RESEARCH ARTICLE

Ensemble learning-based hierarchical retrieval of similar cases for site planning

Pengyue Wang1,2, Maozu Guo2,3,*, Yunsong Han4,*, Lingling Zhao5, Xiaoping Zhou2,3 and Dayu Zhang1

1School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; 2Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing 100044, China; 3School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; 4School of Architecture, Harbin Institute of Technology, Harbin 150001, China and 5School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

*Corresponding author. E-mail: guomaozu@bucea.edu.cn; hanyunsong@hit.edu.cn

Abstract

Site planning not only involves the arrangement of a large number of elements but also considers the complexity of urban systems; hence, the design process involves large workloads and is time consuming. Retrieving similar existing cases helps architects optimize or accelerate the design process. This paper proposes a computational approach that provides valuable references by retrieving similar cases. Three types of attributes are extracted to represent a given parcel: land-use attributes, geometric attributes (area, orientation, etc.), and environmental attributes (greening rate, surrounding facilities, etc.). The complete hierarchical retrieval process is divided into three phases. The first phase selects cases whose land-use attributes are consistent with the target parcel. Then, the similarity distances between the given target parcel and the selected cases are calculated using geometric attributes. The eXtreme Gradient Boosting (XGBoost) classifier is adopted to determine which case is similar to the target parcel. Finally, similarity scores of the retrieved cases are calculated based on the environmental attributes to provide more options during the actual design. In total, 1189 cases with different land-use types in Beijing were collected for the case base. The comparative experimental results confirmed that the proposed ensemble learning-based hierarchical retrieval of similar cases approach improves the accuracy of retrieval results. Furthermore, we use a real-world target parcel to demonstrate the superiority and flexibility of the retrieval process.

Keywords: site planning; similarity retrieval; parcel representation; XGBoost algorithm

1. Introduction

Site planning is an organizational design process that involves the arrangement of buildings, landscape elements, roadways, and vegetation (Beer & Higgins, 2004). Commonly, the design process is complex and time consuming (Sun et al., 2016). Design tasks require not only a general understanding of daily human issues and the surrounding environment but also design expertise gained through extensive design practice (Domeshek & Kolodner, 1992). Observations of human designers frequently reuse the previous design cases as a source of design knowledge (Hua & Faltings, 1993). Existing cases generally integrate the architectural knowledge, experience, intelligence, and creativity of architects (So’mnez, 2018). Therefore, previous cases could provide architects with valuable architectural design assistance.
Architects can solve their problems by referencing previous cases to learn from and adapt them for new problems (Zehtabian et al., 2016), thereby reducing the design time. Architectural case retrieval techniques can retrieve similar previous cases according to the requirements of architects. Similar cases potentially allow architects to capture the knowledge and experience of architectural design easily and thus complete designs more quickly.

The architectural case retrieval technique belongs to case-based design (CBD) and assumes that the experience and knowledge of architects reside in design examples (Lawson, 2004). One type of CBD is based on prototypes, which are common design elements extracted from similar design cases (Gero et al., 1987). The prototypes can be used repeatedly as design experience resources. However, it is inevitable that the uniqueness of the case itself can also be filtered out during the extraction process. Another type of CBD skips the extraction of knowledge and matches the closest cases to a new problem. This approach needs features to describe the characteristics of cases and problems. These features are used to measure the similarity of cases and problems (Rosenman et al., 1991; Watson & Perera, 1997). However, the related research only uses simple features (area, orientation, etc.) without considering the surrounding environment, which is not sufficient to adequately characterize the conditions and constraints involved in the case. Moreover, state-of-the-art retrieval methods usually use a fixed threshold or a weighted average that is determined based on experience. These similarity calculation approaches are too simple to obtain accurate retrieval results.

To address these problems, this paper proposes a computational approach to retrieve cases similar to the target parcel. Here, similar cases/parcels refer to cases/parcels with similar conditions and design constraints. Note that this approach does not generate or find solutions. Instead, it attempts to provide architects with multiple references. Obtaining an accurate feature representation of the conditions and design constraints of previous cases and target parcels is a fundamental prerequisite for effective retrieval. In this study, these aspects are represented by three types of attributes: land-use attributes, geometric attributes (area, orientation, etc.), and environmental attributes (greening rate, surrounding facilities, etc.). We apply a hierarchical retrieval method composed of three phases: selecting cases with the same land-use type as the target parcel, retrieving similar parcels from the case base using the extreme Gradient Boosting (XGBoost) classifier, and evaluating the retrieved cases according to specific requirements.

The main contributions of this paper are as follows:

1. We present feature representation based on various architectural knowledge about parcel conditions and design constraints. These features consist of three types of attributes related to architects' concerns regarding the actual design.
2. We propose an ensemble learning-based classifier to improve the accuracy of the retrieval results. This model is a supervised learning method and can learn directly from labeled data containing the experience of experts. In addition, the model outputs important feature ranking, which transforms subjective evaluations into quantifiable indicators.
3. We evaluate the retrieved cases based on environmental attributes to meet different architectural needs. The evaluation of the retrieved cases makes it easier for architects to find, browse, and learn from the cases, which can provide valuable references.

The remainder of this paper is structured as follows: Section 2 reviews the related works on CBD. Section 3 provides an overview of the proposed approach and describes each part in detail. Section 4 presents the experimental datasets, the retrieval approach based on the XGBoost classifier, a case study, and a discussion of the experimental results. Section 5 concludes this work and suggests directions for future work.

2. Related Works

CBD is a general paradigm for solving problems based on the recall and reuse of specific experiences from previous cases (Maier & de Silva Garza, 1997). One type of CBD depends on prototypes. The prototypes are extracted from existing cases and defined by feature vectors. These prototypes can then act as reusable design elements for future designs. Hua et al. (1996) implemented a case-based building design system through dimensionality reduction (CADRE). Luca et al. (2007) developed a generative modeling platform that applied prototype designs to dynamic simulations. Okeil (2019) compared sketches, prototypes, and virtual reality to create new hybrid design environments. Li and Han (2011) controlled the building prototype using three parameters, the rectangle, the area, and the aspect ratio, developing a usable package. Ling et al. (2013) arranged buildings based on the common residential styles of determinant arrangement and peripheral arrangement. Sun et al. (2017) developed a hierarchical and parametric prototype system to automatically generate a 3D city model. These studies show the different ways of using prototypes in CBD. However, a prototype extracts common knowledge from similar cases, which suggests that the specific architectural knowledge is filtered during this process. Specific architectural knowledge refers to detailed and unique knowledge in a given case. Common knowledge is not enough for the task of designing complex parcels.

Another type of CBD directly uses the existing cases that are closest to the problem as knowledge. Cases are distinguished from other forms of knowledge in that they model specific, non-generalized instances (Hua & Faltings, 1993). How to represent the characteristics of cases and retrieve valuable previous cases are key issues.

The characteristics of the design constraints and case conditions are rich and multilayered. Zhang and Lu (2004) identified that the Fourier descriptor (FD) and Hu invariant moments are promising methods in contour-based shape representation techniques. Fang et al. (2012) used the circular compactness, the area, the perimeter and the orientation to describe parcels. Sun et al. (2016) described cases using the orientation, the floor area ratio, and other parcel information. Fleischmann (2019) implemented morphometric characters to analyse urban forms that spanned from the individual building level to entire urban regions. Morgenstern et al. (2019) developed a method to estimate perceived shape similarity of images by integrating over 100 shape features. These methods demonstrate different ways to describe cases from various perspectives. However, these approaches consider either only the shape aspect or a simple description. The surrounding environment of the case is not considered. Due to the distinct feature representation, some feature fusion methods can be utilized to capture more accurate and useful information (Zhang et al., 2020; Wang et al., 2021).

After obtaining suitable features, retrieval methods use these features to measure the similarity of previous cases and the target parcel and retrieve relevant similar cases. Traditionally, similar parcels are retrieved by using the weighted averages of various features or setting thresholds. The weighted values and thresholds are set according to expert experience. These approaches are subjective and have limitations, which lead to inaccurate retrieval results (Hao et al., 2008; Zheng & Zhang, 2013).
Recent architectural studies on machine learning (ML) appear to provide potential approaches. ML techniques can learn from data and self-optimize to obtain better accuracy. The XGBoost method was first introduced by Chen and Guestrin (2016). XGBoost has shown promising results regarding stability, accuracy, and efficiency compared to other ML methods, such as logistic regression (LR; Hernesniemi et al., 2019), support vector machine (SVM; Li et al., 2020), K-nearest neighbor (KNN; Naghibi et al., 2019), and random forest (RF; Speiser et al., 2019) algorithms. XGBoost provides a parallel tree boosting approach that can solve classification problems quickly and accurately. Moreover, XGBoost uses less computing resources than other methods while producing superior results in less time.

3. Methodology

This section provides an overview of the proposed approach and several essential aspects, including parcel representation, the XGBoost classifier, and hierarchical retrieval.

3.1. Overview of the proposed approach

The proposed approach is divided into two main phases: a training phase and a retrieval phase. Feature representation is crucial in both phases. The dataset used during the training phase is composed of pairwise parcels labeled by experts. After geometric representation, we obtain a similarity distance vector for each pairwise parcel. Then, these distance vectors are used to train the XGBoost classifier. The parameters of the optimal XGBoost classifier are saved and applied during the retrieval phase. The retrieval phase retrieves cases similar to a target parcel based on three types of attributes. First, we retrieve potential cases that have the same land-use type as the target parcel. Then, we calculate the similarity distance vector between the target parcel and each potential case. The distance vector is input into the trained XGBoost classifier, and the output 1 indicates that the pairwise parcel is similar, while the output 0 indicates that the pairwise parcel is dissimilar. Then, cases with an output of 1 are retrieved as references. Finally, these retrieved cases are evaluated by different aspects based on environmental attributes to assist the architect in choosing cases for reference or modification. Figure 1 depicts the complete process.

3.2. Parcel feature representation

According to experts’ concerns regarding parcel design, three types of attributes are considered to characterize parcels: land-use attributes (the state-of-the-art), geometric attributes, and environmental attributes, as listed in Table 1. The most common land-use attributes are residential, commercial, and industrial. The geometric attributes are separated into three categories: basic descriptors, simple descriptors, and complicated descriptors. The geometric attributes are used to calculate the similarity vector between each case in the case base and the target parcel. The environmental attributes include parcel information, surrounding facility information, and the surrounding natural
| Attributes                          | Categories                | Characteristics                      | Additional comments                      | Distance calculation method |
|------------------------------------|---------------------------|--------------------------------------|------------------------------------------|----------------------------|
| Land-use attributes                | Type                      | Land-use type                         | Residential, commercial, and industrial  | \                           |
| Geometric attributes               | Basic descriptors         | Area                                  | Area of each parcel                      | d_{AB} = | abs(A - B) |
|                                   |                           | Perimeter                             | Perimeter of each parcel                 |                            |
|                                   |                           | Longest axis                          | Length of the longest axis of each parcel|                            |
|                                   |                           | Orientation                           | From Yan and Wang (2009)                 |                            |
|                                   |                           | Corners                               | Number of parcel corners                 |                            |
|                                   |                           | Compactness                           | from Jacob et al. (2017)                 |                            |
|                                   |                           | Convexity                             | (minimum bounding rotated rectangle area) from Jacob et al. (2017) |                            |
|                                   |                           | Rectangularity                         | (minimum bounding rotated rectangle area) from Jacob et al. (2017) |                            |
|                                   |                           | Square compactness                    | \left\{ \frac{4\pi\text{area}}{\text{perimeter}} \right\} from Alessandra (2018) |                            |
|                                   |                           | Compactness weight axis               | longest axis \times \left\{ \frac{4\pi\text{area}}{\text{perimeter}} \right\} from Fleischmann (2019) |                            |
|                                   | Simple descriptors        | Elongation                            | \left\{ \frac{\text{perimeter} - \sqrt{\text{perimeter}^2 - 16\text{area}}}{4} \right\} from Gil et al. (2012) |                            |
|                                   |                           | Equivalent rectangular index          | \left\{ \frac{\text{area of bounding rectangle}}{\text{perimeter of bounding rectangle}} \right\} from Basaraner and Cetinkaya (2017) |                            |
|                                   |                           | Fractal dimension                     | 2\log_{4} \left( \frac{\text{perimeter}}{\text{area}} \right) from McGarigal (1995) |                            |
|                                   | Complicated descriptors   | FDs                                   | From Rafael and Richard (2011)           |                            |
|                                   |                           | Hu moments                            | From Hu (1962)                           |                            |
| Environmental attributes           | Parcel information        | Floor area ratio (FAR)                | The maximum floor space that can be constructed on a parcel |                            |
|                                   |                           | Entrances                             | Number of entrances in the parcel        |                            |
|                                   |                           | Greening rate                         | Green spaces inside the parcel           |                            |
|                                   |                           | Parking space                         | Parking space inside of the parcel       |                            |
|                                   |                           | Build date                            | Date of building completion              |                            |
|                                   |                           | Longitude                             | Geographic coordinates                   |                            |
|                                   |                           | Latitude                              |                                        |                            |
|                                   |                           | Mall                                  |                                        |                            |
|                                   |                           | Restaurant                            |                                        |                            |
|                                   |                           | Hospital                              |                                        |                            |
|                                   |                           | Subway                                |                                        |                            |
|                                   |                           | Sightseeing spot                      |                                        |                            |
|                                   | Surrounding facilities (1 km) | Park                                  | Descriptions of facilities within 1 km of the parcel |                            |
|                                   |                           | Water                                 | Descriptions of natural resources such as water resources and parks within 1 km of the parcel |                            |
|                                   | Surrounding natural environment (1 km) | Park                                  | Descriptions of natural resources such as water resources and parks within 1 km of the parcel |                            |
environment. The environmental attributes are used to calculate similarity scores between the retrieved cases and target parcel with respect to different aspects. Then, the architect can select cases that meet the actual design requirements based on the similarity scores.

3.2.1. Land-use attributes
Land use is the characterization of the parcel’s intended use. It determines what types of structures can be built on a parcel and what that parcel can be used for. Land use also describes the type of community, environment, or settlement for which a particular type of parcel can be used. The three most common land uses are residential, commercial, and industrial. The site planning process begins by determining the land-use type of the parcel.

3.2.2. Geometric attributes
Considering the calculation methods and information contained in geometric attributes, all the geometric attributes are divided into three categories: basic descriptors, simple descriptors, and complicated descriptors. The calculation methods of simple descriptors are listed in Table 1. Each basic and simple descriptor can only discriminate shapes with large differences and is not suitable for standalone use. Therefore, basic and simple descriptors usually need to be used in combination.

An FD is a way of representing the shape of a closed curve at varying levels of detail. As shown in equation (1), the contour points of the parcel \((x(k), y(k))\), \(k = 0, 1, \ldots, K - 1\) can be represented by the one-dimensional complex numbers \(s(k)\). A one-dimensional Fourier transform is then applied to \(s(k)\) to obtain FDs \(a(u)\), as shown in equation (2):

\[
s(k) = x(k) + j y(k) , \quad k = 0, 1, \ldots, K - 1 . \tag{1}
\]

where \(k\) is the \(k\)th contour point of the parcel and \(K\) indicates the total contour point number of the parcel. Then, the FDs are calculated as follows:

\[
a(u) = \sum_{k=0}^{K-1} s(k) e^{-j2\pi u k /K} , \quad u = 0, 1, \ldots, K - 1 . \tag{2}
\]

The normalized FDs \(f_n\) are computed as shown in equation (3). The high-frequency components of the FD are susceptible to noise. The normalization coefficient is already close to 0 when \(n = 20\). Therefore, we utilize \(n \leq 20\) to represent the parcel shape, and these values correspond to the low-frequency components of the boundary.

\[
f_n = \left| \frac{a(n)}{a(1)} \right| , \quad n = 1, 2, \ldots, K - 1 . \tag{3}
\]

Hu moments perform well on similarity transformed and affinely transformed contour-based shapes. The raw moment, \(M_{pq}\), of the order \((p+q)\)th moments is defined as shown in equation (4).

\[
M_{pq} = \sum_{x} \sum_{y} x^p y^q I(x, y) , \quad p, q = 0, 1, 2, \ldots \tag{4}
\]

where \(p\) and \(q\) are nonnegative integers, \(x\) and \(y\) are the positions of the contour points of the parcel, and \(I(x, y) = 1\) for contours on the object.

The central moments \(\mu_{pq}\) are defined in equations (5) and (6) as follows:

\[
\begin{align*}
\hat{x} &= \frac{M_{10}}{M_{00}}, & \hat{y} &= \frac{M_{01}}{M_{00}}, \\
\mu_{pq} &= \sum_{x} \sum_{y} (x - \hat{x})^p (y - \hat{y})^q I(x, y) . \quad p, q = 0, 1, 2, \ldots \tag{5}
\end{align*}
\]

The normalized central moments are defined in equation (7). A set of seven Hu moments is calculated using the central moments, as shown in equations (8)–(14). The first six moments have been proven to be invariant to translation, scale, rotation, and reflection, but the sign of the seventh moment changes for reflection.

\[
\begin{align*}
\eta_{pq} &= \mu_{pq} / (\mu_{00}^{1/2}) , & \rho &= \left( \frac{p + q}{2} \right) + 1 , \tag{7}
\end{align*}
\]

\[
\begin{align*}
h_1 &= \eta_{20} + \eta_{02} , & h_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 , \tag{8}
\end{align*}
\]

\[
\begin{align*}
h_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 , & h_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 , \tag{9}
\end{align*}
\]

\[
\begin{align*}
h_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \\
&+ (3\eta_{21} - \eta_{03})(\eta_{03} + \eta_{12}) \tag{10}
\end{align*}
\]

\[
\begin{align*}
h_6 &= (\eta_{20} - \eta_{02})((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2) \\
&+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{12} + \eta_{03}) , \tag{11}
\end{align*}
\]

\[
\begin{align*}
h_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2) \\
&- (\eta_{30} - 3\eta_{12})(\eta_{03} + \eta_{12}) \tag{12}
\end{align*}
\]

\[
\begin{align*}
&\times (3\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 . \tag{13}
\end{align*}
\]

Based on these parcel representations, the similarity between these parcel descriptors is calculated according to the following equations. The similarity measurements for basic and simple descriptors are calculated as shown in equation (15), where \(A\) denotes each descriptor of parcel \(A\) and \(B\) denotes each descriptor of parcel \(B\):

\[
d_{AB} = \left| a - b \right| . \tag{15}
\]

We use the Euclidean distance to calculate the similarity between the two FD vectors as defined in equation (16):

\[
d_{FD} = \left( \sum_{n=1}^{20} \left| f_n^A - f_n^B \right|^2 \right)^{1/2} . \tag{16}
\]

The similarity distance between the Hu invariant moments is calculated according to equations (17) and (18) as follows:

\[
m_i = \text{sign}(h_i) \cdot \log_{10} h_i , \quad i = 1, \ldots, 7 , \tag{17}
\]

where \(h_i\) is the Hu moment of the parcel.

\[
d_{Hu} = \sum_{i=1}^{7} \left| m_i^A - m_i^B \right| . \tag{18}
\]

After measuring the similarity distance, these distance values are concatenated into a 1×18 distance vector and used to train the XGBoost model. It would be useful to capture the ways which an architect observes and compares the similarity between pairwise parcels by integrating these multiple characters into a multidimensional representation.

3.2.3. Environmental attributes
The environmental attributes include three aspects: parcel information, the surrounding facilities, and the surrounding natural environment. Parcel information includes essential
information about the parcel. The surrounding facilities and natural environment describe various unique and specific attributes in the surrounding environmental. These environmental attributes are used to calculate the similarity scores for different aspects to assist architects in selecting cases to refer to or modify.

The parcel information similarity distance is calculated by equation (19):

\[ S_{\text{parcel}} = 1 - \frac{|C - G|}{G}, \]  

(19)

where \(C\) refers to the character of retrieved case \(C\) and \(G\) refers to the character of a given target parcel.

Each item of surrounding information is binarized into a value of 0 or 1. Because the surrounding information is converted into a binary array, the distance between the retrieved cases and the target parcel can be calculated based on the Hamming distance, as shown in equation (20):

\[ S_{\text{surround}} = 1 - \frac{\sum |V_C - V_G|}{T}, \]  

(20)

where \(T\) is 8, which refers to the total number of surrounding information, and \(V_C\) represents the surrounding information vector for case \(C\). For example, if case \(C\) matches the first and third features but does not match the second and fourth features, then \(V_C = (1, 0, 1, 0, \ldots)\). \(V_G\) represents the surrounding information vector for the given target parcel.

### 3.3. The XGBoost classifier

The XGBoost classifier is a sequential ensemble learning-based algorithm in which new models are created to correct the mistakes of prior models and then their predictions are combined to form a final prediction. Figure 2 shows the XGBoost computational process. The \(\hat{y}^{(0)}\) value is calculated by equation (21):

\[ \hat{y}^{(0)} = \sum_{k=1}^{t} f_k(x) = \hat{y}^{(t-1)} + f_t(x). \]  

(21)

where \(\hat{y}^{(0)}\) is the final tree model; \(\hat{y}^{(t-1)}\) is the previously generated tree model; \(f_t(x)\) is the newly generated tree model; and \(t\) is the total number of base tree models.

The optimal classifier model can be acquired through hyperparameter tuning. Hyperparameter tuning is intended to reduce the loss function. The target loss function is defined in equation (22). The regularization component is dependent on the number of leaves. The regularization term \(\Omega(f)\) is calculated by equation (23) to reduce the model’s complexity and avoid overfitting:

\[ L^{(0)} = \sum_{i=1}^{t} L\left(y, \hat{y}^{(i)}\right) + \sum_{i=1}^{t} \Omega(f_i). \]  

(22)

where \(y\) is the actual value, \(\hat{y}^{(i)}\) is the predicted value, \(L(y, \hat{y}^{(i)})\) is the loss function, and \(\Omega(f)\) is the regularization term, calculated as follows:

\[ \Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2. \]  

(23)

where \(T\) is the number of leaves, \(\omega\) represents the weights of the leaves, and \(\gamma\) and \(\lambda\) are coefficients whose default values are \(\gamma = 0\) and \(\lambda = 1\), respectively.

After training the XGBoost model, we also obtain the importance feature ranking. Then, we can learn directly from the data and transform subjective evaluations into quantifiable indicators.
3.4. The hierarchical retrieval process

After developing and validating the XGBoost classifier, we obtain an optimized classifier. Then, we use the saved optimal XGBoost classifier during the retrieval phase. The hierarchical retrieval process includes the following steps:

Step 1: Specify a target parcel T and its land-use type L_T. In addition, the expected design parcel information (FAR, greening rate, etc.) and the surrounding environment of the target parcel are specified.

Step 2: From case base C, select cases C_i = (C_i \subset C, i = 1, 2, \ldots, K) whose land-use type L_{C_i} = L_T. Then, the target parcel T and each case C_i form K pairwise parcels (T, C_1), (T, C_2), \ldots, (T, C_K), where K is the total number of selected cases.

Step 3: Geometry represents the target parcel T and each selected case C_i as shown in Table 1. Then, the similarity of each pairwise parcel (T, C_i) is measured using equations (15–18) and K distance vectors d_{TC_i} are obtained.

Step 4: Input the distance vectors d_{TC_i} into the trained XGBoost classifier. The output, \hat{y}_{TC_i} = \begin{cases} 0, & \text{dissimilar} \\ 1, & \text{similar} \end{cases}, specifies whether C_i is similar to the given target parcel T.

Step 5: Select all the cases C_j whose corresponding output \hat{y}_{TC_j} = 1 as the retrieved cases C_j = (C_j \subset C_j, j = 1, \ldots, m), where m is the total number of retrieved cases.

Step 6: The similarity scores \rho_j for each retrieved case C_j in terms of parcel information and surrounding environment are calculated with equations (19–20). Finally, architects can flexibly choose from the appropriate references based on their actual needs.

4. Experimental Results and Discussion

To verify the performance of the proposed approach, we first constructed a real dataset with labels. Then, we compared various evaluation criteria with different classifiers to evaluate the effectiveness of the XGBoost classifier. Finally, we demonstrate the entire hierarchical retrieval process using a real-world target parcel.

4.1. Dataset

In total, 1189 parcels in Beijing were collected to construct the real case base. Among these, 1102 residential parcels were adopted from Long (2016). Fifty one commercial parcels and 36 industrial parcels were selected using Open Street Map. We randomly selected some parcels of the residential land-use type to form pairwise parcels, for a total of 2553 pairs. These pairwise parcels were manually identified as similar or dissimilar. Each pairwise parcel was labeled by three architecture experts to ensure the correctness of the retrieval results. A pairwise parcel identified to be similar by all three experts was regarded as truly similar, i.e. a positive sample. Similarly, negative samples consisted of pairwise parcels identified as dissimilar by all three experts. Simultaneously, we expanded the negative samples by selecting some pairwise parcels with large differences. Finally, our dataset contained 1067 positive samples and 1642 negative samples. Then, we combined these two types of samples and shuffled them to construct the final dataset.

4.2. Evaluation criteria and experimental settings

To evaluate the performance of the XGBoost classifier, we used the metrics accuracy (ACC), recall, precision, F1-score, confusion matrix, receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC). The prediction results of XGBoost on the testing dataset fall into four categories: true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs). A confusion matrix provides a more detailed analysis. ACC is used to evaluate the model’s prediction accuracy. Recall represents the fraction of samples correctly identified among all the positive samples, while precision indicates the fraction of true positive samples among those the model predicts as positive; thus, precision and recall are contradictory measures. The F1-score is used to evaluate the model’s overall predictive performance by combining the recall and precision results. Accuracy, recall, precision, and F1-score are calculated as follows:

\[
\begin{align*}
\text{ACC} & = \frac{TP + TN}{TP + TN + FP + FN} \quad (24) \\
\text{Recall} & = \frac{TP}{TP + FN} \quad (25) \\
\text{Precision} & = \frac{TP}{TP + FP} \quad (26) \\
\text{F1} & = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (27)
\end{align*}
\]

The ROC curve is a graph of the false positive rate (FPR) vs. the true positive rate (TPR). The curve is drawn by traversing all thresholds. The AUC quantifies a model’s performance across the entire range of different classification thresholds.
all possible classification thresholds. An AUC value closer to 1 indicates a more effective method.

We used 90% of the overall dataset as a training dataset to optimize the XGBoost classifier and used the remaining 10% of the dataset as a testing dataset for prediction. We adopted fivefold cross-validation to evaluate the classifier’s performance. The parameter settings were determined by a grid search. The parameter settings for the optimal XGBoost classifier are shown in Table 2.

We conducted a set of comparative experiments to evaluate the effectiveness, robustness, and generalizability of the XGBoost classifier and implemented four commonly used
classification methods for comparison purposes: RF, SVM, KNN, and LR. The RF and SVM were employed because they are commonly used and perform well for classification problems. The KNN uses the Euclidean distance to calculate similarity and retrieve the closest cases. LR has been employed frequently in recent similar parcel studies. To ensure an unbiased and consistent experiment, we trained the compared classifiers on the same dataset and tuned their parameters.

4.3. Experimental results

The performances of the XGBoost classifier and the baselines were evaluated using the abovementioned evaluation criteria. The prediction accuracies of these classifiers are shown in Table 3. The XGBoost classifier achieved the highest prediction accuracy of 94.46%, which was 0.73%, 5.16%, 2.95%, and 2.95% higher than the accuracies of the RF, SVM, KNN and LR.
Table 5: Ablation studies on removing different features. "w/o" indicates without the corresponding feature.

| Proposed | ACC     |
|----------|---------|
| w/o Equivalent rectangular index | 0.9336  |
| w/o Rectangularity | 0.9336  |
| w/o Hu moments | 0.9299  |
| w/o Perimeter | 0.9252  |
| w/o Convexity | 0.9336  |
| w/o Fractal dimension | 0.9373  |
| w/o Compactness weight axis | 0.9336  |
| Full features | 0.9446  |

Table 6: Ablation studies on the different combinations of features. The ✓ indicates that the subset was used and ✗ indicates that the subset was not used.

| Basic descriptors | Simple descriptors | Complicated descriptors | ACC   |
|-------------------|--------------------|-------------------------|-------|
| ✓                 | ✗                  | ✓                       | 0.8893|
| ✓                 | ✗                  | ✗                       | 0.9225|
| ✓                 | ✓                  | ✗                       | 0.7970|
| ✓                 | ✓                  | ✓                       | 0.9373|
| ✓                 | ✗                  | ✓                       | 0.9114|
| ✓                 | ✓                  | ✓                       | 0.9188|
| ✓                 | ✓                  | ✓                       | 0.9446|

models, respectively. Table 4 provides the detailed results for the XGBoost classifier, including a confusion matrix and recall and precision scores. The precision and recall values both exceed 92%, showing outstanding performance.

The detailed comparison results are shown in Figs 3 and 4. In terms of the recall rate and F1-scores, the XGBoost classifier outperformed the other classifiers. Although the precision of XGBoost was slightly lower than that of the RF classifier, it achieved a higher recall. The XGBoost curve is the closest to the upper left corner, and its AUC value is the highest. The AUC value indicates that the XGBoost classifier achieved good performance across all possible classification thresholds. Overall, these experimental results provide supporting evidence for adopting the XGBoost classifier to judge parcel similarity. The XGBoost classifier improves the accuracy of retrieval results.

In addition, the ensemble learning-based models XGBoost and RF can provide feature importance ranking, as shown in Fig. 5, to identify which features are crucial when measuring parcel similarity. The 7 shared features in the top 10 features of these models are the compactness weight axis, length, convexity, rectangularity, fractal, Hu moment invariants, and equivalent rectangular index. These features are the main factors that affect the internal arrangement of a parcel.

We used 7 shared features from the top 10 features of XGBoost and RF to conduct ablation studies to evaluate the effects of different features on the accuracy of our model; the results are shown in Table 5. It can be found that the ACC decreases when removing each shared feature by at least 0.77% and at most 1.95%, which suggests that each shared feature is conducive to the performance of the model.

We also carried out ablation studies under different feature subset combinations to assess the impact of basic descriptors, simple descriptors, and complicated descriptors. Table 6 compares the results with different subset combinations. The combination of three subsets achieves the best ACC, meaning that each subset is crucial to retrieving similar parcels. The model with the combination of basic and simple descriptors obtains the second best ACC, indicating that basic and simple descriptors have the greatest impact on the model. In addition, the model with simple descriptors achieves higher accuracy than the one with basic descriptors or complicated descriptors, which suggests that the simple descriptors make more contribution on the performance of the proposed model.

4.5. The case study

Based on the above experiments, we obtained an optimal XGBoost classifier suitable for achieving retrieval results with improved accuracy. In this subsection, we report the results of retrieved similar cases to demonstrate the complete hierarchical retrieval approach using a real-world parcel.

The user provides data regarding the target parcel in a shapefile format. In this case, the land-use type is residential. The design information on the target parcel and its surrounding information are shown in Fig. 6.

First, we selected all 1102 cases with residential land-use types from the case base. After feature representation, we obtained 1102 distance vectors between the target parcel and each selected case. These distance vectors were input to the saved XGBoost classifier. All corresponding cases whose output result was 1 were considered the retrieved cases. It should be noted that the number of retrieved cases is uncertain.

As shown in Fig. 7, three retrieved cases similar to the target parcel were retrieved, with the building layouts shown in the figure. Real-world building layouts contain various types of knowledge, and architects can use them as design references. From Fig. 7, we can intuitively discover the following:

1) The buildings’ colors indicate three different building heights: low-rise, mid-rise, and high-rise buildings. Darker colors indicate taller buildings.

![Figure 6: Detailed information on the given target parcel.](https://academic.oup.com/jcde/article/8/6/1548/6425803)
Table 7: The environmental attributes of the retrieved cases.

| Attributes \ ID | 363 | 660 | 755 |
|---------------|-----|-----|-----|
| Parcel information | FAR | 3   | 1.3 | 2.3 |
| Price (rmb)    | 72K | 80K | 69K |
| Entrances      | 3   | 7   | 2   |
| Greening rate  | 40% | 30% | 40% |
| Parking space  | 200 | 150 | 312 |
| Surrounding environment | School | 1 | 1 | 1 |
| Mall           | 0   | 1   | 0   |
| Restaurant     | 1   | 1   | 1   |
| Hospital       | 1   | 1   | 1   |
| Subway         | 1   | 1   | 1   |
| Sightseeing spot | 0   | 1   | 1   |
| Park           | 0   | 1   | 1   |
| Water          | 1   | 0   | 0   |

2) The distance between each building is visually displayed in the figure. Architects can modify the building layout slightly to design the target parcel and save design time.

3) The three cases represent three different architectural styles used in residential areas. Because Beijing is an ancient capital, the design span is overwhelmingly large; thus, the architectural styles are quite different. Different types of residential buildings from different periods may be retrieved, providing a diverse set of design references for architects.

The retrieved cases can be further analysed based on their environmental attributes, which are listed in Table 7. Architects should carefully consider these environmental attributes during the actual design process. The similarity scores of the retrieved cases are calculated based on six aspects (FAR, greening rate, etc.) as shown in Fig. 8. Similarity scores indicate the similarity between the design conditions and constraints of these retrieved cases with the target parcel faced by the architects. The architects can select from and revise these retrieved cases to assist in reaching a unique confirmed solution for the target parcel design.

To better evaluate the effectiveness of our hierarchical retrieval approach, we compare our approach with Sun’s work (Sun et al., 2016). Their method retrieves similar design cases based on a fixed threshold. Figure 9 shows the comparison results.

From Fig. 9, it can be found that our hierarchical retrieval method provided more similar cases for target parcels a and b, which suggests that more references were provided. For target parcel c, Sun’s method could not capture the special shape and the retrieval results were quite different from the target parcel. For parcel d, Sun’s retrieval result was also different because the retrieval method is limited by retrieving shapes with the same corners. Therefore, the hierarchical retrieval method is more accurate, even if the target parcel has a complex shape or a simple shape but many corners.

5. Conclusions

In this paper, we propose a computational approach for providing site planning references through hierarchical retrieval of similar cases. All cases are represented by land-use attributes, geometric attributes, and environmental attributes. These features are used to calculate the similarity distances of case conditions and design constraints. The similarity distances are input to the ensemble learning model (XGBoost classifier), which provides predictions regarding the similarity of existing cases to the target parcel. The final prediction accuracy is 94.46%, which...
Figure 8: The similarity scores of the retrieved cases (the numbers in the figure are the IDs of cases in the case base).

Figure 9: Retrieval results using Sun and our hierarchical retrieval approaches.

demonstrates the superiority of the XGBoost classifier over prior baseline ML models. The real-world case study demonstrates that this approach can provide a diverse set of reference cases. Architects can then select the most appropriate cases by considering the surrounding environment and other design requirements. The model’s ability can be enhanced as the case base grows by collecting additional cases from different cities and building types.

It should be noted that there is still a distance between the retrieved cases and the final design solution. To complete the design process, retrieved cases need to be adapted to fit the requirements of the new situations. Case adaptation can be implemented by the architects themselves through creative ideas or by automatic computing techniques following given rules or constraints. Automatic case adaptation is also a challenging problem that has been addressed in recent decades. Regarding artificial case adaptation, there are several pathways that can be flexibly utilized by architects based on the outputs of our tool. For example, architects can select one or multiple cases according to the detailed similarity scores and then fine-tune or fuse them to complete the final plan. Architects can also use these retrieved cases for redesign by sampling some components of these cases. In addition, the architects can utilize the retrieved cases as expected but ignore the scores evaluated by our method. An alternative for case adaptation is using automatic learning algorithms, which is also addressed in our future work.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61871020), the High Level Innovation Team Construction Project of Beijing Municipal Universities (No. IDHT20190506), the National Key Research and Development Plan Task of China (No. 2020YFF0305501), and the Key Science and Technology Plan Project of Beijing Municipal Education Commission of China (No. KZ201810016019).

Conflict of interest statement

None declared.
References
Alessandra, F. (2018). Resilience and urban design: A systems approach to the study of resilience in urban form. [Doctoral dissertation, University of Strathclyde].
Basaraner, M., & Cetinkaya, S. (2017). Performance of shape indices and classification schemes for characterising perceptual shape complexity of building footprints in GIS. International Journal of Geographical Information Science, 31(10), 1–26. https://doi.org/10.1080/13658816.2017.1346257.
Beer, A., & Higgins, C. (2004). Environnemental planning for site development: A manual for sustainable local planning and design. Routledge.
Chen, T., & Guestrin, C. M. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785.
Domeshek, E. A., & Kolodner, J. L. (1992). A case-based design aid for architecture. In J. S. Gero, & F. Sudweeks (Eds.), Artificial Intelligence in Design (pp. 497–516). Springer. https://doi.org/10.1007/978-94-011-2787-5_25.
Fang, A., Yan, H., & Zhang, L. (2012). Description and calculation of similarity degrees between individual buildings in multi-scale map space. Science of Surveying and Mapping, 37(1), 98–100. https://doi.org/10.16251/j.cnki.1009-2307.2012.01.028.
Fleischmann, M. (2019). Momepy: Urban morphology measuring toolkit. Journal of Open Source Software, 4(43), 1807. https://doi.org/10.21105/joss.01807.
Gero, J. S., Coyne, R. D., Rosenman, M. A., & Radford, A. D. (1987). Innovation and creativity in knowledge-based CAD. Expert Systems in Computer-Aided Design, 435–471.
Gil, J., Montenegro, N., José, N. B., & José, P. D. (2012). On the discovery of urban typologies: Data mining the multi-dimensional character of neighbourhoods. In Proceedings of the 27th Computer Aided Architectural Design Conference (pp. 27–40).
Hao, Y., Tang, W., Zhao, Y., & Li, N. (2008). Areal feature matching algorithm based on spatial similarity. Acta Geodaetica et Cartographica Sinica, 37(4), 103–108. https://doi.org/10.1177/111786.
Hernesiemi, J. A., Mahdian, S., Tynkkynen, J. A., Lyytikainen, L. P., Mishra, P. P., Lehtimaki, T., Eskola, M., Nikus, K., Antila, K., & Oksala, N. (2019). Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome—the Maddex study. Annals of Medicine, 51(2), 156–163. https://doi.org/10.1080/07853890.2019.1596302.
Hu, M. K. (1962). Visual pattern recognition by moment invariants. IRE Transactions on Information Theory, 8(2), 179–187. https://doi.org/10.1109/TIT.1962.1057692.
Hua, K., & Faltings, B. (1993). Exploring case-based building design—CADRE. Artificial Intelligence for Engineering, Design, Analysis and Manufacturing, 7, 135–143. https://doi.org/10.1017/S0950667196000822.
Hua, K., Fairings, B., & Smith, I. (1996). CADRE: Case-based geometric design. Artificial Intelligence in Engineering, 10(2), 171–183. https://doi.org/10.1016/0954–1810(95)00026–7.
Jacob, D., Alexis, P., Ombretta, R., Mattia, Z., Emanuele, S., Mark, P., & Sergio, P. (2017). On the origin of spaces: Morphometric foundations of urban form evolution. Environment and Planning B: Urban Analytics and City Science, 46(4), 707–730. https://doi.org/10.1068/b339908317725075.
Lawson, B. (2004). What designers know. Architectural Press.
Li, B., & Han, D. (2011). Technical comprehension of architectural generative design and its prospects. Architectural Journal, 6, 91–100. https://doi.org/10.3969/j.issn.0529-1399.2011.06.016.
Li, L., Zhao, X., Tseng, M. L., & Tan, R. R. (2020). Short-term wind power forecasting based on support vector machine with improved dragonfly algorithm. Journal of Cleaner Production, 242, 118447. https://doi.org/10.1016/j.jclepro.2019.118447.
Ling, L., Wang, X. B., & Li, C. G. (2013). An automatic building-layout algorithm based on inner available region. Computer Technology and Development, 8, 1–4. https://doi.org/10.3969/j.issn.1673–629X.2013.08.001.
Long, Y. (2016). Redefining Chinese city system with emerging new data. Applied Geography, 75, 36–48. https://doi.org/10.1016/j.apgeog.2016.08.002.
Luca, C., Francesco, G., & Elena, M. (2007). Generative platform for urban and regional design. Automation in Construction, 16(1), 70–77. https://doi.org/10.1016/j.autcon.2005.10.008.
McGarigal, K. (1995). FRAGSTATS: Spatial pattern analysis program for quantifying landscape structure (Report No. PWN-GTR-351). U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. https://doi.org/10.2737/PNW-GTR-35-4.
Maher, M. L., & de Silva Garza, A. G. (1997). Case-based reasoning in design. IEEE Expert, 12(2), 34–41. https://doi.org/10.1109/64.585102.
Morgenstern, Y., Schmidt, F., Hartmann, F., Tiedemann, H., Prokott, E., Maiello, G., & Fleming, R. (2019). An image computable model of visual shape similarity. Journal of Vision, 19(10), 37c. https://doi.org/10.1167/19.10.37c.
Naghibi, S. A., Vafakhah, M., Hashemi, H., Pradhan, B., & Alavi, S. J. (2019). Water resources management through flood spreading project suitability mapping using frequency ratio, k-nearest neighbours, and random forest algorithms. Natural Research, 29, 1–19. https://doi.org/10.1007/s11053–019–09530–4.
Okeil, A. (2010). Hybrid design environments: Immersive and non-immersive architectural design. Journal of Information Technology in Construction, 15(16), 202–216.
Rafael, C. G., & Richard, E. W. (2011). Digital image processing. (3rd ed.). Publishing House of Electronics Industry.
Rosenman, M. A., Gero, J. S., & Oxman, R. E. (1991). What’s in a case: The use of case bases, knowledge bases and databases in design. In Proceedings of CAAD Futures’ 91: Computer Aided Architectural Design Futures: Education, Research, Applications (pp. 285–300).
Schirmer, P. M., & Axhausen, K. W. (2015). A multiscale classification of urban morphology. Journal of Transport and Land Use, 9(1), 101–130. https://doi.org/10.5198/jtlu.2015.667.
Söhnmez, N. O. (2018). A review of the use of examples for automating architectural design tasks. Computer-Aided Design, 96, 13–30. https://doi.org/10.1016/j.cad.2017.10.005.
Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028.
Sun, C., Luo, Q., Tu, P., & Duan, R. (2016). Study on generation method of 3D building forms for street blocks in urban design applications. Urbanism and Architecture, 138(01), 114–117. https://doi.org/10.19892/j.cnki.csjz.2016.01.022.
Sun, C., Luo, Q., Song, X., Xie, J., & Rao, J. (2017). A practical approach to generating 3D city models: A case study of the density zoning project in Beihai. Architecture Journal, 8, 77–81. https://doi.org/10.3969/j.issn.0529–1399.2017.08.014.
Wang, S., Celebi, M. E., Zhang, Y.-D., Yu, X., Lu, S., Yao, X., Zhou, Q., Miguel, M.-G., Tian, Y., Gorriz, J. M., & Tyukin, I. (2021). Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects.
Zhang, Y.-D., Dong, Z., Wang, S.-H., Yu, X., Yao, X., Zhou, Q., Hu, H., Li, M., Jiménez-Mesa, C., Ramirez, J., Martinez, F. J., & Gorriz, J. M. (2020). Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. Information Fusion, 64, 149–187. https://doi.org/10.1016/j.inffus.2020.07.006.

Zheng, Y., & Zhang, Q. (2013). An approach to identical areal entity matching based on topology and spatial similarity. Journal of Geomatics Science and Technology, 30(5), 510–514. https://doi.org/10.3969/j.issn.1673-6338.2013.05.016.

Zofie, C. (2017). Urban morphology in Prague: Automatic classification in GIS. [Doctoral dissertation, Czech Technical University].