clustering results. Second clustering was done for generating class information. In this paper, first clustering processing is called "Segmentation" and the produced cluster is called "Segment" for distinguishing from the second processing. Figure 9 shows the segmentation result and the result of generated class information.

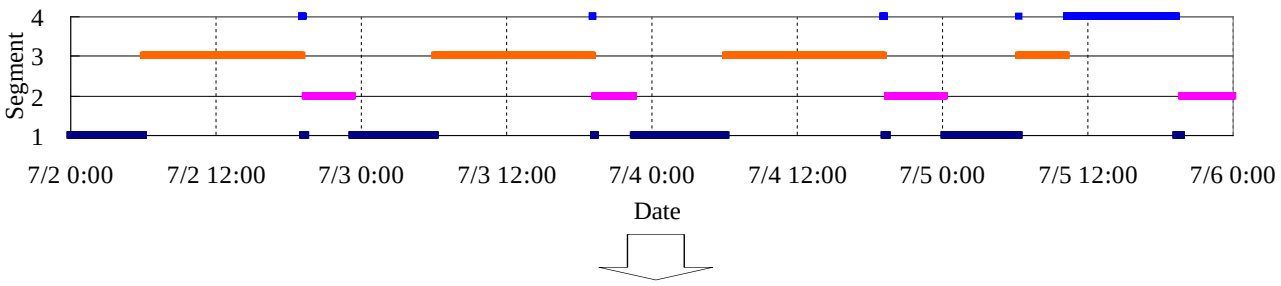
Data mining was executed using generated class information and contextual data. In this case, contextual data was divided into two types, i.e., inner and outer contexts. Inner context is composed of environmental data where sensors are set, e.g., occupants, office automation apparatus, etc. Outer context consists of information of the space adjacent to the office the sensor is located, i.e., space name and direction. Although inner context can represent

| Site | Simulation Period |
|---|---|
| Osaka, Japan | 7/2-7/5 |

| Zone | Air conditioner | |
|---|---|---|
| | Setting temperature (°C) | Operating time |
| Office | 24.00 | 6:00-19:00 |
| Server room | 20.00 | 24 hours |
| Experiment room | - | - |

| Data source | Content |
|---|---|
| SensorML | Indoor air temperature; Air flow tempereature from an outlet; Electricity of air handling unit; Indoor air humidity; Electricity of office auto mation appartuses; Electricity of office lighting; Generated data from simlation |
| O&M | Auto mation appartuses; Door; Shading; Air handling unit; Occupancy; Lighting; Scheduled data for simulation |
| IFC | Building; Story; Space; Building data |

Figure 8. Detailed information about simulation (left two tables) and data sources for data mining (right table)




| Segment | Office 1 | Office 2 | Office 3 | Office 4 | Office 5 | Office 6 | Office 7 | Office 8 | Office 9 | Office 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Segment 1 | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 1 |
| Segment 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Segment 3 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| Segment 4 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |

Figure 9. Results of segmentation (above) and generation of class information (below)

data in a table format, outer context is difficult to represent in a table format because more than two values may correspond to one attribute. Therefore, transaction data format was applied to outer context.

For analyzing contributions of attribute against class information, “Relief”, which is one of the attribute selection algorithms, was applied to the inner context. The strength of relationship between class information and attributes is evaluated and rank values, of which highest value is one, are output as a result of evaluation of this algorithm. On the other hand, “Apriori” algorithm, which can process transaction data, was applied for outer context. In this step, “Segment 2” was excluded for investigation since all offices were classified in the same class (Figure 9). Figure 10 shows the processes of data mining. Since Apriori algorithm produced a large amount of rules, filtering conditions were used to select rules applicable to only one class.

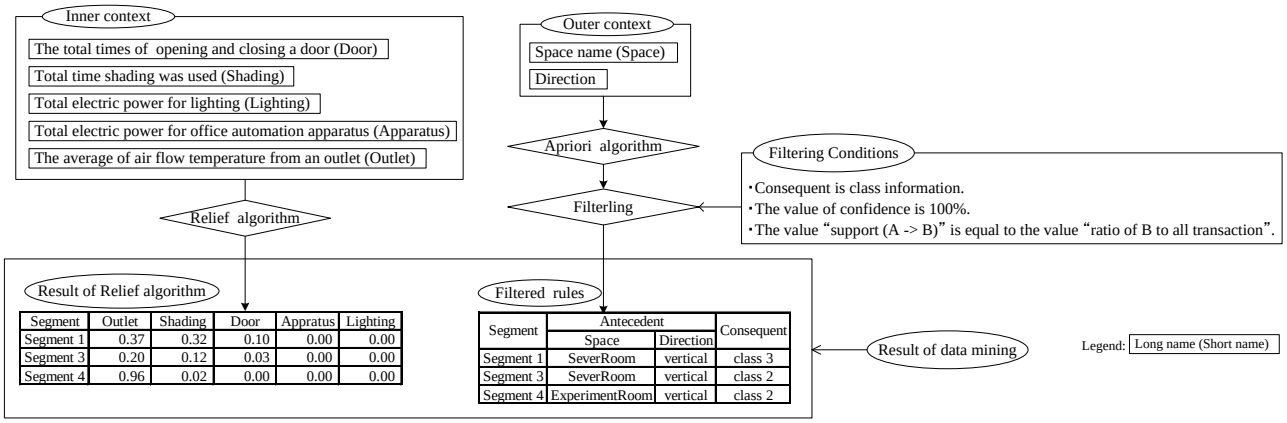
## 5. DISCUSSION

First, data of the state of air handling unit, i.e., ON or OFF, and occupancy rate, i.e., =0 or 0<, was collected in order to judge whether the inner context can be ignored or not for each segment. Note the outer context must be investigated.

- Segment 1: Air-handling unit was mostly OFF and occupancy rate was almost always =0. Thus, the inner context can be ignored.
- Segment 3 and 4: Air-handling unit was almost always ON and occupancy rate was mostly 0<. Thus, the inner context must be investigated.

Next, each segment was investigated using inner and/or outer contexts. By the result of Relief algorithm, “Outlet” obtained the highest rank. Thus, “The average of air flow temperature from an outlet” was selected as the primary item of the inner context. On the other hand, the filtered rules were used as for the outer context. The result of each segment is as follows.

- Segment 1: Office 2 and office 9 were in class 3 because the outer context of them was “there is a server room up or down adjacent to the office.”
- Segment 3: Since the averages of “The average of air flow temperature from an outlet” (Table 1) are about the same, the inner context can be ignored. Office 2 and office 9 were in class 2 because the outer context of them was “there is a server room up or down adjacent to the office.”
- Segment 4: Office 4 was in class 2 and other offices were in class 1. Result of Relief algorithm shows that “Outlet” obtained the highest value. From “The average of air flow temperature from an outlet” (Table 1), office 4 was a particularly high temperature value compared to other offices. Thus, outlet of office 4 can be judged as malfunctioning.




| Segment | Outlet | Shading | Door | Appratus | Lighting |
|---|---|---|---|---|---|
| Segment 1 | 0.37 | 0.32 | 0.10 | 0.00 | 0.00 |
| Segment 3 | 0.20 | 0.12 | 0.03 | 0.00 | 0.00 |
| Segment 4 | 0.96 | 0.02 | 0.00 | 0.00 | 0.00 |

| Segment | Antecedent | | Consequent |
|---|---|---|---|
| | Space | Direction | |
| Segment 1 | SeverRoom | vertical | class 3 |
| Segment 3 | SeverRoom | vertical | class 2 |
| Segment 4 | ExperimentRoom | vertical | class 2 |

Figure 10. Processes and results of data mining

Table 1. “The average of air flow temperature from an outlet” (Segment 3 and 4)

| Segment | Office 1 | Office 2 | Office 3 | Office 4 | Office 5 | Office 6 | Office 7 | Office 8 | Office 9 | Office 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Segment 3 | 24.00 | 23.94 | 24.00 | 24.20 | 24.02 | 24.02 | 24.03 | 24.01 | 23.94 | 24.01 |
| Segment 4 | 24.21 | 24.13 | 24.24 | 28.30 | 24.25 | 24.29 | 24.36 | 24.26 | 24.19 | 24.31 |

## 6. CONCLUSION

In order to find meaningful information and knowledge from a large amount of sensor data, “Addition & Absorption (A&A) Method” was developed in this research. This method enables development of an integrated building and sensor data model with contextual data, i.e., “A&A Data Model”. Then, the developed data model was verified storing sensor and contextual data in a database. Finally, data mining was executed using the data stored in the database. And the test showed that some meaningful knowledge was found in the data mining. Future work includes application of A&A Data Model for actual sensor data.

## REFERENCES


WU, S., CLEMENTS-CROOME, D., Understanding the indoor environment through mining sensory data—A case study, 2007, *Energy and Buildings*, 39(11), 1183–1191.

YABUKI, N. AND YOSHIDA, Y., 2006. A Data Model for Storing a Large Amount of Sensor Data, *Proceedings of the First Asia-Pacific Workshop on Structural Health Monitoring*, Yokohama, Japan, December 4-6, Paper No.36, pp.1-8.

YABUKI, N., 2007. An Intelligent Framework for Knowledge Discovery from a Large Amount of Data in SHM, *Proceedings of the World Forum on Smart Materials and Smart Structures Technology*, Chongquing, China.

PELLEG, D. AND MOORE, A., 2000, Extending K-means with Efficient Estimation of the Number of Clusters, *Proceedings of the 17th Conference on Machine Learning*, Stanford University, Stanford, CA, USA, pp.727-734.