Hybrid Information Mining Approach on BIM-based Building Operation and Maintenance

Yang Peng[a], Jia-Rui Lin[a], Jian-Ping Zhang[a], Zhen-Zhong Hu[a,b,*]

[a] *Department of Civil Engineering, Tsinghua University, Beijing 100084, China*

[b] *Graduate School at Shenzhen, Tsinghua University, Guangdong 518055, China*

Abstract: Huge amounts of data are generated daily during the operation and maintenance (O&M) phase of buildings. These accumulated data have the potential to provide deep information that can help improve facility management. Building Information Model/Modeling (BIM) technology has proven potential in O&M management in some studies, making it possible to store massive data. However, the complex and non-intuitive data records, as well as inaccurate manual inputs, raise difficulties in making full use of information in current O&M activities. This paper aims to address these problems by proposing a BIM-based Data Mining (DM) approach for extracting meaningful laws and patterns, as well as detecting improper records. In this approach, the BIM database is first transformed into a data warehouse. After that, three DM methods are combined to find useful information from the BIM. Specifically, the cluster analysis can find relationships of similarity among records, the outlier detection detects manually input improper data and keeps the database fresh, and the improved pattern mining algorithm finds deeper logic links among records. Particular emphasis is put on introducing the algorithms and how they should be used by building managers. Hence, the value of BIM is increased based on rules, extracted from data of O&M phase, that appear irregular and disordered. Validated by an integrated on-site practice in an airport terminal, the proposed DM methods are helpful in prediction, early warning, and decision making, leading to the improvements of resource usage and maintenance efficiency during the O&M phase.

Keywords: data mining; building information modeling; operation and maintenance; cluster analysis; pattern analysis; outlier detection

1. Introduction

The operation and maintenance (O&M) phase of buildings is the longest phase within the lifecycle. It always involves sophisticated interactions among various stakeholders, facilities, professionals, and management activities, as well as some multifarious works, such as scheduling, space planning, repairing, and emergency managing. From daily activities, huge amounts of disordered data are generated. In order to manage these data, introducing building information model/modeling (BIM) technology in building O&M is currently in practice. BIM provides a parametric and detailed model with the related components of buildings, as well as integrated model views that enable constant synchronization of any changes [1], within a unified information repository, supporting the requirements of information integration [2] for collaboration among different stakeholders. In addition, BIM enables better information exchange from the design and construction phases towards the O&M phase, on the other hand, makes it possible to store mass information generated during the O&M phase. Therefore, BIM can perform as the data layer in applications [3]. However, when the data in BIM increase to a certain volume, some features of "big data" emerge. It is believed that big data have the potential to find latent patterns as well as to help prediction [4], but the problems in the information requirement within big data arise in at least two aspects.

(1) The increasing volume of data in BIM is now challenging the outdated method of data usage and experience-based decision-making paradigms [5]. The heterogeneity in information, the complexity in storage, and the specialized functions of users lead to more and more non-intuitive data. Current BIM standards usually represent building elements and their relationships by complex structures. For example, the practical as-built BIM of an airport terminal, as discussed in the case study section, is approximately 50 GB in database size with over 10 million building entities that are deeply linked to each other. Only managers with enough professional knowledge can access information via BIM. Ways to extract useful information from BIM data and represent those patterns in an understandable form are worth exploring.

(2) Inaccurate data may hamper management activities. Manual input, which is an error-prone procedure, still plays an important role in data input and management in current practice. Wrong input can lower the data quality and lead to negative implications for management activities. For example, if an improper repair instruction is

attached to a pump, workers may incorrectly perform repair tasks. However, manual checks are almost impossible, because handling so much data will be tedious and costly [6]. Those inaccurate records may also confuse the data analysis process [7].

The above two problems exist as the gap between BIM data and the use of information. In order to maximize the strength of big data to find valuable patterns, some data mining (DM) approaches are introduced in this paper as useful tools to address these problems. It has been validated that the value of BIM data can be enhanced by DM processes [8]. DM is a knowledge discovery method from big data. Retrieving information is an important component of artificial intelligence (AI) that was developed in the 1960s, and has been a well-developed research area in computer science since then [9]. When using DM, three main steps are usually involved: (1) data sets are transformed and standardized into appropriate forms when invalid and missing data are eliminated at the same time. (2) Core mining algorithms are then executed to find information from the data. (3) Mining results are represented in an understandable form for users.

This paper utilizes a DM process in handling BIM data within the O&M phase to extract useful laws and patterns. In section 2, related studies and applications of BIM-based O&M and DM are reviewed. The following three sections introduce cluster analysis, outlier detection, and pattern analysis methods, respectively. Then, a validation of the proposed approach is given. The last two sections provide a discussion and a conclusion.

2. Literature Review

2.1. Application of BIM-based big data in O&M

With a strong ability to manipulate huge data, BIM supports the information requirements in O&M applications. As a result, an increasing amount of data from various sources are accumulating routinely in BIM, forming big data. These data are mainly gathered from the following channels.

(1) When establishing BIM, basic information is input manually or derived from standard component libraries. Records generated in daily management are integrated by methods into existing models, such as schedule and work package information [10]. The accumulation of big visual data, such as pictures, videos, point cloud, etc., is discussed as well [11].

(2) Many entities are transformed by algorithms from outside the BIM environment via some building management tools. For example, a framework was developed to store and distribute knowledge in the management process [12]. This approach could acquire lessons from previous projects and then map to the corresponding elements in BIM. A semantic material matching system [13] transformed the names of materials in BIM according to standard libraries. This mainly addressed the semantic conflicts among stakeholders and brought rich information of material properties from the ontology web, as well. Most algorithms were capable of gathering dense data, but further usages of transformed data were limitedly discussed. Another algorithm was reported that point cloud data were converted to BIM objects together with semantic information [14].

(3) Documents were delivered from the design or construction phase to O&M models. More documents were gradually attached to building components during daily management, such as checking lists and operation history.

(4) Sensors (indoor/outdoor) collect huge amounts of samples and send them back to the data repository [15]. Location/orientation via BIM [16] and RFID technology [17,18] were used as data sources. They usually send fragmentary data in BIM when executing assignments. In addition, cloud technology enables workers to establish dynamic data in BIM [19].

Some cases on big data usage in O&M were reported. A decision-making support system for large facility management was built based on data and knowledge [20]. For the supporting Building Energy Model (BEM), a framework was proposed to integrate relative data into BIM databases [21]. A tool for energy monitoring and optimization was developed based on BIM [22]. The sensor data attached to BIM objects were analyzed. These data sets were first visualized for human observation, and then transformed into a fault detection and diagnosis process. A standard parametric model was used to recognize certain factors among retrieved data when there were abnormal activities. Such a method was intuitive, especially when targeting patterns were already defined before analysis. For example, “turn down the air conditioner when the room is too hot” is common knowledge to be predefined in the system. In contrast, other research introduced data mining skills into the building automation system for finding unknown relationships in observed data [23]. Tens of thousands of sensor data were recorded. Such big data were classified into three groups and summarized individually. Then, two kinds of unusual patterns were detected leading to prominent energy saving. As energy source data are typical big data sets, finding common patterns and discovering unknown relationships can support decision-making on financial measures. Some other studies, such as research on public safety management [24] and evacuation behaviors [25] have made full use of data about the building environment. For example, route planning for escape should utilize relationships among spaces to calculate the distance. A smart grid big data framework to summarize and output patterns of electricity consumption was designed [26]. The above-mentioned studies indicated the central position of big data when addressing issues about O&M activities.

Two BIM-based O&M systems [27,28] utilized huge databases that stored records about properties, locations, and three-dimensional (3D) information of components. These two systems used different but smart schemes to simplify the volume of BIM. They both worked efficiently, taking advantage of the Relational Database Management System (RDBMS), but neither of them was able to provide further patterns among records. Moreover, a web-based RDBMS was used in a facility management system [29]. This work provides a WebGL viewer to help on-site management. However, these developed systems stored massive data mainly for viewing and searching, rather than supporting data mining or knowledge discovery.

2.2. Practices of DM in building industry

Strategies in DM were reviewed, as DM is utilized to analyze large data sets generated in the O&M phase. Some innovative research has been reported recently. As reviewed, clustering, regression, and pattern analysis were the most commonly used in projects. The factors of steam load were mined as a regression model to predict in every month [30]. Miller et al. [31] demonstrated a novel method called Symbolic Aggregate approXimation (SAX). Pattern analysis was executed on facility operational data, and found rules to save energy [32]. A recent trend shows that researchers prefer multiple methods rather than an individual method in their works. For instance, clustering was used to discover basic operations of doors and windows, and then five abnormal patterns were discovered through pattern analysis for ventilation design [33]. Two studies utilized DM on the energy consumption data of skyscrapers, and both gave examples of multiple analyses: one combined cluster and patterns [23], whereas the other used both classification and regression [34]. Moreover, classification and decision tree have been employed in finding factors of injuries and accidents [35], and in predicting the cost-saving potential of houses [36]. The daylight metrics were analyzed by an existing DM software through even more methods; their performance varied, but was never bad [37]. Four algorithms formed a model for predicting the performance of green building projects [38]. Clustering pattern analysis was used to find daily behavior from sensors in two buildings [31]. These studies indicate that several DM methods may work in a sequence to dig deeper information step by step. Several studies mainly used text mining [39,40], with other predicting algorithms for predicting cost overruns in construction. This research demonstrated the value of detecting special patterns. Common methods, such as outlier detection, usually work well with other methods. For example, a rule-based approach was also used for detecting abnormal patterns [15]. Since clustering and associated patterns can describe data from different aspects, they tend to be combined in parallel to provide more comprehensive views of the hidden patterns [7].

2.3. Discussion

Data are heavily gathered from various sources by different methods. These data are analyzed by means, showing their power in practice. Many successful cases were reported regarding DM towards building management. Some of them were BIM-based, taking advantage of interoperability. DM skills have already been widely used in buildings, in which most cases focused on building behavior analysis. DM results have been used in prediction as well as finding abnormal patterns. The basic strategy is that multiple DM methods often complement each other. However, applications of big data remain relatively shallow and simple, and have not yet followed a unified workflow. Further usage of those data should be exploited. Besides, most developed BIM systems lack support for DM functions. Among DM studies, there is not much research on specialized algorithms towards BIM-based O&M of buildings. Table 1 summarizes and compares the key related studies on DM in buildings. As indicated in the table, only two of them were supported by BIM platforms, and the majority utilized classic algorithms but proposed no improvements or new methods.

In this paper, targeted algorithms for BIM data are proposed, trying to address the problems in information requirements and support decision-making during O&M phase. The key characteristics of the proposed approach and platforms compared to related studies in Table 1 are: (1) The BIM-based approach exploits the natural ability of collaboration of BIM platforms. Raw data can be directly extracted from the design and the construction phase, and the mining results can be shared among stakeholders through a unified working platform. (2) The proposed approach made some improvements based on classic algorithms. For example, the time complexity of pattern analysis for decision-making in O&M occasions was improved. (3) Validated by the high volume of data, this hybrid DM approach was proven effective.

Table 1. A non-exhaustive list of related studies on DM in buildings

	interoperability: supported by a BIM platform	the algorithms used: Classic	the algorithms used: Existing software	the algorithms used: New methods	validated by big volume of data	benefits from combined methods
Chen et al. [19]	●		●	●	Sufficient	
Costa et al. [22]	●	●			Sufficient	
Xiao et al. [23]		●			Sufficient	●
Miller et al. [31]		●			Sufficient	●
Yu et al. [32]			●	●	Simple	
Son et al. [38]		●			Simple	●
proposed approach	●	●		●	Sufficient	●

3. Cluster Analysis

Large construction projects generate various kinds of data from different disciplines every day. Currently, the data are usually input to, stored in, and retrieved from a BIM repository. It is difficult to summarize the deep relationship among those data. For example, managers usually have to hold a meeting for one hour with workers to check the 3D model and the data lists in order to obtain the spatial distribution of repairs—a task that tends to be slow and tedious. Statistical methods, such as charting and regression, are not sufficient because the relationship between repair records and spatial structures is not obvious. Cluster analysis is able to find information about similarity relationships in the data [7]. Therefore, managers can benefit from the information to make timely and reasonable decisions. For example, if they find some similar records containing repaired electric units in the same region, workers can then carry out a thorough investigation in the region. In addition, cluster analysis (an unsupervised algorithm) does not require manual interactions to obtain training sets; therefore, information can be generated automatically. This section proposes a clustering approach towards structured data from BIMs for giving valuable information on hidden relationships behind the data. This approach first establishes a data warehouse through data extraction and transformation. On this basis, a cluster analysis algorithm, where the parameter k_c should be carefully determined (Section 3.4), is executed for classification. After clustering, a coefficient is introduced to evaluate the quality of these clusters.

3.1. Establishing the data warehouse: data extraction and transformation

The Industry Foundation Classes (IFC) standard is widely used in representing entities in BIM. The IFC is a kind of object-oriented, rich and neutral schema, and in most cases, different implementers choose either an object-oriented database or a relational database as backing storage. Generally speaking, an object-oriented database is better at expressing IFC entities and their logics, while a relational database works better for processing large data sets. In this research, the relational database is selected, and thus, the IFC file is imported and concurrently transformed to relational expressions. However, records in such a database are not yet ready for DM, especially when the data are distributed in many data tables. Thus, extracting data from the relational database and organizing them in a new form aimed at a more efficient DM is necessary. A data warehouse is an integrated and stable container of data, with an explicit scene of application and essential analyzing tools [41]. In this study, the warehouse is built following two steps, described below. It should be emphasized here that some information would be lost during this data processing, because the warehouse is only a temporary and concise storage for future DM process. Careful definitions of rules about extraction and transformation are required to keep useful data in the warehouse, ensuring that the missing information are of no importance for current problems.

(1) The data controller first changes the strategy of data storage. As shown in Fig. 1, the repair record is drawn as LIST_1 in the relational database. This list has several properties and two foreign references. Properties are directly put into the relevant record on the right, and the two references point at LIST_2 and LIST_3, respectively. This process is recursive until all references are extracted (when LIST_4 is reached). All records in the database are transformed similarly. When calculating large amounts of records, there would be no need to expand references in the database. For example, three references have to be searched in the database to retrieve the same record in Fig. 1, while the data warehouse can provide this record in only one operation. This strategy saves much time in performing common operations in DM such as massive calculation and high-speed analysis. When data size increases, it takes more time to perform this transformation even in a nonlinear way, but has no effect on the following DM process since it is carried out before data analysis.

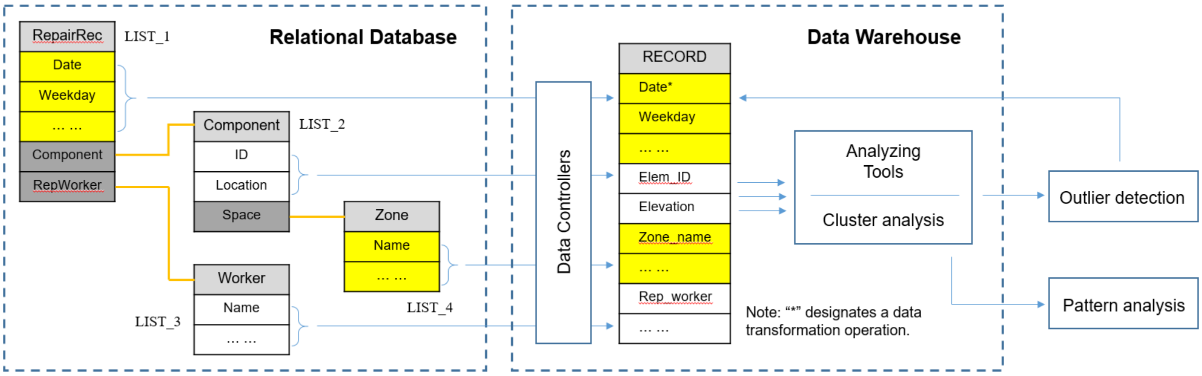

Fig. 1. Data extraction from BIM database

(2) BIM data have the nature that categorical and numeric (discrete and continuous) records are usually mixed. For example, a repair record may include the date and time, the type of the objective equipment, the operator's name, as well as operation log, etc. Thus, the data controllers perform two transformation operations on different kinds of BIM data in this approach. First, the data controller performs normalization on some numeric properties by mapping onto a certain numeric interval. For example, positive numeric property X should be mapped from ($x_{\min}$, $x_{\max}$) to (0, 1) as Eq. (1).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Second, the data controller performs discretization on some continuous numeric properties. When they are corresponding to daily concepts, they are reduced to discrete values making them more understandable for users. For example, "time of day" has a value from 0:00:00 to 24:00:00 according to the definition in the database, and it is reduced into "morning," "afternoon," and "night." The asterisks in Fig. 1 designate those transformed properties.

3.2. Clustering algorithm

After establishing the data warehouse, the cluster algorithm will be carried out to divide all original data records into k_c[1] clusters automatically, with a final goal of making records in the same cluster as similar as possible, but any two different clusters should be dissimilar [42]. This study adopted the popular "k-means" clustering algorithm. In this algorithm, the similarity between records $\boldsymbol{p}$ and $\boldsymbol{q}$ can be measured by the distance

[1] k_c refers to the number of the clusters that should be determined before analysis. It will be further discussed in section 3.4.

scale:

$$dist(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{n_N} \Big| Nu_i(\boldsymbol{p}) - Nu_i(\boldsymbol{q}) \Big| + \sum_{j=1}^{n_D} \delta\Big(Ds_j(\boldsymbol{p}), Ds_j(\boldsymbol{q})\Big) \tag{2}$$

where Nu_i is the i^{th} normalized numeric property value, and Ds_j is the j^{th} discrete property value. $\delta(x,y)$ is 0 when $x=y$, and 1 when $x\neq y$. After the parameter k_c is given, the algorithm works as the pseudocode below.

Algorithm Clustering

Input: k_c

1. Randomly selected k_c records as the initial cluster centers.
2. Initial clustering: Each record is distributed to the nearest cluster center (with the smallest $dist(\cdot,\cdot)$).
3. **Do loop**

   (1) Refreshing centers: Within each cluster, the average property values are calculated as new cluster centers.

   (2) New centers are used to redistribute.

   Until All new centers and distributions remain the same.

Output clusters $C_1, \ldots C_{k_c}$

The average property value in step 3 is determined by types: for a continuous numeric property, the arithmetic average is used; for a discrete or discretized numeric property, the value that appears most of the time is chosen.

3.3. Quality of clusters

Then, the cluster silhouette coefficient (S) [43] is used to evaluate the quality of cluster results. A larger S indicates a cluster with better quality.

First, the internal distance and the external distance are calculated for a record $\boldsymbol{o}$ in cluster C_i. Here, $d_{\text{in}}(\boldsymbol{o})$ is the internal distance—the average distance from $\boldsymbol{o}$ to other records in C_i, while $d_{\text{ext}}(\boldsymbol{o})$ is the external distance—the minimum average distance from $\boldsymbol{o}$ to records in other clusters:

$$d_{\text{in}}(\boldsymbol{o}) = \frac{\sum_{\boldsymbol{o}'\in C_i, \boldsymbol{o}\neq\boldsymbol{o}'} dist(\boldsymbol{o},\boldsymbol{o}')}{|C_i|-1} \qquad d_{\text{ext}}(\boldsymbol{o}) = \min_{C_j: j\neq i}\left\{\frac{\sum_{\boldsymbol{o}'\in C_j} dist(\boldsymbol{o},\boldsymbol{o}')}{|C_j|}\right\} \tag{3}$$

where $|C_i|$ is the total amount of records in the i^{th} cluster. Then, record $\boldsymbol{o}$'s silhouette $S_{\text{obj}}(\boldsymbol{o})$ is defined as

$$S_{\text{obj}}(\boldsymbol{o}) = \frac{d_{\text{ext}}(\boldsymbol{o}) - d_{\text{in}}(\boldsymbol{o})}{\max\{d_{\text{in}}(\boldsymbol{o}), d_{\text{ext}}(\boldsymbol{o})\}} \tag{4.1}$$

Finally, the S of a cluster is the arithmetic average of all $S_{\text{obj}}(\boldsymbol{o})$ of its own records.

$$S(C_i) = \frac{1}{|C_i|} \sum_{o \in C_i} S_{\text{obj}}(\boldsymbol{o}) \tag{4.2}$$

A smaller $d_{\text{in}}(\boldsymbol{o})$ and a larger $d_{\text{ext}}(\boldsymbol{o})$ makes S larger, and a large S means the records in this cluster are close to each another, but far away from all other clusters (examples are shown in Fig. 2).

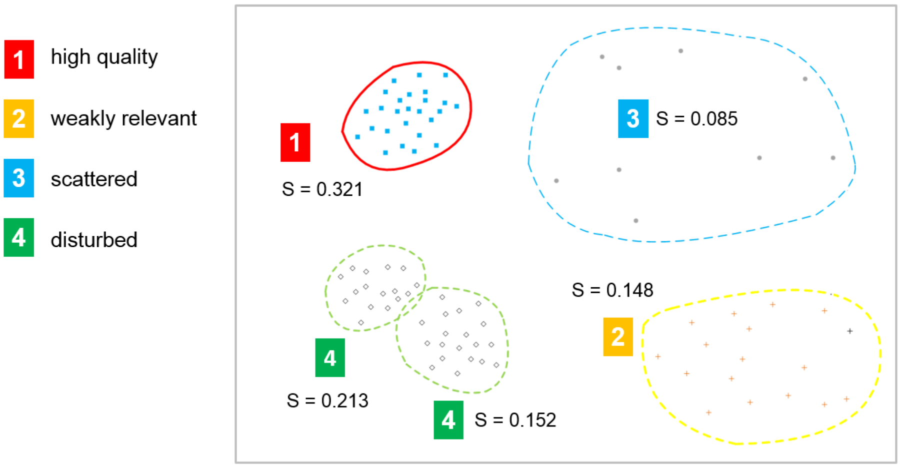

Fig. 2. Four typical kinds of clusters and their typical silhouette coefficients

S ranges from -1 to 1, according to its mathematical definition. In common situations, *S* of ~0.30 to 0.40 is good enough to be considered as high quality. Meanwhile, a low *S* that is close to zero cannot be avoided. Depending on the value of *S*, all clusters can be divided into four typical types: high quality, weakly relevant, scattered, and disturbed, as shown in Fig. 2. High-quality clusters contain complete information and strong rules that managers can directly use in decision-making. Weakly relevant clusters produce less information. Scattered clusters that have the smallest *S*, and consist of separate, individual records. These scattered clusters will be discussed in outlier detection in the next section. Disturbed clusters behave differently, showing that they are the consequence of a high-quality cluster which is improperly divided into several parts. Each part has similar records ($d_{\text{in}}(\boldsymbol{o})$ is small, like high-quality clusters), but it will be disturbed by some close neighbors ($d_{\text{ext}}(\boldsymbol{o})$ is also small). Managers should combine these clusters and reform a high-quality cluster. In addition, $S < 0$ is a poor result, making the clustering process invalid.

3.4. Determining the number of clusters (k_c)

k_c is the only parameter pre-defined in the cluster algorithm. It decides the overall presentation of the result. Therefore, k_c should be carefully determined. This study proposes a three-step method to determine k_c by introducing the professional background of O&M management. The method is illustrated in Fig. 3. The horizontal axis represents the value of k_c. In the case study, the total record amount n_r is 2281. The following part demonstrates how to calculate the appropriate k_c.

Step 1 derives a point estimation according to Eq. (5) [7]

$$k_c^* = \left[\sqrt{n_r/2}\right] + 1 = \left[\sqrt{2281/2}\right] + 1 = 34 \tag{5}$$

This number is presented as a single point on the horizontal axis.

Step 2 involves professional knowledge about the scenes of application from O&M managers. A rough range of k_c is drawn after browsing the data. This range should not be too far away from k_c* in Step 1. For example, when finding clusters related to spatial structures, managers should be familiar with every region in the building. As for the airport terminal in the case study, every floor was divided into eight regions. Therefore, at least eight clusters were needed. On the other hand, 40 clusters were determined as the upper bound, because too many clusters were inconvenient for observation. Finally, the range is roughly given from 8 to 40 (marked by an arrow strip in the left chart).

Step 3 is a parametric analysis. Clustering runs for each k_c and S is recorded for each run. The functional relationship between k_c and S is then drawn in the charts within the range from Step 2. In the left chart, a wide line is used to plot the average S_{ave} and the vertical lines are used to mark S_{max} to S_{min}. In the right chart, the slope (rate of change) of S_{ave} and S_{max} are also plotted. In terms of overall tendency, the average $S(k_c)$ is roughly increasing with k_c. Therefore, k_c is not determined by a large S. In this research, the criterion suggests that a better k_c should make S grow faster than its neighbors. This kind of k_c lies on the zero points of the second derivative of $S(k_c)$, or the extremums of the curve of the slope of $S(k_c)$ in the right chart (where k_c =18, 22, 28 and 34 are represented by four arrows).

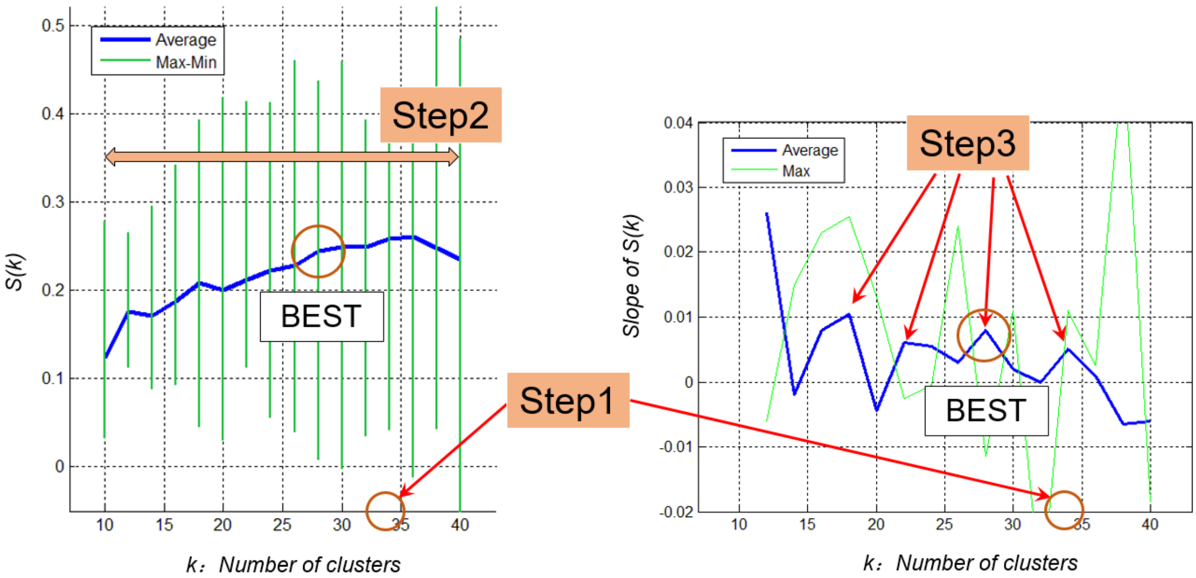

Fig. 3. Three steps to determine k_c

After these three steps, the best k_c can be determined: when k_c = 28 or 34, S_{ave} and S_{max} are both satisfying. However, the gap between S_{max} and S_{min} is larger around 34, and the curve of the slope of maximum $S(k_c)$ becomes much more unstable after k_c = 30 (the thin line in the right chart). Therefore, k_c = 28 is chosen. O&M managers can determine k_c by using this three-step method.

4. Outlier Detection

As the O&M phase covers a lengthy time span and involves numerous management activities, an increasing amount of structured and non-structured properties are added in BIM. For example, non-structured files, including design drawings, monitoring reports, repair logs, videos, and pictures, are usually attached to BIM elements. Most properties have to be manually imported or linked to building elements, usually causing a considerable error rate. The problem is, manual work in detecting improper properties is obviously tedious because of the huge amount of records involved. In order to automatically correct this kind of mistake and keep a clean database for further analysis, a detection process should be adopted. In this study, improperly matched properties or files are considered “outliers.” Outlier refers to records that are far away from the common ones (see the distance definition in Eq. (2)). An outlier detection towards improper files can be executed using DM methods.

4.1. Outlier detecting method

Outlier detection and clustering are closely related to each other. Usually, records from different clusters have different inner rules. For a certain cluster, records behave similarly, and are thus expected to have similar properties. A local density-based algorithm is utilized to find outliers. This algorithm contains four steps:

(1) For record $\boldsymbol{o}$, its k_n nearest neighbors ($\boldsymbol{o}_1$, …, $\boldsymbol{o}_{kn}$) are found, and their distances from $\boldsymbol{o}$: $dist(\boldsymbol{o}, \boldsymbol{o}_i)$ (i=1, …, kn) are calculated as defined in Eq. (2).

(2) The local density l_d of record $\boldsymbol{o}$ is calculated as:

$$l_d(\boldsymbol{o}) = \frac{k_n}{\sum_{i=1}^{k_n} dist(\boldsymbol{o}, \boldsymbol{o}_i)} \tag{6}$$

(3) Those kn nearest neighbors' local densities $l_d(\boldsymbol{o}_i)$, i=1, …, k_n, are calculated similarly as Eq. (6), using their respective neighbors.

(4) Record $\boldsymbol{o}$'s outlier coefficient u is defined as:

$$u(\boldsymbol{o}) = \frac{\sum_{i=1}^{k_n} l_d(\boldsymbol{o}_i)}{k_n \cdot l_d(\boldsymbol{o})} \tag{7}$$

A larger outlier coefficient of a record indicates this record is more probable to be an outlier. The distance scale $dist(\boldsymbol{o}, \boldsymbol{o}_i)$ is defined as in Eq. (2).

The outlier detection method works on every property that is involved in calculating distances. To further demonstrate the detecting process, a specific property "File" extracted from the BIM database is selected as an example. The distances between files are defined in the next section, and added in step (1) when calculating distances of neighbors.

4.2. Vector-based file distance

The similarity of files is measured by their identical keywords. The basic idea is that if two files have some common keywords, they are considered similar, and therefore contents of document files are transformed into word series before calculation. Considering that image and video content recognition is difficult, only title names and extensions are considered for multimedia files in this approach. To support this strategy, well-defined document management rules are required when establishing origin BIMs, and efforts should be taken to ensure the integrity and quality of these rules. For example, in the following discussions, naming the media files should observe the following rule: "[Discipline]-[Zone]-[Content title]-[Name of the objective element]-[Date].[Extension]", where "Discipline" is one of predefined department names, "Zone" is the corresponding spatial zone and "Date" is an eight-digit number. In this manner, "HVAC-ZoneC-Unusual flow curve-Pipe 195-20160322.jpg" is considered a proper file name.

First of all, keywords and relative themes are defined regarding involved professionals. A theme contains several keywords. The keyword definitions are stored in an Extensible Markup Language (XML) file (a segment is shown in Fig. 4). In the case study, more than 300 keywords of 105 themes were gathered from electrical, HVAC, water supply, and other common glossaries. The "WholeMatch" attribute of a keyword, as shown in the XML definition in Fig. 4, marks whether the word should be matched by all the letters or not. If WholeMatch=False, the keyword only needs to match the beginning letters of a word.

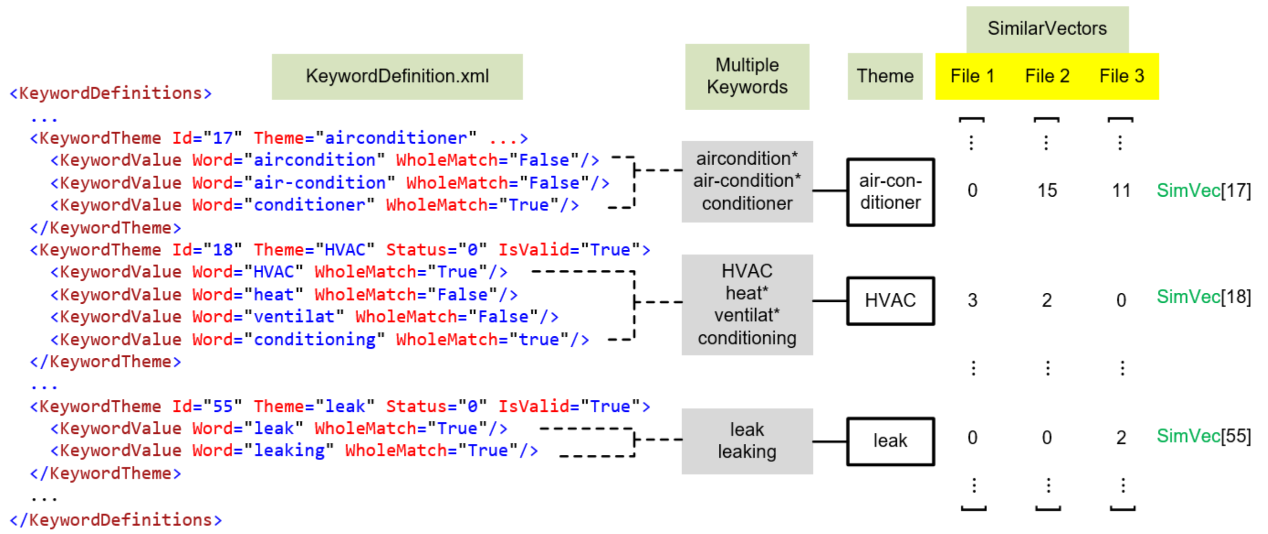

Fig. 4. The theme/keyword definitions and similarity vectors

The execution steps of the program are described in the following. Files are all transformed to word series. For each word in the series, if the word matches any keyword in a theme, the occurrence times of that series is added by one. The occurrence times of all defined themes are then arranged in a "similarity vector" (see the right part of Fig. 4). Let **x** and **y** stand for two files, and x_i and y_i are the i^{th} element in their similarity vectors. The distance of these two files is calculated as:

$$file\_dist(\mathbf{x},\mathbf{y}) = \frac{\sum not\_match(x_i, y_i)}{\sum positive\_match(x_i, y_i) + \sum not\_match(x_i, y_i)} \tag{8}$$

In the equation, *positive_match()* and *not_match()* are

$$\begin{array}{lll} \text{If } x_i > 0, y_i > 0 & : & positive\_match(x_i, y_i) = \min\{x_i, y_i\} \qquad not\_match(x_i, y_i) = 0 \\ \text{If } x_i y_i = 0 \text{ but not both } 0 & : & positive\_match(x_i, y_i) = 0 \qquad not\_match(x_i, y_i) = 1 \\ \text{If } x_i = 0, y_i = 0 & : & positive\_match(x_i, y_i) = not\_match(x_i, y_i) = 0 \end{array} \tag{9}$$

This measurement of distance is similar to the "Jaccard index" [44], frequently used in other research, except that negative matches are weakened in Eq. (8). The distances of files in Fig. 4 are (assume that all other elements in their similarity vectors are zero)

$$file\_dist(\mathbf{File}1, \mathbf{File}2) = \frac{1}{2+1} = 0.333 \qquad file\_dist(\mathbf{File}2, \mathbf{File}3) = \frac{1+1}{11+(1+1)} = 0.154$$

In addition, distance equals 1.000 when there are no positive matches (such as File 1 and File 3). The equation restricts the distance between two files must be between 0 and 1 (already normalized).

4.3. Evaluation and information interpretation method

All records are sorted by their outlier coefficients into a list. Managers can then check from the list from top to bottom to detect real mismatches of properties. To quantify this kind of fake detection and evaluate the accuracy of the algorithm, a method using two parameters, Universal Detection Order Rate (*UDOR*) and 80% Detection Order Rate (*80DOR*), is proposed. Let n be the total amount of real outliers, assuming their order in the list are r_i (i=1, …, n). First, the detection order (d_o) is defined as the geometric average of the orders that appears on the list:

$$d_o = \left(\prod_{i=1}^{n} r_i\right)^{1/n} \tag{10}$$

Then the best detection order ($d_{o,\text{best}}$) is defined as:

$$d_{o,\text{best}} = \left(\prod_{i=1}^{n} i\right)^{1/n} = \sqrt[n]{n!} \tag{11}$$

$d_{o,\text{best}}$ indicates the best situation that, all real outliers are detected in the front part of the list.

Finally, *UDOR* is defined as the ratio of $d_{o,\text{best}}$ to d_o:

$$UDOR = \frac{d_{o,\text{best}}}{d_o} \times 100\% \tag{12}$$

When only the front 80% of the real outliers are considered (i.e., i = 1, …, [0.8N] + 1), *80DOR* is similarly calculated. Since the last 20% may appear considerably irregular, *80DOR* can eliminate their influence, thus give a better estimation of detection accuracy. If *UDOR* and *80DOR* are both close to 100%, the detection result is of high reliability. For example, in the case study, UDOR was 74% and 80DOR was 91%, proving that the detection of improper files was valid for use. In summary, the outlier detection improves the quality of clusters and makes them ready for frequent pattern mining.

5. Cluster-based Frequent Pattern Mining Algorithm

After clustering analysis and outlier exclusion, data in BIM are divided into clusters with high quality. In this way, O&M managers can deal with only a few clusters, instead of thousands of individual records. Relationships of similarity among data records are provided, helping fast management decisions. However, apart from similarity, two kinds of patterns still exist: (1) causalities, in which one event is the result of another event; and (2) some events are related to one another. In other words, some events can increase the probability of other events. Given that the two relationships provide further comprehension about data, finding these frequent patterns is important for decision making.

Since first proposed two decades ago [45], the classic Apriori algorithm has been widely used in finding frequent patterns. The basic principle of the classic Apriori is that all subsets of a frequent set is naturally frequent. Therefore, the core issue is to find out largest frequent sets. This process in fact involves complex operations, which is not discussed in detail in this paper but can be found in other bibliographies such as reference [7]. However, the classic Apriori algorithm involves some extremely expensive temporal steps. For example, generating and testing all the subsets is an exponential calculation. Frequent pattern mining algorithm based on cluster improves the classic Apriori especially on temporal complexity.

This study proposes a cluster-based frequent pattern mining algorithm based on Apriori. First, some basic definitions are given:

Definition 1. A ***status*** is a 'property=value' pair.

Definition 2. $\mathscr{S}$ is the universal set of all possible statuses.

Definition 3. A ***status set*** $S_k \subseteq \mathscr{S}$ is a set of k statuses.

Subfunction The ***support count*** of a status set S

$s_c = \boldsymbol{sup\_count}(S)$:

  $s_c = 0$ // initialize

  foreach *record* in all records

    if *record* has status S

      then $s_c = s_c + 1$

  return s_c

Definition 4. Giving a positive integer *minsup*, S is ***frequent*** if $sup\_count(S) \geq minsup$. ***Frequent status sets*** are represented by the symbol F.

Definition 5. A ***frequent pattern*** is a logic implication $(F)_\mathrm{A} \Rightarrow (F)_\mathrm{B}$ only when $(F)_\mathrm{A} \cap (F)_\mathrm{B} = \emptyset$

The logic implication of a frequent pattern means that if a record has all statuses in $(F)_\mathrm{A}$, it will have status in $(F)_\mathrm{B}$. Once the largest frequent set is founded, all candidate frequent patterns can be generated using the subsets of the largest frequent set. If a pair $\{(F)_\mathrm{A}, (F)_\mathrm{B}\}$ are exclusive, they can form a frequent pattern (see Definition 5). Sometimes, $(F)_\mathrm{A}$, $(F)_\mathrm{B}$ are irrelevant, and are thus meaningless. In order to find those meaningful patterns, they must pass the "confidence test" (Eq. (13)) and "correlation coefficient test" (Eq. (14)). Only when a pattern $(F)_\mathrm{A}$=>$(F)_\mathrm{B}$ passes both tests is it output as a strong pattern.

$$C((F)_\mathrm{A} \Rightarrow (F)_\mathrm{B}) = \mathrm{P}((F)_\mathrm{B}|(F)_\mathrm{A}) = \frac{sup\_count((F)_\mathrm{A} \cup (F)_\mathrm{B})}{sup\_count((F)_\mathrm{A})} > C_{\min} \tag{13}$$

$$R((F)_\mathrm{A}, (F)_\mathrm{B}) = \frac{\mathrm{P}((F)_\mathrm{A} \cup (F)_\mathrm{B})}{\sqrt{\mathrm{P}((F)_\mathrm{A}) \times \mathrm{P}((F)_\mathrm{B})}} = \frac{sup\_count((F)_\mathrm{A} \cup (F)_\mathrm{B})}{\sqrt{sup\_count((F)_\mathrm{A}) \times sup\_count((F)_\mathrm{B})}} > R_{\min} \tag{14}$$

Where P() is the probability: support count divided by the amount of all records. The limitation $C_{\min}$ and $R_{\min}$ are both given before analyzing. In practice, analysts should try various combinations of $C_{\min}$ and $R_{\min}$ to obtain acceptable results. Finally, $(F)_\mathrm{A}$ is marked as the condition, and $(F)_\mathrm{B}$ is the consequence.

Then the main idea of the cluster-based algorithm is that clustering can be preprocessed before pattern analysis. The algorithm to generate frequent status sets with cluster centers is described in the pseudocode below.

Algorithm Cluster-based Frequent Pattern Mining

Input: $s_{c,\min}, C_{\min}, R_{\min}, s^{\mathrm{c}}_{c,\min}$

Step 1. Generate the clusters' largest frequent status sets

After clustering, all records are divided into clusters. As for the ith cluster:

1. Let $(S)_i = \emptyset$ // initialize
2. Make single-element status sets from this cluster's center

   $(S_1)_1 = \{prop_1 = value_1\}, ..., (S_1)_{n_{prop}} = \{prop_{n_{prop}} = value_{n_{prop}}\}$
3. **foreach** $j = 1, ..., n_{prop}$

   if $sup\_count(S_1)_j$ in this cluster $> s^{\mathrm{c}}_{c,\min}$

   then merge $(S)_i$ and $(S_1)_j$

Step Output $(S)_i$ as the ith cluster's longest frequent status set

Step 2. Find strong patterns

foreach S in all clusters' longest frequent status sets

1. Let $\mathscr{F} = \emptyset$ // initialize
2. **foreach** subset $S_k \subseteq S$

   if $sup\_count(S_k)$ in all records $> s_{c,\min}$

   then S_k is F_k, add it to $\mathscr{F}$
3. **foreach** $(F)_{\mathrm{A}}, (F)_{\mathrm{B}} \in \mathscr{F}$

   Do the confidence test and the correlation coefficient test

Output each qualified $(F)_{\mathrm{A}} \Rightarrow (F)_{\mathrm{B}}$

First of all, based on a cluster's center, some single-property status sets are generated. Each set contains one property from the cluster center. Then, each set's support count inside this cluster is calculated. Finally, those sets whose counts are less than $s_{c,\min}{}^{\mathrm{c}}$ are eliminated, and the remaining sets are merged as one of the largest frequent status sets. Time complexities before and after improvement are shown in Table 2, where n_{prop} is the number of properties. Generating the largest status sets is the speed-determining step in the classic Apriori, while cluster-based processing makes it much faster because exponential calculations are avoided. Although testing strong patterns is slower than the classic Apriori, the overall time cost is obviously still decreased. Considering only the speed-determining steps, the proposed algorithm is approximately $n_{prop} \cdot 2^{n_{prop}}/k_c$ times faster than the classic Apriori algorithm.

Table 2. Time complexity before and after improvement

Steps:	Generating longest sets	Testing strong patterns
Classic Apriori	$n_r \cdot n_{prop} \cdot 2^{2n_{prop}}$ (speed-determining)	$k_c \cdot 2^{n_{prop}}$
Cluster-based algorithm	$n_r\ n_{prop}$	$k_c \cdot n_r \cdot 2^{n_{prop}}$ (speed-determining)

However, a cluster center only contains the main information of this cluster, and cannot cover all the records in this cluster. Therefore, the improved algorithm may miss some information. Only when clustering quality is acceptable, the center record is enough qualified for representing the whole cluster. This indicates the further value of high-quality clusters.

6. A Case Study of an Airport Terminal

With the three information mining approaches mentioned above, an integrated application on a real BIM data set from a large public building is then implemented. After the DM process, the output results are evaluated, and the process in which O&M managers can utilize the mined information is discussed.

6.1. Case overview

The proposed information mining approach was applied to the new terminal of Kunming Changshui International Airport. The terminal, with a total building area of 435,400 square meters, is one of the largest airport terminals in China. It consists of four floors above the ground and three floors underground. The modeling and application steps are introduced below and illustrated in Fig. 5.

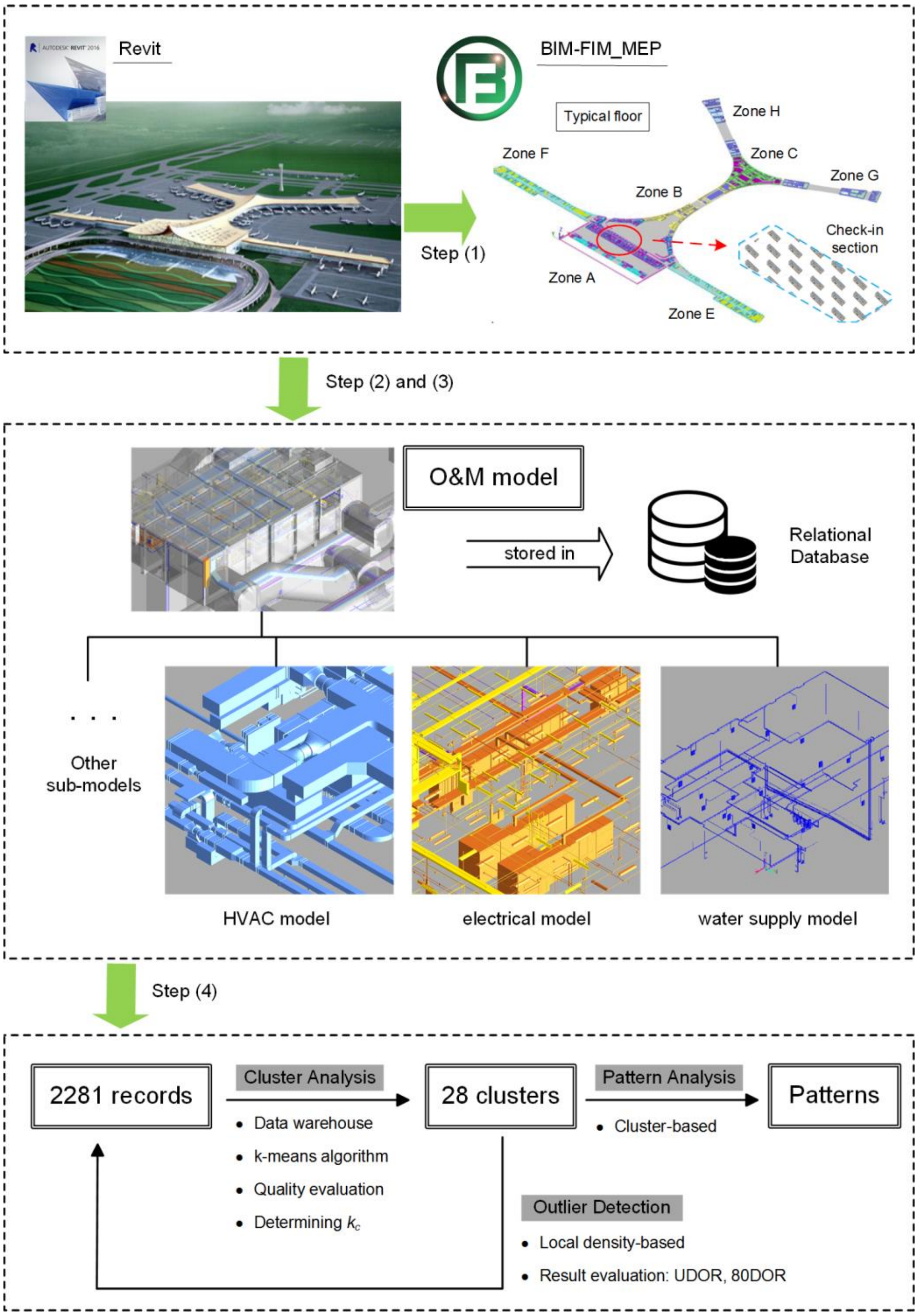

Fig. 5. The modeling and application steps

(1) An as-built BIM was established by the constructors of the project according to the design model (3D architecture model and structure model) in Autodesk Revit™.

(2) All data were transferred and imported into the BIM-FIM_MEP system [46], a BIM-based facility management system, to build an O&M model. The BIM-FIM_MEP system realized the integrated delivery of the

Mechanical, Electrical, and Plumbing (MEP) model from the construction phase to the O&M phase. Moreover, it provided a platform that enabled O&M functions and ensured the safe operation of all MEP systems. Some crucial information, including O&M records and upstream/downstream relationships, were also integrated into the O&M model.

(3) Three sub-models were mainly examined, namely, HVAC, electrical supply, and water supply models. The analyzed data were obtained from the database of the BIM-FIM_MEP. The core data repository was stored in a typical relational database.

(4) As described in Section 3.1, 2281 records were transformed before analysis. Then all three DM methods were executed in a predefined flow and output the final result.

Some necessary preprocessing, including normalization and discretization, were performed. Each record contained 19 properties after data preprocessing. In Table 3, all properties and three examples of original data from the data warehouse are listed. These properties mainly came from three data tables in the database: "repair records", "maintenance records" and "spatial structures". The logic chain among MEP elements was also important when finding related elements; thus, two properties (upstream and downstream[2]) were included in indexing to upstream and downstream elements of the current record. Property "File" was read from file data tables (binary files).

Table 3. Properties of records and some examples

Properties (Type)	Example 1	Example 2	Example 3
Elem_ID (uint)	1290335	1290337	423826
Date (DateTime)*	2016/2/26	2016/5/11	2016/1/16
Weekday (enum)	Wednesday	Sunday	Thursday
Time (enum)*	Morning	Night	Night
Repair_ID (uint)	0	1344	0
Rep_worker (string)	null	TianPL	null

[2] Upstream and downstream both refer to the connection relationships inside MEP systems. For example, if the water system supplies from A->B->C, A is the upstream component of B, and C is the downstream component of B.

Rep_severity (enum)*	Not Rep	Slight	Not Rep
Maintenence_ID (uint)	247	0	148
Maint_worker (string)	WangXing	null	ZhuJT15
Maint_severity (enum)*	Serious	Not Maint	Slight
Storage (bool)*	Not Used	Used	Not Used
Type (enum)	Water_pump	Air_cond	Elec_appliance
Department (enum)	Facility_mgnt	HVAC	Electrical
Elevation (double)	13.15…	-5.21…	19.30…
Zone_name (enum)	ZoneA	ZoneB	ZoneB
GUID (Guid)	1f2e5216-030d-4e01-a3b5-0c3de10a3676	dfac0432-5b36-41e1-8fad-191ecdfc0e13	95888cc4-a08e-471a-abab-ab1ec0611ccf
Upstream (uint)	1290334	1290336	0
Downstream (uint)	1290336	1290338	290763
Status (bool)	Finished	Finished	Finished
File (binary)	(some files)	(some files)	null

Note: * designates a property that was transformed before analysis (described in Section 3.1).

6.2. Information mining results

Cluster analysis was first executed. The value of coefficient k_c was already determined as 28 (see Section 3.4), indicating that these records would be divided into 28 clusters. As the iteration seeds were randomly chosen, the algorithm was run several times to get high-quality clusters. In each run, about 30 iterations were processed, with a total time of 2–4 seconds on a mid-range desktop. In this case, 14 clusters had S above 0.200, and only 4 clusters below 0.100. In total, $S_{\max}$ and S_{ave} were 0.365 and 0.184, respectively, showing that k_c=28 was reasonable. One of the big high-quality clusters was the No. 17 cluster, containing 45 records. Table 4 shows the counts of occurrence of the most common property values and percentages among all records in this cluster. These records were generally 90% similar to each other, especially in time, location and repair contents. Some relationships about similarity could be inferred from this cluster. For instance, many electric appliances in Zone A often stopped working in the afternoon during March and April, and always coincided with repairs of upstream elements (usually near some electric brakes). This piece of information was then sent to the electrical department. They checked for the power flow curve of related power supply system and found that actual electrical load was far higher than designed. In winter, coffee boilers and heaters were used at work, so their upstream element—the magnet protection system—was often tripped. Finally, this was marked as a daily checking task in the BIM-FIM_MEP, and those corresponding workers were informed.

Table 4. Detail of a typical high-quality cluster (No. 17)

Value of properties	Count (total=45)	Percentage
November 15 to December 15	27	60%
Time: afternoon	42	93%
Medium malfunction	43	96%
Storage used	43	96%
Element: electric appliance	27	60%
Major: electrical	44	98%
Elevation: 16m to 24m	28	62%
Location: Zone A	43	96%
Upstream component repaired	44	98%

Other high-quality clusters provided other relationships among records, for example, one indicated a similarity about repair date, operator's name, and severity. This helped optimization in human resource planning. It was estimated by operators that the information from DM saved about half the human work time when conducting repairs in the airport terminal.

After clustering, improper files were detected through methods shown in Fig. 6. The coefficient k_n was 10. The total calculation time was about 30 seconds on a mid-range desktop. Detection results were sorted by outlier factor in a list, and managers then searched backward to related elements and attached files. After detection, managers went on checking the records from the top of the output list and modified improper files. Fifteen outliers were found among the top 100 records. Assuming all others were not outliers, UDOR was 74% and 80DOR was 91%. This result proved the feasibility and validity.

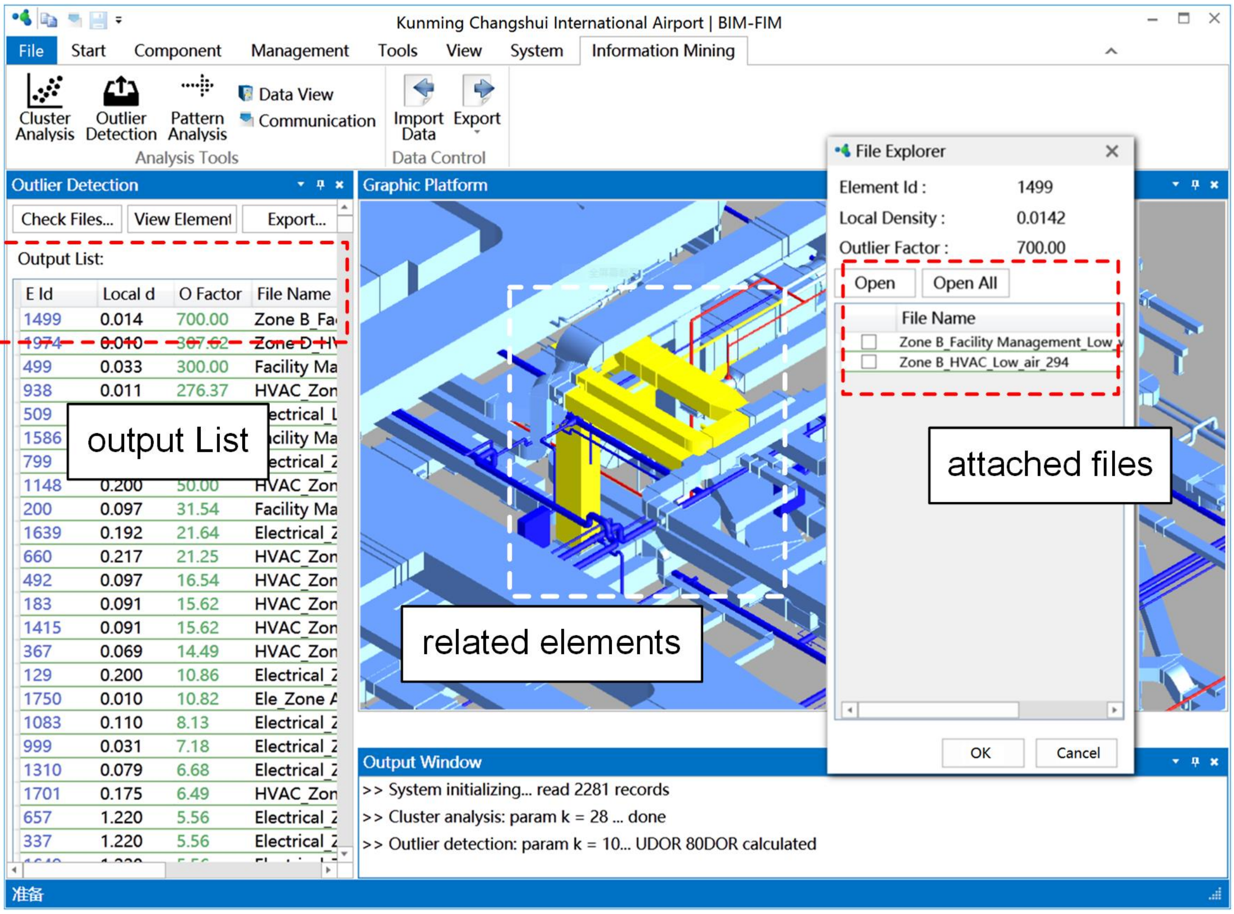

Fig. 6. User interface of outlier detection

Frequent pattern mining was then executed after correcting the improper files. To further accelerate the calculation, clusters with lower *S* were not accepted in pattern analysis. According to the improved algorithm, the centers of these clusters were directly used for generating frequent status sets. To avoid too many patterns being found, strict limitations were chosen: $s_{c,\min}^{c}$, s_c, $C_{\min}$ and $R_{\min}$ were set as 50, 400, 0.900 and 0.800 respectively. In addition, conditions and consequences with more than three and two items are accepted. Finally, 201 frequent patterns were generated after a one-minute run.

Table 5 lists a typical pattern. The three-item condition indicated that when the discipline was facility management, the location was Zone B and the downstream component was repaired, it can be inferred that the malfunction was likely to be slight, the storage was possibly not used, and the upstream component was usually repaired. *C*() and *R*() were calculated at the same time. In this pattern, the consequence had a probability of 93.9% when the condition happened indicating that these two cases were strongly and positively related.

Table 5. Detail of a typical frequent pattern

Output item	Content
Condition	Facility Management, Zone B, Downstream repaired
Consequence	Slight malfunction, Storage not used, Upstream repaired
Confidence	0.939
Cos coefficient	0.945

Managers obtained two pieces of useful management advice through this pattern. First, most repair and maintenance operations of facilities in Zone B were relatively slight and storage was not required, thus the storage room for the facility department was arranged in other zones away from Zone B, in order to make space management more flexible. Second, upstream and downstream records often happened together, indicating that facilities in this section had experienced a large area of failure instead of occasional repairs. Analysts issued a warning according to this information, and managers carried out a complete investigation towards these components. Most patterns had C close to 1.00 and R over 0.90, indicating strong frequency and good relationships of causality. The proposed hybrid DM approach had provided about 100 findings in total for the airport terminal where most of them were proved meaningful in site work. Table 6 shows the amount of useful findings (54 accepted suggestions and 19 handled warnings) which provide suggestions to space management, material optimization, repair and maintenance, and other O&M activities.

Table 6. Amount of meaningful **findings for the airport terminal**

Content	All findings	Handled warnings	Accepted suggestions	Useful findings
Space management	25	8	12	80%
Repair and maintenance	37	6	26	86%
Material planning	19	2	11	68%
Human resource	6	1	4	83%
Other activities	3	2	1	100%

7. Discussion

Three information mining methods have been introduced and implemented in a case study focusing on repair and maintenance data, illustrating that the proposed approach is suitable for analyzing records generated during the O&M phase: (1) cluster analysis can find direct relationship among records; (2) outlier detection improves the qualities of clusters; and (3) the improved pattern mining algorithm helps in finding some implicit logics among management tasks. Currently BIMs have many features of big data. The presented case study is about 50 GB in database size, which may be considered an example of big data. However, except for geometric information and embedded properties of each element, it contains no more than 5000 maintenance records generated in the past 3 years, and only half of those records are valid for DM. From this perspective, it is far from a big data problem. Regardless, DM skills show more advantages particularly in timely knowledge discovery thus the proposed approach is expected to provide basis and useful tools for big data problems.

The facility managers of the terminal appraised the DM for improving not only the work efficiency, but also space utilization. Within these applications, data played a core role throughout the whole DM process with no doubt. However, the way to integrate data sources from O&M remains a problem. BIM data standards are not yet perfect and O&M management is not thoroughly standardized at the present stage. Furthermore, monitoring data is important but the integration between self-contained building automation systems (BAS) and a new BIM platform may pose a challenge, making it difficult to obtain real-time information. Some difficulties also occurred in the pure DM algorithm. Both outlier detection and frequent pattern analysis are time-consuming processes. Although preprocessing by clustering helps accelerate the process, hours may still be needed when the data set is considerably large. In addition, initial condition determination and result interpretation both require expert knowledge, indicating that the proposed method cannot be fully automated. Specialists must participate in the process, and this additional requirement leads to the need for extra budget. In the near future, the proposed approach will be further studied based on these mentioned problems, including four aspects below:

- Algorithm complexity should be further optimized. At the same time, definitions of the data warehouse are expected more flexible in order to limit information loss. Some automatic result interpretation methods should be developed based on the specific application scene of the O&M phase.
- The Internet of Things (IoT) technology and BAS monitoring system should be integrated. Given that

DM is strong in analyzing massive data, data from the Internet of Things and BAS can provide rich information to find more patterns.

- Cloud platforms have provided new mechanisms for BIM. This study, as a possible extension for cloud BIM, focused on deeper data analysis and provided further information by DM methods. With the proposed approach, cloud BIM will be able to represent more valuable information for users.
- Data analysts should grasp AEC knowledge in addition to the acquisition of DM skills. Therefore, professional training is essential for site workers. On the other hand, missing data and lack of discipline in BIM database will severely confuse algorithms. Efforts should be put on to ensure the accuracy of data when establishing the as-built BIM.

8. Conclusion

Big data are heavily generated and gathered from daily O&M activities of buildings. These data can be managed by BIM platforms for better interoperability, and have the potential to provide deep information, helping improve facility management. However, data sets are always non-intuitive due to the complex inside relationships. In addition, inaccurate data in BIM databases can lower the data quality, and negative implications to management activities.

To address these problems, this study proposes a hybrid BIM-based DM approach for extracting meaningful laws and patterns as well as detecting improper records. In this approach, a BIM database is first transformed into a data warehouse. After that, three information mining methods are combined to find useful information from BIM data sets: (1) Cluster analysis can find relationships of similarity among records. A standard clustering process is proposed and the four kinds of clustering results and their features are discussed. A cluster silhouette coefficient is introduced to evaluate the quality of clusters, and the parameter k_c is determined by a three-step method. (2) Outlier detection detects improper manually input data and keeps the database fresh. Two new parameters (*UDOR* and *80DOR*) are also proposed to evaluate the detection. (3) A cluster-based algorithm on temporal complexities is proposed to find deep logic links among records. To improve the slow steps in classic Apriori algorithm, cluster centers are used as sources to generate the largest frequent status sets. Particular emphasis is put on introducing the improved algorithm and how they should be used by building managers.

An integrated on-site case in a real-world airport terminal was conducted to evaluate the proposed approach. O&M data were first transformed into 2281 records in the data warehouse. These records were divided into 28 clusters, in which 14 clusters were considered high quality. As a typical user case, when dealing with a big high-quality cluster, a daily checking task towards the magnet protection system was suggested and the corresponding departments accepted the suggestions. After clustering, improper files as well as other data were detected through outlier detection. Fifteen outliers were corrected among the top 100 records. UDOR was 74% and 80DOR was 91%. Finally, in pattern analysis, 201 useful patterns were found. For example, as indicated by a pattern, the storage room in Zone B was arranged in other zones, making the space management more reasonable.

The proposed approach had provided about 50 suggestions and 20 warnings in total for O&M staffs of the airport terminal. The results demonstrated that the hybrid DM method is helpful in prediction, early warning, and decision making, leading to the improvements of resource usage and maintenance efficiency during the O&M phase.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (No. 51478249 and No. 51778336) and the Tsinghua University-Glodon Joint Research Centre for Building Information Model (RCBIM). The authors would like to acknowledge the FM office of the Kunming Changshui Airport for providing the application case.

References

[1] T. Cerovsek, A review and outlook for a ‘Building Information Model’ (BIM): A multi-standpoint framework for technological development, Advanced Engineering Informatics 25 (2) (2011) 224-244.

[2] R. Volk, J. Stengel, F. Schultmann, Building Information Modeling (BIM) for existing buildings — Literature review and future needs, Automation in Construction 38 (2014) 109-127.

[3] U. Isikdag, J. Underwood, G. Aouad, N. Trodd, Investigating the Role of Building Information Models as a Part of an Integrated Data Layer: A Fire Response Management Case, Architectural Engineering and Design Management 3 (3) (2007) 124-142.

[4] M. Bilal, L.O. Oyedele, O.O. Akinade, S.O. Ajayi, H.A. Alaka, H.A. Owolabi, J. Qadir, M. Pasha, S.A. Bello, Big data architecture for construction waste analytics (CWA): A conceptual framework, Journal of Building

Engineering 6 (2016) 144-156.

[5] M. Bilal, L.O. Oyedele, J. Qadir, K. Munir, S.O. Ajayi, O.O. Akinade, H.A. Owolabi, H.A. Alaka, M. Pasha, Big Data in the construction industry: A review of present status, opportunities, and future trends, Advanced Engineering Informatics 30 (3) (2016) 500-521.

[6] K. Orr, Z. Shen, P.K. Juneja, N. Snodgrass, H. Kim, Intelligent Facilities - Applicability and Flexibility of Open BIM Standards for Operations and Maintenance, Construction Research Congress, 2014, pp. 1951-1960.

[7] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., 2011.

[8] J.R. Lin, Z.Z. Hu, J.P. Zhang, F.Q. Yu, A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM, Computer-Aided Civil and Infrastructure Engineering 31 (1) (2015) 18-33.

[9] S. Liao, P. Chu, P. Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications 39 (12) (2012) 11303-11311.

[10] H. Liu, M. Al-Hussein, M. Lu, BIM-based integrated approach for detailed construction scheduling under resource constraints, Automation in Construction 53 (2015) 29-43.

[11] K.K. Han, M. Golparvar-Fard, Potential of big visual data and building information modeling for construction performance analytics: An exploratory study, Automation in Construction 73 (2017) 184-198.

[12] A. Deshpande, S. Azhar, S. Amireddy, A Framework for a BIM-based Knowledge Management System, Procedia Engineering 85 (2014) 113-122.

[13] K. Kim, G. Kim, D. Yoo, J. Yu, Semantic material name matching system for building energy analysis, Automation in Construction 30 (2013) 242-255.

[14] X. Xiong, A. Adan, B. Akinci, D. Huber, Automatic creation of semantically rich 3D building models from laser scanner data, Automation in Construction 31 (2013) 325-337.

[15] M. Peña, F. Biscarri, J.I. Guerrero, I. Monedero, C. León, Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach, Expert Systems with Applications 56 (2016) 242-255.

[16] N. Li, B. Becerik-Gerber, Performance-based evaluation of RFID-based indoor location sensing solutions for the built environment, Advanced Engineering Informatics 25 (3) (2011) 535-546.

[17] C. Ko, RFID-based building maintenance system, Automation in Construction 18 (3) (2009) 275-284.

[18] A. Krukowski, D. Arsenijevic, RFID-based positioning for building management systems, International Symposium on Circuits and Systems, 2010, pp. 3569-3572.

[19] H. Chen, K. Chang, T. Lin, A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs, Automation in Construction 71 (2016) 34-48.

[20] M. Gajzler, Knowledge Modeling in Construction of Technical Management System for Large Warehousing Facilities, Procedia Engineering 122 (2015) 181-190.

[21] J.A. Abdalla, K.H. Law, A Framework for a Building Energy Model to Support Energy Performance Rating and Simulation, International Conference on Computing in Civil and Building Engineering, 2014, pp. 227-234.

[22] A. Costa, M.M. Keane, J.I. Torrens, E. Corry, Building operation and energy performance: Monitoring, analysis and optimisation toolkit, Applied Energy 101 (2013) 310-316.

[23] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy and Buildings 75 (2014) 109-118.

[24] S. Wang, W. Wang, K. Wang, S. Shih, Applying building information modeling to support fire safety management, Automation in Construction 59 (2015) 158-167.

[25] A. Sagun, D. Bouchlaghem, C.J. Anumba, Computer simulations vs. building guidance to enhance evacuation

performance of buildings during emergency events, Simulation Modelling Practice and Theory 19 (3) (2011) 1007-1019.

[26] J. Chou, N. Ngo, Smart grid data analytics framework for increasing energy savings in residential buildings, Automation in Construction 72 (2016) 247-257.

[27] C. Nicolle, C. Cruz, Semantic Building Information Model and Multimedia for Facility Management, 6th International Conference on Web Information Systems and Technologies, Springer Berlin Heidelberg, Valencia, Spain, 2010, pp. 14-29.

[28] Z. Hu, X. Zhang, X. Chen, J. Zhang, A BIM-based research framework for monitoring and management during operation and maintenance period, 14th International Conference on Computing in Civil and Building Engineering, Moscow, Russia, 2012.

[29] F. Fassi, C. Achille, A. Mandelli, F. Rechichi, S. Parri, a New Idea of Bim System for Visualization, Web Sharing and Using Huge Complex 3d Models for Facility Management., ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-5/W4 (5) (2015) 359-366.

[30] A. Kusiak, M. Li, Z. Zhang, A data-driven approach for steam load prediction in buildings, Applied Energy 87 (3) (2010) 925-933.

[31] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Automation in Construction 49 (2015) 1-17.

[32] Z.J. Yu, F. Haghighat, B.C.M. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, Energy and Buildings 47 (2012) 430-440.

[33] S. D'Oca, T. Hong, A data-mining approach to discover patterns of window opening and closing behavior in offices, Building and Environment 82 (2014) 726-739.

[34] J. Zhao, B. Lasternas, K.P. Lam, R. Yun, V. Loftness, Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining, Energy and Buildings 82 (2014) 341-355.

[35] C. Cheng, S. Leu, Y. Cheng, T. Wu, C. Lin, Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry, Accident Analysis & Prevention 48 (2012) 214-222.

[36] J. Jeong, T. Hong, C. Ji, J. Kim, M. Lee, K. Jeong, C. Koo, Development of a prediction model for the cost saving potentials in implementing the building energy efficiency rating certification, Applied Energy 189 (2017) 257-270.

[37] A. Ahmed, M. Otreba, N.E. Korres, H. Elhadi, K. Menzel, Assessing the performance of naturally day-lit buildings using data mining, Advanced Engineering Informatics 25 (2) (2011) 364-379.

[38] H. Son, C. Kim, Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches, Journal of Cleaner Production 109 (2015) 144-151.

[39] T.P. Williams, J. Gong, Predicting construction cost overruns using text mining, numerical data and ensemble classifiers, Automation in Construction 43 (2014) 23-29.

[40] P. Carrillo, J. Harding, A. Choudhary, Knowledge discovery from post-project reviews, Construction Management and Economics 29 (7) (2011) 713-723.

[41] W.H. Inmon, Building the Data Warehouse,3rd Edition, John Wiley & Sons, Inc., 2002.

[42] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. Mclachlan, A. Ng, B. Liu, P.S. Yu, Top 10 algorithms in data mining, Knowledge and Information Systems 14 (1) (2008) 1-37.

[43] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, DBLP, 1990.

[44] P. Jaccard, THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE, New Phytologist 11 (2) (1912)

37-50.
[45] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, International Conference on Very Large Data Bases, 1994, pp. 487-499.
[46] Z. Hu, X. Zhang, H. Wang, M. Kassem, Improving interoperability between architectural and structural design models: An industry foundation classes-based approach with web-based tools, Automation in Construction 66 (2016) 29-42.