A clustering-based method of typical architectural case mining for architectural innovation

Shuyu Liu, Guangtian Zou and Si Zhang

Abstract
Architectural innovation is important for improving the built environment. In recent years, an increasing number of architects have focused on this field. A comprehensive understanding of existing building design patterns contributes to reasonable innovation. Based on the large number of architectural cases on the internet in the big data era, this study proposes a three-stage case mining method, containing case collection, case analysis and case study, to find typical architectural cases, discover existing design patterns and create new design patterns by using cluster analysis of architectural cases. An extensive architectural design case mining system and a case clustering program are developed to assist in the case analysis. An agglomerative hierarchical clustering algorithm is applied to support the case clustering program. The example shows the complete application process and practical effect of the proposed method. With this intelligent method, architects can make more reasonable innovations in projects. The proposed typical case mining method is also expected to be useful for engineers and planners with similar needs.

1. Introduction
Architectural innovation is an important approach for improving the quality of human settlement (Wu 2012) and one of the core themes of extensive architecture (Wang and Zou 2011). Especially in recent years, innovation has received unprecedented attention from architects (Chenping Han 2015). Since a complete understanding of existing knowledge is an essential basis for reasonable innovation (Quintane et al. 2011), case studies are considered a necessary prerequisite for architectural innovation (Dongqing Han 2007). However, there are too many architectural cases on the internet in the big data era (Furundzic 2003). Due to time and effort limitations, architects are unable to study every relevant case when they need to innovate in a project. Therefore, it would be efficient for innovation if architects could understand the main design patterns of existing buildings by studying few typical cases. Specifically, architects need a method for finding typical cases that represent the main design patterns and discovering the patterns within an acceptable amount of time.

Case searching is a common method for obtaining information for specific purposes. Its process can be described as follows. The user inputs criteria to obtain links to pages or records of spatially delimited units observed at a point or period of time which meet the criteria (Gerring 2007; Weitz 2014). There are a number of methods and systems developed for case searching that are not limited to architectural cases. Schank (1982) proposed a case-based reasoning (CBR) theory to search the most valuable cases according to the similarity between cases and the current project. The CBR theory is currently the core case searching method.

To accurately search for cases, Domeshek and Kolodner (1993) developed the Archie system for architects to search for cases with attributes on five dimensions, including design issues, building space, functional components, stakeholders, and life cycle phases. Chiu and Tsai (2007) introduced a weighted feature c-means clustering algorithm to weight the similarity-related attributes so that the search result would be unbiased. Artificial neural networks (Li and Feng 2004) and genetic algorithms (Huang, Lei, and Li 2004) were also used to optimize the weights of the attributes. Wang and Shao (2000) made the weight of each attribute flexible for different problems by introducing rough set theory into the CBR system. Zidi et al. (2014) proposed a personalization method to provide an efficient personalized search result and improve the user’s satisfaction based on a combination of domain ontology and CBR tools. Wei et al. (2015) designed an Architable system to search for architectural cases according to the rooms and functional layout input by the user. They proposed that spatial relationships are a better basis than building attributes in architectural case searching.
In addition to accuracy, efficiency is another important concern of research in case searching. Stephane, Hector, and Marc (2010) proposed a generic method for case-base indexation coupled with a special research algorithm to reduce the processing time during searching. Nouaouria, Boukadoum, and Proulx (2014) presented an associative memory model based on particle swarm optimization (PSO) to flexibly search for cases according to both similarity and adaptability. Zhao et al. (2017) developed the FCM retrieval model to improve the speed and efficiency of case searching in matching a large-scale case database.

Other efforts have been made to extend the scope of case searching. Heylighen et al. (2007) designed DYNAMO v.5.1 to share architectural projects with students and professional architects through a web browser. Wang, Jia, and Zhu (2014) presented a web-based remote customization platform that can utilize internet resources. Xiong et al. (2015) developed a cloud platform to search the most similar cases from many databases, which makes multisource case searching possible. Chang, Lee, and Wang (2016) established a new framework that includes a collaborative filtering mechanism and a semantic-based case searching agent so that target cases can be found according to natural language as indices and attributes. Cha and Lee (2018) proposed a BIM-based framework to link construction data with 3D/BIM objects. On the basis of construction information that is broken down into various levels and categories, users can search for cases through the details of the project execution phase.

The abovementioned studies have made comprehensive contributions to the effectiveness and application of case searching. However, because most of them only focus on the similarity between the case and the current project due to the limitation of CBR theory, users are supposed to describe the main features of the target case based on the current project before searching. In other words, researchers prefer to search for cases through a retrieval method in these studies. For most architects who do not know exactly which cases represent the main design patterns of existing buildings before searching, the methods mentioned above cannot be used effectively to search typical cases for architectural innovation.

According to the similarity between cases, Yang and Wu (2001) introduced cluster analysis to compress a large case-base into small case clusters. By preferably searching cases from clusters that are similar to the user’s target, the time consumption of case searching was significantly reduced. Based on the same idea, grid clustering (Jia, Huang, and Ma 2009), k-means clustering (Chen et al. 2010) and hierarchical clustering (Li et al. 2014) have also been adopted. Qiao, Jiang, and Jia (2011) presented a user-feedback approach to optimize the case cluster so that the searched case would be more helpful for the user’s problem. Although these works focused on intercase similarity and introduced cluster analysis, clustering was only used to assist rather than replace retrieval. In essence, these methods can be considered clustering-improved versions of case retrieval. For architects who need to innovate in a new field, they still cannot find typical cases without detailed preset criteria using such methods.

This study is a primary part of the research on Extensive Architectural Design Data Mining (EADDM), which studies how to find helpful cases from the internet for the diverse needs of architects (Liu, Zou, and Zhang 2017). The objective of this study is to provide a typical case mining approach for reasonable architectural innovation, which helps architects find typical cases without detailed descriptions in advance, discover the main design patterns of existing buildings, and create new patterns in projects. For specific innovation requirements, typical architectural cases are found by clustering relevant cases from the internet. More specifically, to achieve this goal, a three-stage case mining process is established, including case collection, case analysis and case study. Based on this process, we propose the method with these four steps. (1) Twenty similarity-related attributes of architectural cases are identified and set as the targets of case collection. (2) An extensive architectural design case mining system is developed to store, retrieve and evaluate the collected cases and to calculate the intercase similarities. (3) An agglomerative hierarchical case clustering program is developed to assist the system with architectural case clustering. (4) A guide of the typical case study is presented for design pattern discovery and creation. To validate the method, an innovation requirement of a small museum is assumed as an example to test and illustrate its application effect. The main progress and potential limitations of this method are also discussed.

2. Mechanism
2.1. Design pattern and typical case
Architectural cases consist of many characteristics, such as towerng and transparency, which describe a case in several aspects. If most of the characteristics in different cases are the same or similar, there is a high degree of similarity between these cases. That is, a group of cases with high similarities between each other can be regarded as the same design pattern. In such a group, the central case usually represents this design pattern because it is the most similar to every other case. Therefore, the central case can be considered the typical case of the group. Figure 1 shows the relation of design patterns and typical cases. Considering the typical case, architects can generally understand its design pattern. Furthermore, architects can efficiently discover the main patterns of existing buildings by studying typical cases of major groups.
2.2. Clustering

Clustering is an unsupervised analysis method of data mining. According to the similarities between samples, the samples should be divided into a series of clusters with maximum intracluster similarities and minimum intercluster similarities by clustering (Han, Kamber, and Pei 2012). Obviously, the clusters of similar samples are equivalent to the groups of cases in the same patterns. By applying clustering on architectural cases, the groups of design patterns can be divided automatically. Typical cases can also be found in the groups.

3. The process of typical architectural case mining

The process of typical architectural case mining contains three stages: case collection, case analysis, and case study. In the first stage, similarity-related data of the architectural case are collected from the internet. In the second stage, typical architectural cases are found and recommended to architects according to an analysis of the collected data. In the third stage, existing design patterns are discovered through typical cases, and new patterns are created for reasonable innovation. Figure 2 shows the process in detail.

3.1. Case collection

As discussed before, an architectural case consists of a series of characteristics. Therefore, the attributes that directly or indirectly reflect these characteristics are important and correlated with intercase similarity (Guo and Zou 2017). Values of these attributes are the basis for similarity calculation. However, there are massive attributes of architectural cases on the internet, and not every one of them is suitable for calculating intercase similarity. To be a similarity-related attribute, the attribute must be sufficient and have a certain correlation with intercase similarity. Values of similarity-related attributes are defined as the fundamental data in this study, and the purpose of this stage is to collect them from the internet.

According to the above standards, 20 architectural attributes were identified as similarity-related attributes based on the process shown in Figure 3. Specifically, we implemented the process in the following steps:

1. Based on a survey of 400 architectural cases on the internet, we found 185 attributes that could be included in the internet architectural case. After excluding 141 attributes that obviously did not reflect any architectural characteristics (e.g., project contractor) and 20 attributes that were included in fewer than 4 (1%) cases (e.g., landscape area), 24 remaining attributes (e.g., gross floor area) of building, environmental and designer information were set as candidate attributes for collection. In addition, 4 more attributes that can be inferred from the remaining attributes were added to the candidate attributes, such as building density. Table 1 shows these candidate attributes.

2. Forty-five participants were invited to judge the correlation between the candidate attributes and intercase similarity, including 14 architects, 13 architectural teachers and 18 graduate students.
of architecture (Table 2). As the participants often need to search architectural cases from the internet, the rich experience ensures that their judgment is reliable. The participants were requested to fill out a rating sheet, as shown in Appendix A1, which contains the above 28 candidate attributes and 3 categories to which they belong. If the participant determined an attribute or category of attributes was correlated with intercase similarity, he/she was asked to rate it within 1 to 5 score. If not, an “×” was marked. On the basis of their professional knowledge and experience, each attribute and category was rated. Table 1 shows the rating results. As no attribute received more than 9 (20%) “×”, all candidate attributes were considered to affect the intercase similarity to different degrees.

(3) The \( r_{wg} \) index proposed by James, Demaree, and Wolf (1984, 1993) was chosen as the measurement of interrater agreement because it is useful for evaluating the interrater agreement for judgments of a single target by one group of judges. The value of \( r_{wg} \) is calculated as follows:

\[
r_{wg} = 1 - \frac{S_x^2}{\sigma_{EU}^2}
\]

\[
\sigma_{EU}^2 = \frac{(A^2 - 1)}{12}
\]

where \( r_{wg} \) represents the interrater agreement for participants on a single attribute; \( S_x \) represents the sample variance on rated scores; \( \sigma_{EU}^2 \) represents the expected variance when there is a complete lack of agreement among participants; \( A \) represents the number of score options, which is 5 in our rating sheet. As shown in Table 1, participants received moderate and above agreements (\( r_{wg} > 0.5 \), LeBreton and Senter 2008) for the judgments of 16 *-attributes, which were identified as similarity-related attributes first. In stage 2, the mean scores of such attributes were used as their weights in intercase similarity calculation.

(4) For the other 12 attributes, another agreement test was implemented to provide them another opportunity to remain. As all three attribute categories’ \( r_{wg} \) indices are more than 0.5, these categories were used as the benchmarks for the attributes they contain in this test. For each remaining attribute, agreement between the rated scores and of its category was evaluated with a weighted kappa coefficient (quadratic). As shown in Table 1, the participants’ judgment on **-attribute and on its category showed moderate and above agreement (\( \kappa > 0.4 \), Landis and Koch 1977). Therefore, 4 **-attributes were also identified as similarity-related attributes, and the mean scores of their categories, not their mean scores, were used as their weights in intercase similarity calculation. For the other 8 attributes, because they failed in both agreement tests, their correlations with intercase similarity could not be confirmed. Therefore, we had to remove them from the candidate attributes. Finally, 20 attributes, including 16 *-attributes and 4 **-attributes, were identified as similarity-related attributes, which should be collected in the case collection stage. In addition, considering that “footprint” (\( r_{wg} < 0.5, \kappa < 0.4 \)) is necessary for computing “building density” (\( r_{wg} > 0.5 \)), it was also set as a collection target but not used in similarity calculation.

In this study, professional architectural websites (e.g. Archdaily.com) were set as the main source of data. In
addition, considering that the data included in the main source were usually incomplete, some relevant databases (e.g. World Weather Information Service) were considered supplementary sources of the missed data. By using a web spider, the data from professional architectural websites can be automatically collected. To achieve this goal, two kinds of information are required: (1) the URL of the architectural case and (2) the expression of attributes in the webpage source code. The web spider downloads corresponding data of architectural cases according to this information. The data from the relevant database are matched to the architectural cases according to some known data in the cases. For instance, if an architectural case has the data of location-city/county, the data of annual temperature is matched to it according to its location.

**Figure 3.** Process of similarity-related attribute identification. (Designed by authors).
Table 1. Results of rating and analysis of candidate attributes and categories.

| Category & attribute                | Mean          | 95%CI             | ×   | SD       | Subsimilarity | k   | p     | Weight |
|-----------------------------------|---------------|------------------|-----|----------|---------------|-----|-------|--------|
| Building information*             | 3.9           | 3.637–4.230      | 0   | 0.000   | 0.986         | 0.514 | 4.2   |        |
| Main use                          | 4.2           | 3.924–4.476      | 0   | 0.000   | 0.919         | 0.577 |       |
| Usage right                       | 3.6           | 3.288–3.957      | 0   | 0.000   | 1.114         | 0.380 | 0.371 | 0.009  |
| Gross floor area*                 | 2.9           | 2.624–3.229      | 4   | 0.899   | 0.959         | 0.540 | 2.9   |        |
| Aboveground area*                 | 2.7           | 2.364–3.002      | 4   | 0.899   | 1.011         | 0.489 | 0.200 | 0.015  |
| Space distribution*               | 3.7           | 3.492–3.935      | 0   | 0.000   | 0.821         | 0.663 | 3.7   |        |
| Footprint                         | 2.7           | 2.354–2.980      | 3   | 0.677   | 1.004         | 0.496 | 0.201 | 0.017  |
| Site area*                        | 2.6           | 2.320–2.930      | 5   | 1.114   | 0.952         | 0.546 | 2.6   |        |
| Building density (added)*         | 3.0           | 2.645–3.260      | 3   | 0.677   | 0.987         | 0.513 | 3.0   |        |
| Plot ratio*                       | 2.9           | 2.601–3.213      | 2   | 0.444   | 0.996         | 0.504 | 2.9   |        |
| Story*                            | 3.0           | 2.714–3.286      | 1   | 0.222   | 0.940         | 0.558 | 3.0   |        |
| Height*                           | 3.1           | 2.862–3.324      | 2   | 0.444   | 0.750         | 0.719 | 3.1   |        |
| Height category*                  | 3.1           | 2.830–3.392      | 0   | 0.000   | 0.935         | 0.563 | 3.1   |        |
| Structure*                        | 3.1           | 2.777–3.356      | 0   | 0.000   | 0.963         | 0.536 | 3.1   |        |
| Material*                         | 3.0           | 2.688–3.268      | 0   | 0.000   | 0.965         | 0.534 | 3.0   |        |
| Total cost                        | 2.3           | 1.930–2.607      | 4   | 0.899   | 1.073         | 0.424 | 0.062 | 0.331  |
| Cost/m² (added)                   | 2.3           | 1.962–2.596      | 2   | 0.444   | 1.031         | 0.468 | 0.145 | 0.026  |
| Keywords*                         | 3.1           | 2.792–3.390      | 1   | 0.222   | 0.984         | 0.516 | 3.1   |        |
| Design description                | 2.8           | 2.421–3.170      | 1   | 0.222   | 1.231         | 0.242 | −0.039| 0.673  |
| Year of design*                   | 2.3           | 2.008–2.550      | 2   | 0.444   | 0.882         | 0.611 | 2.3   |        |
| Year of completion*               | 2.1           | 1.834–2.405      | 3   | 0.677   | 0.916         | 0.580 | 2.1   |        |
| Environmental information*        | 3.2           | 2.900–3.411      | 0   | 0.000   | 0.852         | 0.637 |       |
| Location-Country**                | 2.9           | 2.513–3.214      | 1   | 0.222   | 1.153         | 0.335 | 0.544 | 0.000  |
| Location-Province/Region          | 2.5           | 2.190–2.880      | 2   | 0.444   | 1.120         | 0.373 | 0.356 | 0.002  |
| Location-City/County**           | 2.3           | 1.976–2.582      | 2   | 0.444   | 0.984         | 0.516 | 2.3   |        |
| Location-District/Town            | 2.1           | 1.754–2.496      | 5   | 0.677   | 1.159         | 0.379 | −0.031| 0.737  |
| Annual temperature (added)**      | 2.8           | 2.386–3.138      | 3   | 0.677   | 1.206         | 0.273 | 0.472 | 0.000  |
| Annual precipitation (added)**    | 2.6           | 2.194–2.934      | 6   | 1.333   | 1.142         | 0.348 | 0.498 | 0.000  |
| Designer information**            | 2.6           | 2.295–2.861      | 0   | 0.000   | 0.941         | 0.557 | 2.6   |        |
| Design company**                  | 2.6           | 2.248–2.895      | 3   | 0.677   | 1.039         | 0.460 | 0.801 | 0.000  |
| Architect**                       | 3.2           | 2.862–3.440      | 0   | 0.000   | 0.976         | 0.524 | 3.2   |        |

Note: 95%CI = 95% confidence interval for mean, × = number (%) of ×, SD = standard deviation, k = weighted kappa coefficient, p = p-value, * = Subsimilarity > 0.5, ** = Subsimilarity < 0.5 & k > 0.4.

Table 2. Professional characteristics of the participants (n = 45).

| Career & position | % (n) | Working, teaching or studying age |
|-------------------|-------|----------------------------------|
|                   | Mean  | Range   |
| Architect         | 30.4% (14) | 6.7 | 3-10 |
| Assistant engineer| 21.4% (3)  | 6.3 | 5-8  |
| Engineer          | 50.0% (7)  | 7.0 | 5-9  |
| Senior Engineer   | 14.3% (2)   | 8.5 | 7-10 |
| Other             | 14.3% (2)   | 8.5 | 7-10 |
| Teacher           | 28.3% (14)  | 12.8| 4-24 |
| Lecturer          | 30.8% (6)   | 7.5 | 4-9  |
| Associate professor| 46.2% (6)   | 13.0| 7-22 |
| Professor         | 23.1% (3)   | 19.3| 12-24|
| Graduate student  | 41.3% (18)  | 8.0| 5-13 |
| Master            | 66.7% (12)  | 6.6| 5-8  |
| Doctoral          | 33.3% (6)   | 10.8| 8-13 |
Sub(N) = \begin{cases} 
0.5, & P \text{ or } Q \text{ is null} \\
\frac{P}{Q}, & P \text{ and } Q \text{ are not null, } P \leq Q \\
\frac{Q}{P}, & P \text{ and } Q \text{ are not null, } P > Q 
\end{cases} (4)

where Sub(N) represents the subsimilarity on the numerical attribute, and P and Q represent the values of two cases on this attribute, respectively. For exact textual attributes that can only be “same” or “different” between two cases, including main use, height category, location-country, location-city/county, design company and architect, the subsimilarity is calculated as:

Sub(Te) = \begin{cases} 
0, & X \text{ or } Y \text{ is null} \\
1, & X \text{ and } Y \text{ are not null, } X \in Y \text{ or } X \supseteq Y 
\end{cases} (5)

where Sub(Te) represents the subsimilarity on the exact textual attribute, and X and Y represent the contents of two cases on this attribute. For fuzzy textual attributes that can be similar to varying degrees between two cases, including space distribution, structure, material and keywords, the Jaccard index (Renu and Mocko 2016) is used to calculate the subsimilarity as follows:

Sub(Tf) = \begin{cases} 
0.5, & A \text{ or } B \text{ is null} \\
\frac{A \cap B}{A \cup B}, & A \text{ and } B \text{ are not null} 
\end{cases} (6)

where Sub(Tf) represents the subsimilarity on the fuzzy textual attribute, and A and B represent the contents of two cases on this attribute.

### 3.2.3. Case clustering & evaluation

According to the intercase similarity of each two cases, a similarity matrix can be created and exported by the system, as shown in Figure 4. Based on this matrix, we developed a case clustering program to cluster architectural cases with an agglomerative hierarchical clustering algorithm (Algorithm 1, Tan et al. 2019).

As shown in Algorithm 1, the distance between pairwise clusters is vital to hierarchical clustering. At the very beginning of clustering, because each cluster has only one architectural case, the initial distance between any pairwise clusters equals the distance between their respective cases, which can be denoted as:

| C1 | C2 | C3 | C4 | ... | Cn |
|----|----|----|----|-----|-----|
| C1 | 1  | IS12 | IS13 | IS14 | ... | IS1n |
| C2 | IS12 | 1  | IS23 | IS24 | ... | IS2n |
| C3 | IS13 | IS23 | 1  | IS34 | ... | IS3n |
| C4 | IS14 | IS24 | IS34 | 1  | ... | IS4n |
| ... | ... | ... | ... | ... | 1  | ... |
| Cn | IS1n | IS2n | IS3n | IS4n | ... | 1  |

Note: C = Case, IS = Intercase similarity.

Figure 4. Diagram of the intercase similarity matrix. (Designed by authors).
Algorithm 1. Agglomerative hierarchical clustering

1. set each architectural case as individual cluster
2. repeat
3. merge the two nearest clusters into a new cluster
4. update the distances between the new cluster and others
5. until only one cluster remains

\[
\text{initial intercluster distance} = \text{intercase distance} = 1 - \text{intercase similarity} \quad (7)
\]

Then, after the first two clusters are merged, the Ward linkage (SciPy community 2019) is used to measure the intercluster distance. For a newly merged cluster \( u \) consisting of clusters \( s \) and \( t \), its distance from an existing cluster \( v \) is computed as:

\[
d(u,v) = \sqrt{\frac{|v| + |s|}{T}d(v,s)^2 + \frac{|v| + |t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}
\]

(8)

\[T = |v| + |s| + |t| \quad (9)\]

where \( d(u,v) \), \( d(v,s) \), \( d(v,t) \), and \( d(s,t) \), respectively, represent the intercluster distance between clusters \( u \) and \( v \), between clusters \( v \) and \( s \), between clusters \( v \) and \( t \), and between clusters \( s \) and \( t \); and \(|v|\), \(|s|\), and \(|t|\) represent the sizes of clusters \( v \), \( s \), and \( t \), respectively.

At the end of case clustering, the developed program prints a clustering tree. Architects can cut the tree and obtain several case groups on different clustering degrees, as shown in Figure 5. Normally, we suggest architects cut the tree between two combinations that have a large difference in clustering degree. This strategy is helpful for reaching the balance between the number of design patterns and the representative level of typical cases.

By importing the case groups back into the case mining system, each architectural case is evaluated on the basis of its average intracluster similarity, the average similarity with other cases in the same group. In each group, the case with the highest average similarity is defined as the central case. This definition implies that the central case can best reflect the design pattern of the group, as it is most similar to all other cases. Finally, the central cases of all case groups are considered the typical cases and recommended to the architect.

3.3. Case study

In this stage, the main existing design patterns are discovered from the typical cases first, and new patterns are created based on the existing patterns. Figure 6 shows the guide of pattern discovery and creation.

3.3.1. Pattern discovery

The existing design patterns are supposed to be discovered from the typical cases’ architectural factors, which can be grouped into three aspects (Zhuang 2000), as shown in Figure 6. It is necessary to note that the architectural factors do not equal the similarity-related attributes discussed before, although some parts overlap. The similarity-related attributes are essentially a set of abstract indices that reflect the design to some extent and are suitable for calculation. In this stage, architects should discover the design patterns of architectural factors directly through specific images and texts of the typical cases.

Because the typical cases were retrieved with the same criteria that were determined at the beginning of the design, they would be the same or very similar on related factors. For architects who need to innovate in
a project, more focus should be on those indeterminate factors. In other words, the design patterns should be discovered from the factors that were not involved in case retrieval. For example, if the floor area, main material and style of a project were determined and used for case retrieval in the case analysis stage, the architect should focus on the structure and layout of the typical cases and discover their main patterns.

In this way, architects can discover the design patterns of factors of concern on two levels. First, the design of one factor of typical cases can be regarded as a single-factor pattern. For example, red, yellow and blue can be regarded as three patterns of a color factor. Second, the combinations of design of several factors can be regarded as multi-factor patterns, such as brown wood facades and gray metal facades.

3.3.2. Pattern creation
Corresponding to the two levels of existing design patterns, there are two strategies to create new patterns as follows:

(1) Based on the single-factor pattern, architects can make a new design of the factor that was not used in typical cases. For instance, if the main materials of three typical cases are stone, glass and wood, the architect can mainly use metal or membrane in the current project.

(2) Based on the multi-factor pattern, architects can make a new combination of design that was not used in typical cases. For instance, if the forms of two typical cases are square plans with pitched roofs and round plans with flat roofs, the architect can design the building into round plans with pitched roofs or square plans with flat roofs in the current project.

The above strategies may help architects create many new patterns that were not used in the buildings within the same scope as the current project. With consideration of other factors and the actual situation of the project, one or some of the new patterns can be used for a reasonable innovation.

4. Example
To demonstrate the effect of the proposed method, a small museum project was assumed as a representative innovation requirement. In this project, the gross floor area (GFA) was approximately 2,000 m², and a new design pattern of exhibition space was demanded.
Therefore, the architect of this project needed to find the typical cases that represent the main design patterns of existing small museums first.

To find the typical cases, the architect should first collect a number of architectural cases from the internet. In this example, 32,716 architectural cases were collected from Archdaily.com by Locoy Spider, a common web spider program. In addition, the annual temperature and precipitation from the World Weather Information Service were matched to 11,930 cases according to their locations. These architectural cases were stored in the case mining system, containing 254,535 pieces of similarity-related attribute data.

By entering the criteria of "Main use: museum" and "GFA: 1,500–2,500 m²" into the developed system as shown in Figure 7, fifty-four qualified architectural cases were retrieved as shown in the upper section of Figure 8. Additionally, a matrix of the similarity between every two cases of the fifty-four cases was calculated and exported by the system, as shown in Figure 9. According to this matrix, the developed case clustering program hierarchically aggregated these cases into one cluster and output the clustering tree. Figure 10 shows the clustering tree, and five main clusters are marked in blue. After importing these clusters back into the system as shown in the lower section of Figure 8, the system ranked the cases of each cluster according to their average intracluster similarity (Table 3) and recommended the first case of each cluster as a typical case (Figure 11). These typical cases are Renewal of Stedelijk Museum Hof van Busleyden (Case 32), Easter Sculpture Museum (Case 29), Design Wing (Case 37), New Maritime Museum and Exploratorium (Case 28) and Museum Tonofenfabrik Lahr (Case 19). Figures 12–16 show the floor plans of the main exhibition spaces.

According to the observation of the exhibition space of the typical cases, the architect can find that two factors, position and integrity, are related to the design pattern of the exhibition space. As shown in the third row of Figure 17, there are four single-factor patterns of position, including separate (Case 28), beside auxiliary space (Cases 19 & 29), crossed with auxiliary space (Case 32) and surrounded by auxiliary space (Case 37), and two single-factor patterns of integrity, including integral (Cases 28 & 29) and scattered (Cases 19, 32 & 37), are used in the typical cases. Furthermore, five multi-factor patterns combined by the above two factors were discovered as follows: (1) integral exhibition space without auxiliary space, represented by Case 28; (2) integral exhibition space beside auxiliary space, represented by Case 29; (3) scattered exhibition space beside auxiliary space, represented by Case 32; (4) scattered exhibition space crossed with auxiliary space, represented by Case 32; and (5) scattered exhibition space surrounded by auxiliary space, represented by Case 37. The fourth row of Figure 17 shows the multi-factor patterns.

Based on the existing patterns of position of exhibition space, the architect can create a new single-factor pattern, exhibition space surrounding auxiliary space, as shown in the third row of Figure 17. In addition, four new multi-factor...
Figure 8. Analysis interface. (Screenshot of the developed system).

Figure 9. The similarity matrix of fifty-four small museums (Designed by authors).

|      | C_1  | C_2  | C_3  | C_4  | C_5  | ... | C_54 |
|------|------|------|------|------|------|-----|------|
| C_1  | 1    | 0.392| 0.422| 0.428| 0.538| ... | 0.377|
| C_2  | 0.392| 1    | 0.455| 0.490| 0.468| ... | 0.383|
| C_3  | 0.422| 0.455| 1    | 0.483| 0.504| ... | 0.372|
| C_4  | 0.428| 0.490| 0.483| 1    | 0.515| ... | 0.373|
| C_5  | 0.538| 0.468| 0.504| 0.515| 1    |     | 0.376|
| ...  | ...  | ...  | ...  | ...  | 1    | ... | ...  |
| C_{54}| 0.377| 0.383| 0.372| 0.373| 0.376| ... | 1    |

Note: C = Case
Figure 10. The clustering tree of fifty-four small museums (Output by the developed program, colored by authors).

Table 3. The average intracluster similarity ranking of five case clusters.

| Rank | Cluster 1 ID | AIS   | Cluster 2 ID | AIS   | Cluster 3 ID | AIS   | Cluster 4 ID | AIS   | Cluster 5 ID | AIS   |
|------|--------------|-------|--------------|-------|--------------|-------|--------------|-------|--------------|-------|
| 1    | 32           | 0.475 | 29           | 0.409 | 37           | 0.541 | 28           | 0.472 | 19           | 0.460 |
| 2    | 34           | 0.473 | 49           | 0.400 | 42           | 0.538 | 22           | 0.472 | 5            | 0.460 |
| 3    | 33           | 0.472 | 11           | 0.400 | 38           | 0.534 | 17           | 0.470 | 8            | 0.444 |
| 4    | 10           | 0.399 | 3            | 0.531 | 41           | 0.469 | 1            | 0.435 |             |       |
| 5    | 36           | 0.398 | 39           | 0.530 | 2            | 0.469 | 30           | 0.424 |             |       |

Note: ID = case ID, AIS = average intracluster similarity

Figure 11. Results interface. (Screenshot of the developed system).
patterns, (1) scattered exhibition space without auxiliary space, (2) integral exhibition space surrounded by auxiliary space, (3) integral exhibition space surrounding auxiliary space and (4) scattered exhibition space surrounding auxiliary space, can be created by making new combinations of single-factor patterns, as shown in the fourth row of Figure 17.
5. Discussion

Our example suggests that finding typical architectural cases by clustering is a feasible approach for architects to find the representatives of the main design patterns of existing buildings. Faced with a large number of architectural cases, the proposed method may help architects to have a general overview of existing buildings and make reasonable innovations within an acceptable time.

Using cluster analysis to search cases is not a new idea. Cheng et al. (1997) proposed a similar idea more than two decades ago. However, this study used clustering differently. In prior relevant studies, researchers have preferred to use clustering to assist in improving the efficiency of case retrieval. In contrast, the point of
this study is to use clustering to find the typical case. Therefore, the case mining process established in this study is the opposite of the prior studies. For instance, in Yang and Wu’s (2001) study and other similar studies, clustering is executed before retrieval to narrow the case scope. In this study, case retrieval is executed first to narrow the case scope for clustering. This makes it possible to find unexpected cases without a clear target. In terms of the authors’ opinions, intercase similarity should be calculated on the basis of architectural attributes. Alternatively, Wei (2013) advocated calculating similarity according to the topological relation of architectural space. This difference may occur because the data used in our and his studies are from the internet and BIM, respectively.

Hierarchical clustering is used to cluster architectural cases by authors because it does not require users to specify the number of clusters in advance. This feature is appropriate for architects who do not know how many design patterns are actually there. Li et al. (2014) made the same choice as ours in their work. Furthermore, because many architectural attributes are not numerical, the intercase similarity must be calculated with customized rules. As Hastie, Tibshirani, and Friedman (2009) suggested, hierarchical clustering allows users to flexibly define the intercase similarity (or dissimilarity). This advantage is another reason why we use it to cluster architectural cases in this study.

This study has several limitations. The data of some potential similarity-related attributes may be missed because the data were only collected from architectural websites. The missing data might affect the similarity calculation and change the result of case mining. Although the collected data cover the main aspects of architectural cases, including building, environment and designer, the possible negative effects from missing data cannot be ruled out. A comprehensive cross-sectional study is required to identify every variable that has a significant association with intercase similarity.

Another limitation is that most collected cases do not include the data of all similarity-related attributes. Due to the lack of data, some qualified cases would be incorrectly excluded during the case retrieval phase. For some design patterns, the excluded cases might be more typical than the retrieved cases. In the future, the lacking data must be supplemented from other sources, such as government public databases and open access cloud platforms.

6. Conclusion

Architectural innovation is gaining increasing attention from architects. To reasonably innovate in new projects, architects must discover the main design patterns of existing buildings by case study first. As there
are too many architectural cases in the big data era, studying the typical cases of these patterns is considered an efficient approach to reasonable innovation.

Focusing on intercase similarity, this study proposes a clustering-based case mining method to find the typical cases from a great number of architectural cases on the internet, discover the main existing design patterns from the typical cases, and create new design patterns based on the existing patterns. For this purpose, twenty similarity-related attributes of the architectural case were identified first. Based on these attributes, an extensive architectural design case mining system was developed to calculate and compare the similarities between architectural cases. In addition, a program of agglomerative hierarchical case clustering was developed to cluster the cases according to their intercase similarities. Finally, a guide of the typical case study was presented to discover and create design patterns. The practical application effect of the method was demonstrated by an assumed innovation requirement in a small museum.

The main contributions of this study can be summarized as follows.

First, with this method, architects can efficiently find typical architectural cases and discover the main design patterns of existing buildings, thereby seeking a reasonable approach to innovate in architectural design.

Second, since typical cases can be found from cases that have not been previously viewed, architects’ dependence on experience is reduced by this method. For architects who need to design unfamiliar buildings, this feature is helpful.

Third, this study provides an intelligent tool for academic research on architecture from a big data perspective. In addition to innovation, the proposed method can also be used in typological analyses of architectural cases in different regions and periods.

Fourth, the proposed method presents a generalized mechanism for typical case mining. This mechanism can be flexibly used in other professions. Engineers, planners, and other designers are encouraged to make reasonable innovation in their fields with the same idea.

In the big data era, with the rapid increase in data on the internet, application of data mining technology on architectural and relevant data is a promising method for promoting intelligent support for architectural design. In addition to typical cases and design patterns, further data-related studies will be performed under
the EADDM framework to discover more kinds of knowledge for architects. Future efforts to develop a system that integrates architects’ requirements, internet sources and analysis programs are also expected.

Acknowledgments

The authors are grateful for the data of architectural cases from Archdaily.com and the World Weather Information Service. Their selfless sharing helped to achieve the objective of this study. The authors are also appreciative of the open source codes from Scipy, Pandas, and Matplotlib. Our case clustering program was developed with these excellent codes.

Data Availability Statement

The data that support the identification of similarity-related attributes in Section 3.1 are openly available in Figshare at https://figshare.com/articles/Evaluation_of_Architectural_Attributes_Weight_in_Inter-case_Similarity_Calculation/8216678 [10.6084/m9.figshare.8216678] (Liu 2019). The data that support the example in Section 4 are from two sources. Most of the data were collected from Archdaily.com, and the copyright belongs to the website. Restrictions apply to the availability of the data, which were collected and used under official permission for this study. The supplementary data on annual temperature and precipitation were derived from the World Weather Information Service (WWIS), which is a resource available in the public domain. According to its permission and requirement at http://worldweather.wmo.int/en/pilot.html (accessed on September 15th, 2019), the data are available in Figshare at https://figshare.com/articles/Weather_Data_of_308_Cities_Worldwide_from_WWIS_Accessed_on_2019_9841391 [10.6084/m9.figshare.9841391], and the national institutes that provided these data are credited in the dataset.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [51178132].

Notes on contributors

Shuyu Liu gained his BS degree of Architecture and MS degree of Engineering from Nanjing Tech University, Nanjing, Jiangsu, China. He is currently a PhD candidate at School of Architecture, Harbin Institute of Technology, Harbin, Heilongjiang, China. His major study field includes architectural data mining and affordable housing design.

Guangtian Zou gained his BS and MS degrees of Engineering from Harbin Institute of Civil Engineering and Architecture, Harbin, Heilongjiang, China and gained his PhD degree of Technical Science from Tokyo University, Tokyo, Japan. He is currently a professor and PhD supervisor at School of Architecture, Harbin Institute of Technology (HIT), Harbin, Heilongjiang, China. He also serves as the director of Architectural Planning and Design Institute (APDI), HIT, vice director of Environment and Behavior Research Association (EBRA), deputy director of Architectural Programming Association (APA) of Architectural Society of China (ASC), council member of urban design branch of ASC, deputy director of Extenics Association of the Chinese Association for Artificial Intelligence (CAAI), etc. His major study field includes environmental behavior psychology, architectural programming, architectural design innovation, extension architecture, architectural data mining, elderly facilities and polar architectural design. Prof. Zou has hosted and completed 2 projects supported by National Natural Science Foundation of China, has published a monograph on architecture which is "Architectural Planning", has compiled the part of Welfare Facilities of "Architectural Design Data Set (3rd Edition)" and "Development Report 2016 of Extenics, Chinese Original Discipline" as deputy editor, has edited 2 journal special issues as guest editor and 4 proceedings of 2 national and 2 international conferences.

Si Zhang gained her BS and MS degrees of Architecture from Southwest Jiaotong University, Chengdu, Sichuan, China and Harbin Institute of Technology (HIT), Harbin, Heilongjiang, China, respectively. She is currently a PhD candidate at School of Architecture, HIT and was a visiting scholar at Center for Health Systems & Design, Texas A&M University, College Station, TX, United States. Her major study field includes architectural data mining, architectural programming and inclusive built environment.

ORCID

Shuyu Liu http://orcid.org/0000-0003-1336-1094
Si Zhang http://orcid.org/0000-0002-2539-1138

References

Cha, H., and D. Lee. 2018. “Framework Based on Building Information Modelling for Information Management by Linking Construction Documents to Design Objects.” Journal of Asian Architecture and Building Engineering 17 (2): 329–336. doi:10.3130/jaabe.17.329.

Chang, J. W., M. C. Lee, and T. I. Wang. 2016. “Integrating A Semantic-based Retrieval Agent into Case-based Reasoning Systems: A Case Study of an Online Bookstore.” Computers in Industry 78: 29–42. doi:10.1016/j.compind.2015.10.007.

Chen, Q., Y. Xiang, X. Guo, and D. Wang. 2010. “Rough-set-based K-means Clustering Algorithm in case Retrieval.” Computer Science [in Chinese] 37 (12): 161–164. doi:10.3969/j.issn.1002-137X.2010.12.035.

Cheng, C. H., J. Motwani, A. Kumar, and J. Jiang. 1997. “Clustering Cases for Case-based Reasoning Systems.” Journal of Computer Information Systems 38 (1): 30–37. doi:10.1080/08874417.1997.11647305.

Chiu, C.-C., and C.-Y. Tsai. 2007. “A Weighted Feature C-Means Clustering Algorithm for Case Indexing and Retrieval in Cased-Based Reasoning.” In New Trends in Applied Artificial Intelligence, edited by H. G. Okuno and M. Ali, 541–551. Berlin, Berlin Heidelberg: Springer-Verlag. doi:10.1007/978-3-540-73325-6_54.

Domeshek, E., and J. Kolodner. 1993. “Using the Points of Large Cases.” AI Edam 7 (2): 87–96. doi:10.1017/S089006400000780.

Furundzic, D. S. 2003. “Website Applications in Urbanism and Architecture.” Spatium 9: 34–41. doi:10.2298/SPAT0309034F.
Attribute weights for similarity calculation between architectural cases

Hi, we are studying on the similarity between architectural cases. In order to achieve this goal, the weights of architectural attributes for similarity calculation need to be rated first. We look forward to your professional advice.

Please fill in the form according to the following:

1. Give weights to the attributes shown in the first box. If you think an attribute is correlated with the similarity between architectural cases, please mark it in the range of 1-5. The meaning of each score is shown in the diagram below. If not, please place an X by it.
2. Give weights to the categories of attribute according to the same rule as before.
3. Select your career and position shown in the second box, and fill in your working, teaching or studying age behind them. We guarantee that your professional information will only be used for academic statistics and research.

Thank you for your contribution!

### Give weights to attributes

| Building information: | Environmental information: | Designer information: |
|-----------------------|-----------------------------|----------------------|
| Main use | Usage right (public/private) | Gross floor area | Aboveground area |
| Space distribution | Site area | Footprint (ground floor area) | Story |
| Building density | Structure | Total cost | Cost/m² |
| Height (how many meters) | Material | | |
| Height category (e.g., high-rise) | | | |
| Keywords | Design description | Year of design | Year of completion |
| Location-Country | Location-Prov/Region | Location-City/County | Location-Dist/Town |
| Annual temperature | | | |
| | | | |
| Design company | Architect | |

### Participant characteristics

| Career (select in options) | Position (select in options) | Working age: |
|-----------------------------|-------------------------------|--------------|
| Architect | Assistant engineer | | |
| Teacher | Engineer | | |
| Graduate student | Senior engineer | | |
| | Professor of Engineering | | |
| | Assistant | | |
| | Lecturer | | |
| | Associate professor | | |
| | Professor | | |
| | Master | | |
| | Doctoral | | |

Teaching age: 
Studying age: (from undergraduate stage)

No. Date: / / (YYYY) (MM) (DD)

**Appendix A1.** The rating sheet used for attribute evaluation. (Designed by authors).