www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

MULTI-SCALE INFORMATION RETRIEVAL FOR BIM USING HIERARCHICAL STRUCTURE MODELLING AND NATURAL LANGUAGE PROCESSING

SUBMITTED: December 2020
REVISED: April 2021
PUBLISHED: July 2021
GUEST EDITORS: Kirti Ruikar, Ketan Kotecha, Sayali Sandbhor, Albert Thomas
DOI: 10.36680/j.itcon.2021.022

Jia Wang, Ph.D.,
School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture
E-mail: wangjia@bucea.edu.cn

Xinao Gao, M.E.,
School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture
E-mail: 1473452944@qq.com

Xiaoping Zhou,Ph.D.,
School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture
E-mail: lukefchou@gmail.com

Qingsheng Xie, M.E.
School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture
E-mail: 1954868773@qq.com

SUMMARY: *Building Information Modelling (BIM) captures numerous information the life cycle of buildings. Information retrieval is one of fundamental tasks for BIM decision support systems. Currently, most of the BIM retrieval systems focused on querying existing BIM models from a BIM database, seldom studies explore the multi-scale information retrieval from a BIM model. This study proposes a multi-scale information retrieval scheme for BIM jointly using the hierarchical structure of BIM and Natural Language Processing (NLP). Firstly, a BIM Hierarchy Tree (BIH-Tree) model is constructed to interpret the hierarchical structure relations among BIM data according to Industry Foundation Class (IFC) specification. Secondly, technologies of NLP and International Framework for Dictionaries (IFD) are employed to parse and unify the queries. Thirdly, a novel information retrieval scheme is developed to find the multi-scale information associated with the unified queries. Finally, the retrieval method proposed in this study is applied to an engineering case, and the practical results show that the proposed method is effective.*

KEYWORDS: *Building Information Modelling (BIM); Multi-scale Building Information; Information Retrieval; Hierarchy Structure; Natural Language Processing (NLP)*

REFERENCE: *Jia Wang, Xinao Gao, Xiaoping Zhou, Qingsheng Xie (2021). Multi-scale Information Retrieval for BIM using Hierarchical Structure Modelling and Natural Language Processing. Journal of Information Technology in Construction (ITcon), Special issue: 'Next Generation ICT - How distant is ubiquitous computing?', Vol. 26, pg. 409-426, DOI: 10.36680/j.itcon.2021.022*

1. INTRODUCTION

Building Information Modelling (BIM) is a pivotal technology, which has been developed since the early 2000s in Architecture, Engineering and Construction (AEC) (Ghaffarianhoseini et al., 2017). BIM captures multi-dimensional building information, i.e. spatial geometry information (Liu et al., 2017), quantity and attributes of buildings, related drawings, procurement details, submission process and other construction documents. BIM can be used in the facility planning, design, construction and operations, changing the traditional work flow and project delivery process (Abualdenien et al., 2020), reducing project cost, and improving production quality and efficiency (Jones, 2020). According to the survey (Yalcinkaya and Singh, 2015), BIM has been widely recognized in the construction industry worldwide, and BIM usage will continue to increase.

A BIM model contains a large amount of multi-scale information (component-level, attribute-level, etc.), and the information retrieval requirements differ from users and applications. For instance, the managers have to check both the components and the attribute values of the building (e.g., the width of the court) in conformity checking (Zhang and El-Gohary, 2015a). In the process of indoor navigation, it is necessary to extract geometric data from BIM for the construction of path navigation algorithm (Zhou et al., 2020). In the operation and maintenance phase, equipment managers need to extract information in various aspects of the operation performance from the BIM model (Chen, 2019), while maintenance staff tends to extract failure pattern information in building systems and components from CMMS databases (Gunay et al.,2019).

Most of the current retrieval systems are component-level. To reduce the threshold of using the BIM object database for practitioners and the general public, Wu et al. (2019)proposed a retrieval engine for BIM object database, which mainly retrieved and recommended component-level information. Xie et al. (2019) proposed a method of matching real-world facilities to BIM data, which is also on component-level. The BIM model contains information of the whole life cycle of the building. Most retrieval systems can only return component-level information. If the retrieval system can also directly return attribute information and other information, then it can assist construction, equipment management, code compliance inspection, and other processes. The current retrieval system can no longer meet the needs of users, so a system capable of retrieving multi-scale information is urgently needed.

The study aims to development of a new building information retrieval scheme by using BIM hierarchical structure and NLP technology. The contributions include:

- A BIM hierarchical tree model (BIH-Tree) that hierarchically represents the BIM model data and establishes the association between the data was provided in the model. Since BIH-Tree integrates multi-scale information, it is possible to facilitate the multi-scale information retrieval from a BIM model.

- A multi-scale building information retrieval scheme was proposed. The scheme jointly utilized the NLP and BIH-Tree, it can understand the query and return the correct search results.

- The method proposed in this study is applied to engineering projects to meet the needs of multi-scale information retrieval of space, equipment and management data in BIM model in the process of rapid decision.

This article proceeds as follows. Section 2 introduces the extensive work in the field of BIM and proposes the defects of current BIM retrieval systems in detail. Section 3 analyzes the tree structure, which can improve the understandability of the data, and gives the definition and principle of the BIM hierarchical tree. Section 4 systematically explores the process of parsing and unifying queries using NLP and IFD. Section 5 introduces a novel information retrieval scheme for searching for multi-scale information associated with a unified query. Section 6 conducted experiments on a real project model to illustrate the effectiveness of the BIM retrieval scheme. Section 7 provides conclusions and future works.

2. RELATED RESEARCHES

2.1 Extensive work in BIM

In recent years, BIM Technology in the field of construction continues to develop. Increasing researchers have begun to employ the BIM technology to address problems. Energy performance analysis based on BIM was presented, quantifying the impact of building envelope conditions on energy use (Jeon et al., 2019). Research has been conducted on integrated BIM and real-time data from Internet of Things (IoT) equipment to improve the efficiency of construction and operation (Tang et al., 2019). Currently, some efforts have been made on BIM visualization, such as the online BIM visualization system based on IFC and WebGL (Zhou et al., 2018). The OutDet algorithm was proposed to select the representative geometric features of BIM models, to eliminate most unnecessary features, and to reduce the computational burden of model visualization (Zhou et al., 2019). To greatly reduce the calculation time and space the IFC file comparison process, Shi proposed a content-based automatic comparison approach, named IFCdiff (Shi et al., 2018).

Automatic compliance check is also one of the important tasks of BIM. The accuracy of information retrieval greatly affects the process of automatic compliance check and model check. In order to extract the information of concept, relation and instance from BIM, and to automatically infer and keep consistency with the information in regulatory documents, the automated extraction of information from building information models into a semantic logic-based representation was proposed (Zhang and El-Gohary, 2015b). To enhance BIM's support for automatic compliance checking, a semi-automatic extension of building information model using semantic natural language processing technology was suggested (Zhang and El-Gohary, 2016), and automated code compliance checking based on a visual language and BIM was raised (Preidel and Borrmann, 2015).

The studies discussed above utilized multi-scale information in the BIM model. Information retrieval is the foundation of all the above processes. If the information extraction is correct and convenient, the processes of energy analysis, facility management (FM), BIM visualization and automated code compliance are easy to implement. Therefore, this paper proposes a BIM retrieval system which can retrieve multi-scale information, so that the retrieval results could better meet the users' demands and increase the BIM value.

2.2 BIM retrieval system

This section mainly introduces two types of BIM retrieval system: the retrieval of components in building component library (Duddy et al., 2013), and the retrieval of information in BIM (Preidel et al., 2017).

2.2.1 Retrieval from Building Component Library

The building component library based on IFC standard can provide information from different suppliers, which is helpful for information sharing and exchange in building life cycle (Wei et al., 2010). To improve the retrieval accuracy, some researchers have conducted researches on the components retrieval in the database. Fleming introduced the Department of Energy (DOE) Building Components Library (BCL), an online repository of building components that could be directly used to create energy models (Fleming et al., 2012). This building component library had a standardized list of components, which made it convenient for designers to retrieve building components. The INNOVance (Name of a national research project) project (Pasini et al., 2017) in Italy proposed a solution to manage construction information. In the INNOVance project, a classification system, an information structure, a database and BIM objects had been developed for construction work products, spaces characterized buildings, and involving actors along the whole building process. To effectively retrieve online BIM resources, Gao et al. (2015) developed a prototype semantic search engine, named BIMSeek. Gao et al. (2017) developed a prototype annotation system called BIMTag, which reduced the ambiguity and unclearness of natural language in BIM documents, and combined it with search engines to improve the retrieval efficiency.

These studies have improved the accuracy of the search in some respects, but most of these studies lack the exploration of the extraction of internal information in the BIM model. Because the data in the BIM model is multi-scale information (e.g., the area of the floor, etc.), there is a relationship (e.g., affiliation, etc.) between these data. The method of retrieving information in the component library is not suitable for retrieving information in the BIM model.

2.2.2 Retrieval of information into BIM Model

Accurately and effectively extracting information from BIM models is the key to implementing all BIM processes.

Information retrieval is the basis for extracting information from BIM models. However, seldom studies explore the multi-scale information retrieval from a BIM model. To efficiently retrieve the needed BIM model, a 3D building model retrieval method was proposed, using airborne LiDAR (Lidar is a laser, global positioning system (GPS) and inertial navigation system (INS) technology in one system, used to obtain point cloud data and generate accurate digital 3D model) point cloud as input query (Chen et al., 2017). However, this retrieval method uses Airborne LIDAR point cloud as input query, so it is necessary to have matching point cloud files of specific buildings to use this retrieval tool. Preidel et al. used visual programming language QL4BIM and VCCL (The name of a visual programming language) in BIM environment to process building model information, and to improve the efficiency of information retrieval from building information model (Preidel et al., 2017). However, the retrieval efficiency of this method depended on the proficiency of the query language, so only those who were proficient in the two query languages could retrieve the results efficiently. To reduce the complexity of data model and improve the retrieval efficiency, Gui et al. (2019) proposed a partial data model retrieval method based on IFC, which extracted the required attribute information from the central IFC model, and used the local model instead of the overall model to complete the retrieval. In addition, a cloud BIM intelligent data retrieval and representation method based on natural language was proposed (Lin et al., 2016). However, this method led to the lack of correlation between different types of documents, so that the retrieval system was difficult to retrieve multi-scale information.

The researches discussed above indicate that BIM retrieval system has made some progress in retrieving component level information, but it is still a challenge to use natural language to retrieve multi-scale information in BIM model. Therefore, a multi-scale information retrieval system based on hierarchical structure and NLP is proposed.

2.3 Problems in BIM retrieval system

At present, BIM is widely used in the field of AEC. Information retrieval is the basis of the BIM application process. Most of the current researches on BIM retrieval systems are for the retrieval of components in the component library, and there are few researches on the retrieval of information in models (As shown in Fig. 1). The above research (Tang et al., 2019) shows that BIM applications require multi-scale information, thus it is urgent to develop a retrieval system capable of retrieving multi-scale information in the model. Additionally, With the continuous development of the computer technology, the demands of using natural language to communicate with computers are mounting. Many parsers for processing various texts have been developed, such as Stanford parser (De et al., 2006), NLTK (Hardeniya et al., 2016), Minipar (Lin, 2003), etc. While most current retrieval systems still retrieve information based on keywords, this study uses natural language processing technology and IFD (Bell et al., 2008) to complete the parse and to unify query language.

FIG. 1: Building Information Query in BIM

Therefore, different from the previous method of retrieving component level in BIM component library, this research combines the data of hierarchical structured representation BIM model with NLP to develop a retrieval system that can retrieve multi-scale information in BIM model to meet the needs of users in different projects and to improve the productivity of AEC industry.

2.4 Research methods and contents

The purpose of this study is to implement a retrieval method that can query multi-scale information in BIM model. By consulting a large number of literature on building information retrieval methods, we know that the existing BIM retrieval methods can not meet the needs of rapid decision-making in building projects for accurate query of multi-scale information such as space, equipment and management data in BIM data. By studying the existing information retrieval methods and theories applied to BIM component library and BIM model, we know that the application of natural language processing technology to multi-scale information retrieval in BIM model can make the retrieval system better understand the user's query intention and return the building information that meets the user's needs. Moreover, the hierarchical structured representation of multi-scale information in BIM model can help the retrieval system to understand the association between data information and make correct processing. Therefore, this study combines hierarchical tree and natural language processing technology to help BIM retrieval system better understand the user's query intention and return the correct query results. The main research methods are: (1) analysis of data information in BIM model. IFC Standard defines the reference relationship between data such as building components and attribute information. IFC Standard is used to analyze the data information in BIM model to build the hierarchical relationship between data information. The construction of this relationship is helpful to the management and retrieval of data. (2) Natural language processing technology is used to analyze the query. Because the data stored in BIM is written in formal description language, it can not be obtained directly by natural language. Therefore, on the basis of the built BIH tree, we use natural language processing technology to map the query to BIH tree to analyze the user's query intention and make the retrieval system return the query results that meet the user's needs.

3. BIM HIERARCHY TREE MODEL

This section introduces the hierarchical representation strategy of BIM data based on IFC, then gives out the building principle of BIH-Tree (BIM Hierarchy Tree).

3.1 Hierarchical Representation Strategy for IFC-Based BIM Data

IfcRelContainedInSpatialStructure (Inclusion relation of spatial structure) is introduced in the IFC Standard to assign elements to a certain level of the spatial project structure. Any element can only be assigned once at a certain level. Moreover, building elements must be hierarchical, and elements can only be contained in a single spatial structure element. An element can refer to many spatial structure elements, and reference relationships are represented by the IfcRelContainedInSpatialStructure. The predefined spatial structure elements for which elements can be assigned include: IfcSite, IfcBuilding, IfcBuildingStorey and IfcSpace (The technical terms used to describe the spatial information of engineering project in IFC Standard). The same element type can be assigned to different spatial structure elements depending on the context of the occurrence. Fig. 2 shows the use of IfcRelContainedInSpatialStructure to assign a stair, a wall, a bookshelf and a desk to three different levels within the spatial structure. The right side of the graph shows the attribute information corresponding to each element.

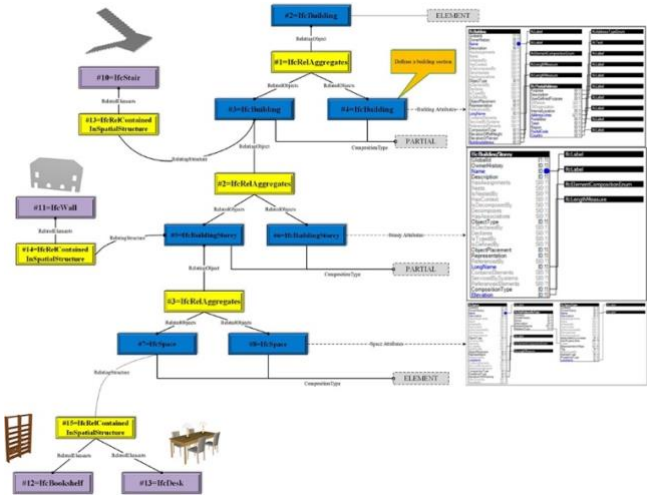


FIG. 2: Relationship for Spatial Structure Containment

3.2 Definition and Principle

As a data structure, the tree structure can be used to represent the hierarchical relationship between data elements. The tree-based visualization is expected to improve the understandability and reading speed of the logic-based representation (Zhang, 2017). Information in BIM can be divided into different hierarchical structures based on IFC standard and be expressed by tree structure. Ifcrelcontainedinspatialstructure in IFC Standard is used to identify the hierarchical relationship between components in BIM model. Thus the tree structure based on IFC standard is defined as the BIM Hierarchy Tree, or BIH-Tree.

BIH-Tree is a hierarchical representation of building information. The root node of the BIH-Tree represents the model of the building. Under the root node, BIH-Tree can be divided into $i > 0$ levels. Each level of the BIH-Tree consists of $n > 0$ sub-nodes, each of which represents a building element, as shown in Fig. 3. This study defines the node of the solid line connection as the Building Node, referred to as 'BN'. The 'BN' are represented by $K^o = \{K_{i1}, K_{i2}, \ldots, K_{in}\}$, where K_{ij} $(1 \le i \le n, 1 \le j \le n)$ is the j-th building node of the i-th level in the BIH-Tree. The building on each node in the BIH-Tree contains $n > 0$ attribute information, which is connected with a dotted line and represented by $M = \{m_1, m_2, \ldots, m_n\}$, and m_i $(1 \le i \le n)$ represents the i-th attribute information of the 'BN'. This study defines dotted-line nodes as Attribute Nodes, referred to as 'AN'. Each BNorAN is identified by a unique key, represented by $Key = \{key_1, key_2, \ldots, key_n\}$, and each node also contains a parent key, represented by $PKey = \{pkey_1, pkey_2, \ldots, pkey_n\}$. If the key_i and the $pkey_j$ are the same, they are considered to be a parent-child relationship. The two nodes ('BN' and 'BN' or 'BN' and 'AN') connected by solid lines or dotted lines in the Fig. 3 express the hierarchical relationship of information in BIM. For example, K_{21} represents a sub-building element of K_{11}, so does the $PKey$ of K_{21} points to the Key of K_{11}. And m_1 represents an attribute information of building element, and level i+1 represents the attribute information of the building node of level i. Thus, the hierarchical relationship of BIM data is expressed as a BIM Hierarchy Tree.

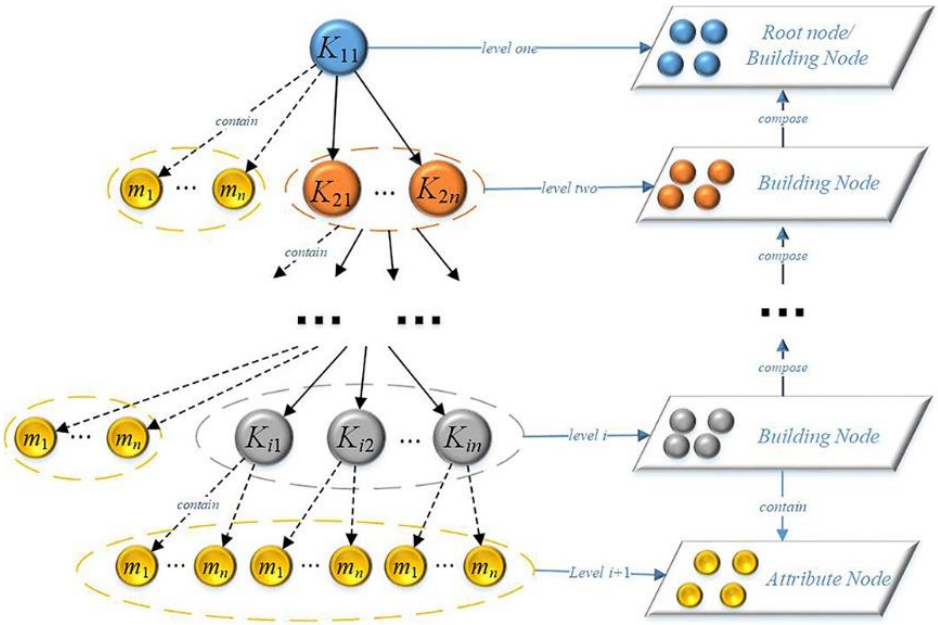


FIG. 3: BIH-Tree Model

Algorithm 1 summarizes the whole construction processes of BIH-Tree.

The flowchart is shown in Fig. 4.

Algorithm 1: Establishment of BIH-Tree

Input: BN and AN
Output: BIH-Tree
Extract
$K^o = \{K_{i1}, K_{i2}, \ldots, K_{in}\}$.
$M = \{m_1, m_2, \ldots, m_n\}$.
$Key = \{key_1, key_2, \ldots, key_n\}$.
$PKey = \{pkey_1, pkey_2, \ldots, pkey_n\}$.
$K_i, K_j \in K^o$, $key_i \in Key$, $pkey_j \in PKey$.
for BN **in** K^o:
  if $key_i = pkey_j$:
    Connect K_i, K_j with solid lines.
  else :
    K_i and K_j are not affiliated
    not connected.
end for
for AN **in** M:
  if $key_i = pkey_j$:
    Connect m_i , K_{ij} with dotted line.
  else:
    m_i and K_{ij} are not affiliated
    not connected.
end for

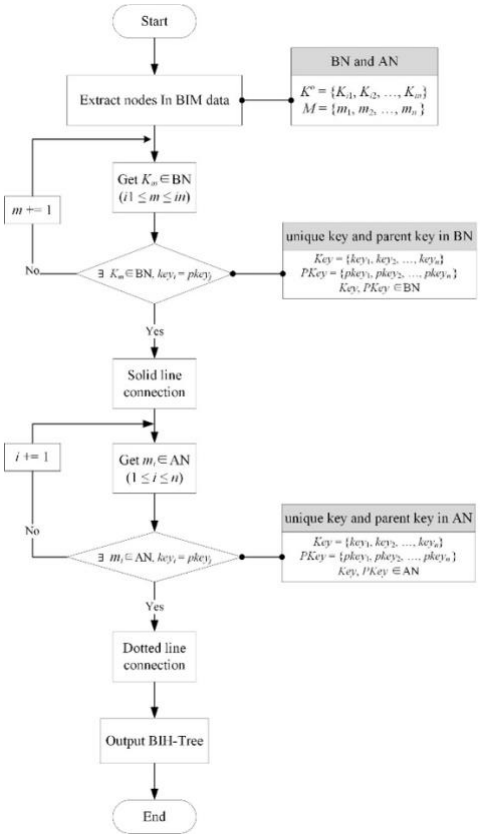


FIG. 4: Flow chart of algorithm 1

4. QUERY PARSING

After using BIH-Tree to represent the data information in BIM model according to IFC Standard, the retrieval system still can not support users to query the multi-scale information in BIM model with natural language, so it is necessary to process the natural language query statements to identify the user's query intention. This section mainly introduces the natural statements of parsing and unifying user input. The whole process includes: 1) word segmentation, to divide the natural sentences input by the user into separate words and punctuation marks. 2) Semantic disambiguation, to map the words or phrases obtained after the word segmentation to the BIM hierarchical tree in realizing the unified representation of words or phrases. 3) Syntactic analysis, to determine the constraint relationship between keywords. The specific steps are detailed below.

4.1 NLP-Based Word Segmentation

Word segmentation is a key step in semantic disambiguation and syntactic analysis. The accuracy of word segmentation plays a decisive role in the retrieval results. Because there are a lot of domain words in architecture, such as "Floor 2" will be divided into "Floor" and "2", the segmentation results cannot meet the system application. Thus we add construction domain dictionary and project dictionary to solve this problem. The natural language input by the user is divided into $n > 0$ words. Segmentation results include nouns, adjectives, adverbs, numerals, prepositions, conjunctions, pronouns, etc. The word segmentation result is expressed by $W = \{w_1, w_2, \ldots, w_n\}$, where w_i $(1 \le i \le n)$ is the i-th word obtained by segmentation.

4.2 Semantic Disambiguation

Due to the complexity and ambiguity of natural language, there are numerous ways to describe the same building information. For example, "student dormitory" and "student apartment" are both used to describe the buildings where students live. Abundant kinds of descriptive forms will be generated in describing the same object due to the difference of country, region, language, culture, etc. Due to the above factors, the existed information retrieval system cannot parse the distinct expressions of the same building component, i.e. it cannot retrieve the corresponding components. Therefore, this study proposes to use IFD to unify the query. IFD defines every architectural concept as the only GUID in the world, ensuring the unification of information. Building information provider only provides the GUID corresponding to the building during information exchange, so it eliminates the ambiguity caused by language, culture, etc.

In this study, IFD is used to disambiguate the segmented words obtained. An IFD-based dictionary of building information is defined, which contains $n > 0$ words and is expressed by $D = \{d_1, d_2, \ldots, d_n\}$, where d_i represents the i-th building information in the dictionary D. The elements representing building nodes and attribute nodes on the BIH-Tree in set W can find their unique corresponding GUID in the dictionary D, which avoids ambiguity. If some words in the W cannot represent the nodes on the BIH-Tree, disambiguation is not necessary. Finally, we get $n > 0$ words semantically disambiguated, being represented by $K = \{k_1, k_2, \ldots, k_n\}$,where k_i $(1 \le i \le n)$ is the i-th word after semantic disambiguation. Each step based on IFD's semantic disambiguation can be written as

$$K(k_1, k_2, \ ..., \ k_n) = \begin{cases} k_i, w_i \text{ can be described by } d_j; \\ w_i, \text{otherwise.} \end{cases} \quad (1)$$

The above expression means that if the extracted i-th segmented word w_i corresponds to the i-th word d_i in the IFD-based dictionary D, then w_i is replaced by d_j and stored as k_i. Otherwise, the value of w_i will be returned to k_i directly. The main process is shown in Fig. 5

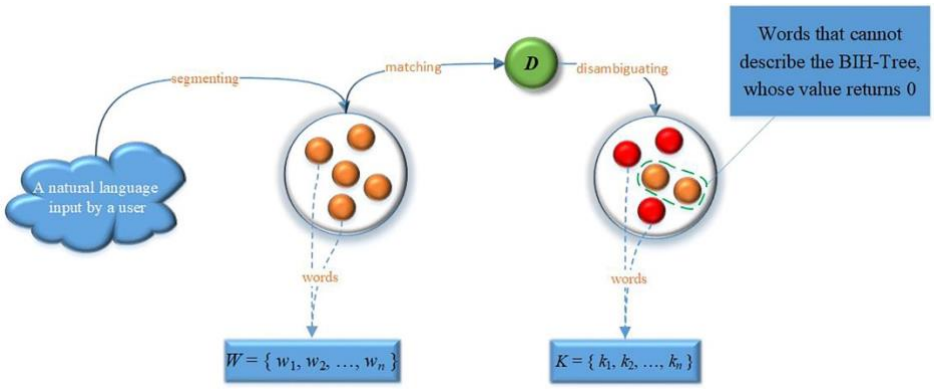


FIG. 5: Word Segmentation and Semantic Disambiguation Process

Algorithm 2 summarizes the whole processes of Word segmentation and semantic disambiguation.

Algorithm 2: Word Segmentation and Semantic Disambiguation.

Input: W, D, K.
Output: K.

$W = \{w_1, w_2, \ldots, w_n\}$.
$D = \{d_1, d_2, \ldots, d_m\}$.
$K = \{\}$.

for w_i **in** D:
  if w_i can be described by D:
    w_i is converted to d_i and stored in K.
  else:
    w_i stored in K.
end for

The flowchart is shown in Fig. 6.

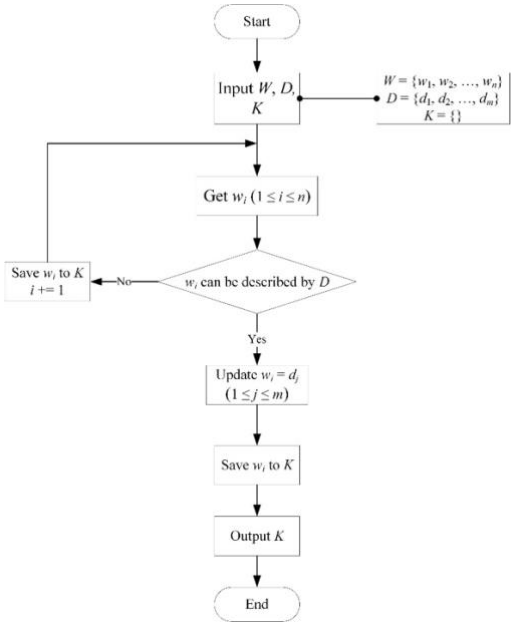


FIG. 6: Flow chart of algorithm 2

4.3 Syntactic analysis

After the work of word segmentation and semantic disambiguation is completed, each word or phrase exists independently. However, individual words or phrases cannot indicate the intent of the user's query. Therefore, it is also necessary to establish a constraint relationship between words or phrases to obtain the user's query intention. This section uses a Stanford Parser for syntactic analysis of sentences.

To illustrate, the natural sentence "图书馆二层的书架的材质（the material of the bookshelf on the second floor of the library）" is taken as a case to illustrate the whole process. Taking the set *K* obtained from word segmentation and semantic disambiguation as the corpus file, the parser uses the “chinesefactored” service in stanford-parser-3.9.2-models.jar to parse the natural language input, and to generate the syntax structure tree shown in Fig. 7. Each word or phrase has a different part of speech. It is marked with abbreviations, such as NN (noun), NP (noun phrase), DNP (phrase composed of '的' to represent the relationship of ownership), CD (cardinal number), etc. After obtaining the syntax tree, the constraint relationship between query words needs to be determined. Since DNP is a phrase composed of 'de' to indicate the affiliation, DNP structure is used to determine the affiliation between words or phrases, as shown in Fig. 7. For example, "二层的书架（Bookshelf on the second floor）", in which "二层" and "书架" form a pair of subordinate relations.

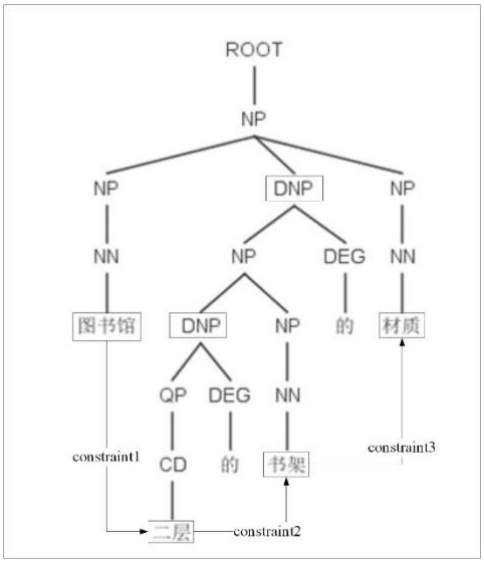


FIG. 7: Syntax tree

In addition to DNP, there are also LC (locators), CP (phrases composed of '的' to indicate modifying relationships), and IN (prepositions or subordinate conjunctions) to mine the affiliation between words or phrases.

5. BUILDING INFORMATION RETRIEVAL

This section develops a novel solution to find the building components or attribute associating with the unified queries.

5.1 Matrix Definition of Information Retrieval

According to the BIH-Tree proposed in the third section, this section assumes that the number of building nodes is n, and each building node contains n different attribute nodes. Then the BIH-Tree is composed of n building nodes and $n \times n$ attribute nodes. The combination of building nodes and attribute nodes from root node to level i+1 attribute node is defined as retrieval path P, as shown in Fig. 8. And this scheme assumes that the node number of level i +1 is N, then the number of retrieval paths is N. The retrieval paths are represented by $P = \{p_1, p_2, \ldots, p_N\}$, where p_i is the i-th retrieval path. Then the nodes on the BIH-Tree are listed in the order of building nodes and their attribute nodes. For example, the first path shown in Fig. 8 can be written as $p_1 = \{K_{11}, m_1, m_2, \ldots, m_n, K_{21}, m_1, m_2, \ldots, m_n, K_{31}, m_1\}$.

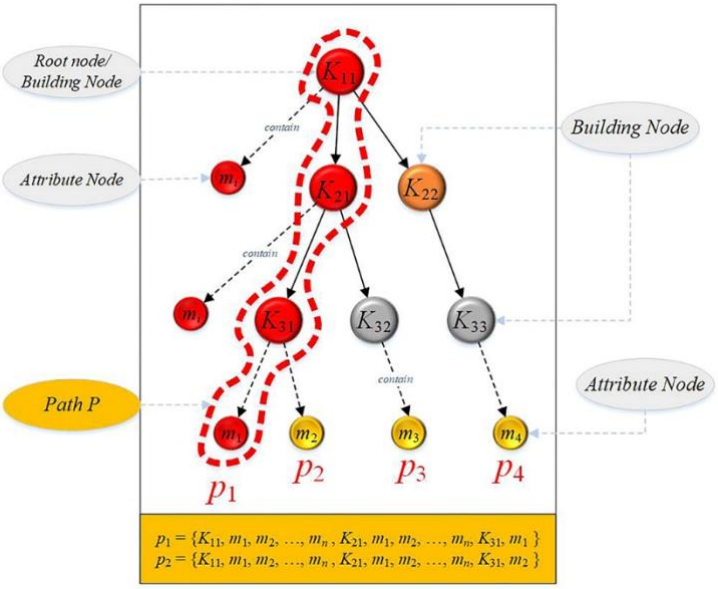


FIG. 8: Retrieval Path P

Therefore, nodes in the BIH-Tree are listed in the order of building nodes and their attribute nodes, which are represented by $A^o = \{(K_{11}, m_1, m_2, \ldots, m_n), (K_{21}, m_1, m_2, \ldots, m_n), \ldots, (K_{in}, m_1, m_2, \ldots, m_n)\}$. The number of nodes in A^o is $n+n\times n$, which is $n\times(n+1)$. Then we take N retrieval paths as the rows of the matrix, and the building nodes and attribute nodes included in the BIH-Tree are taken as columns of the matrix. When the K_{ij} and m_i elements in the retrieval path p_i belongs to A^o, assign 1, otherwise assign 0. The following relations are obtained:

$$A = \begin{cases} 1, K_{ij}, m_i \text{ belong to } A^o; \\ 0, \text{otherwise.} \end{cases} \tag{2}$$

According to the relation (2), the BIH-Tree feature matrix A can be obtained. The BIH-Tree feature matrix A of N rows and $n\times(n+1)$ columns is shown below.

$$A = \begin{matrix} \vdots \\ p_N \end{matrix} \begin{bmatrix} \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{N\times[n\times(n+1)]} \tag{3}$$

The subordination of each word or phrase in a query statement has been identified by the subordination in the syntax tree. The key words in the set K are arranged in the order of principal and subordinate to ensure the accuracy of the analysis. The specific forms are as follows: 1) the principal represents the words or phrases of the building node, and the subordinate represents the attribute node. 2) The principal represents the words or phrases of the building node, and the subordinate represents the child building node. 3) The principal is the word or phrase representing the attribute node, then the subordinate is the attribute value. If the element in K is found in A^o, the value is assigned 1, otherwise the value is assigned 0. Retrieval matrix B is obtained from the following relationships.

$$B = \begin{cases} 1, K_i \text{ belong to } A^o; \\ 0, \text{otherwise.} \end{cases} \tag{4}$$

According to the relation (4), the retrieval matrix B can be obtained. Each row of matrix B represents the parsed and unified words that can be found in set A^o. Retrieval matrix B of $n\times(n+1)$ rows and one column is shown below.

$$B = \begin{matrix} \vdots \\ m_n \end{matrix} \begin{bmatrix} \vdots \\ 0 \end{bmatrix}_{[n\times(n+1)]\times 1} \tag{5}$$

5.2 Retrieval Process

According to the above method, the BIH-Tree feature matrix A and the retrieval matrix B are obtained, then the following retrieval scheme is implemented.

$$\begin{matrix} \vdots \\ p_N \end{matrix} \begin{bmatrix} \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{N\times[n\times(n+1)]} \times \begin{matrix} \vdots \\ m_n \end{matrix} \begin{bmatrix} \vdots \\ 0 \end{bmatrix}_{[n\times(n+1)]\times 1} = \begin{bmatrix} \vdots \\ p_N^o \end{bmatrix}_{N\times 1} \tag{6}$$

A B C

In this scheme, the BIH-Tree feature matrix A and retrieval matrix B are multiplied to get a matrix of N rows and one column, which is represented by $C=[p_1{}^o, p_2{}^o, ..., p_N{}^o]^T$. Each row of matrix C represents the algebraic sum of nodes that can be retrieved in each retrieval path. Then the maximum number in C is the query result, wherefore the elements in matrix C are sorted from high to low. C^o represents a new matrix, in which the first element is the retrieval result.

Algorithm 3 summarizes the whole processes of building information retrieval.

Algorithm 3: building information retrieval

Input: A^o,P, K
Output: C^o
for p_i **in** A^o:
  if K_{ij}, $m_i \in p_i$ **and** K_{ij}, $m_i \in A^o$**:**
    $K_{ij} = 1.$
    $m_i = 1.$
  else:
    $K_{ij} = 0.$
    $m_i = 0.$
The BIH-Tree featurematrix A is obtained.
$k_i \in K$
for K **in** A^o:
  if $k_i \in K$ **and** $k_i \in A^o$:
    $K_{ij} = 1.$
    $m_i = 1.$
  else:
    $K_{ij} = 0.$
    $m_i = 0.$
The retrieval matrix B is obtained.
$A*B = C$
Sort the elements in matrix C from high to low to
get the retrieval result matrix C^o.
C^o = **sort(**C**)**

The flowchart is shown in Fig. 9.

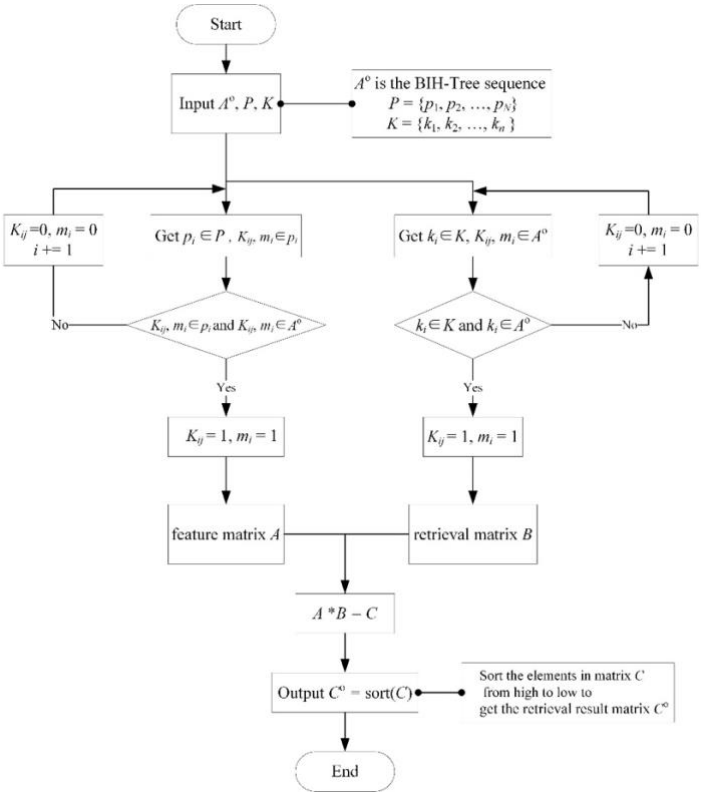


FIG. 9: Flow chart of algorithm 3

6. EXPERIMENTAL

This section takes the library of 'xxxx' University as an experimental model (Model size 1.22GB), investigates the retrieval process of building information in BIM with the new information retrieval scheme proposed in this study. The feasibility of multi-scale information retrieval for BIM using hierarchical structure modeling and Natural Language Processing is verified.

The library is a square building of 32 meters high, 70 meters long and 70 meters wide, with a total building area of 38580 m^2, costing 320 million RMB. Each layer of the seven floors above ground and one underground is divided into four regions, namely A, B, C and D.

- The hierarchical relationship of building information is constructed and the building information model is established according to IFC standard. The specific steps are shown below and in Fig. 10.
- Using Autodesk Revit 2016 to build BIM model of the library. The model includes building information in the fields of building structure, water supply and drainage, heating and ventilation, electricity etc.
- Each floor of the building is divided into four areas, namely Area A, Area B, Area C and Area D, according to the structural characteristics of the building and IFC standards. Several buildings are assigned to these areas, as shown in Fig. 10.
- The BIH-Tree is established according to the hierarchical relationship of different buildings in the same system and IFC standards. The BIH-Tree is imported into each building unit and a hierarchical building information model is designed. For instance, the bookshelf belongs to the B area of the second floor of the library, then the BIH-Tree containing the bookshelf is established and imported into the attribute information of the bookshelf to complete the design of the model.
- BIH-Tree data that represent the hierarchical relationship in the retrieval system are imported to realize the function of finding building components associated with unified query.

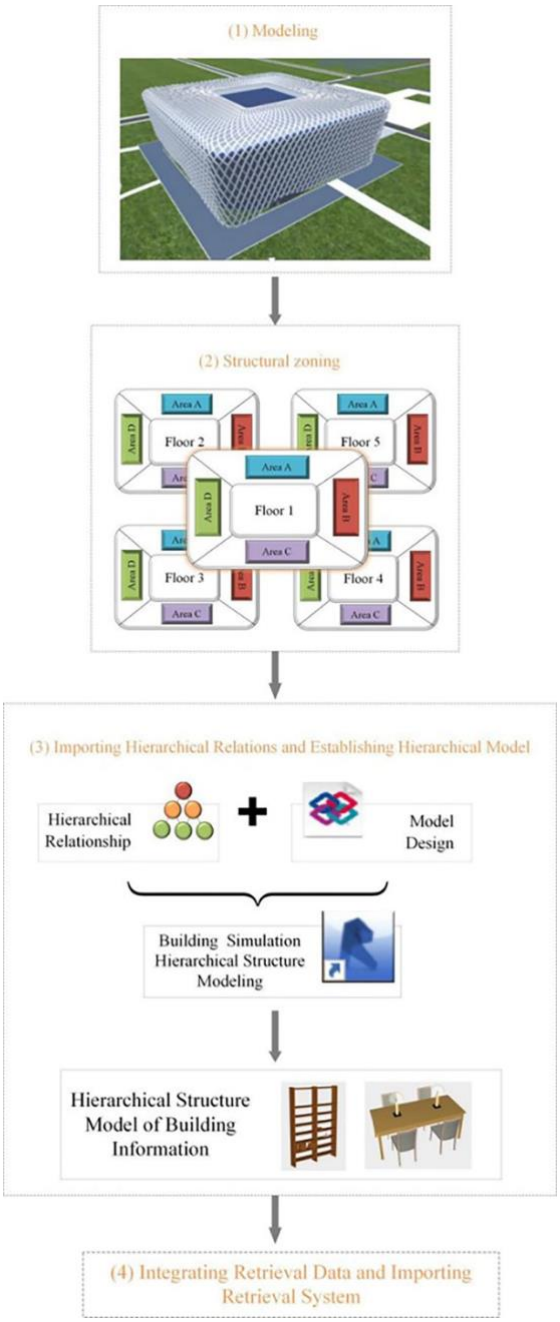


FIG. 10: Modeling Process

In this study, a search engine for retrieving multi-scale information in BIM model is developed by combining BIH tree and natural language processing technology. In this system, the multi-scale information can be queried. By analyzing the real intention of the query statements, the information of building components meeting the needs of users can be returned. The system page is shown in Fig. 11.

The left side of the page is a query box, the right side shows the query results, and the BIM models are visualized at the top of the page. The performance of this system is compared with that of keyword based BIM retrieval system, and the same multi-scale query statement "额定工作温度为 57 度并且响应为快速的直立型喷头" (vertical sprinkler with rated operating temperature of 57 and fast response) is compared in similar keyword based BIM retrieval systems (As shown in Fig. 12). The retrieval system proposed in this study can query multi-scale building information, and can meet the actual needs of practitioners.

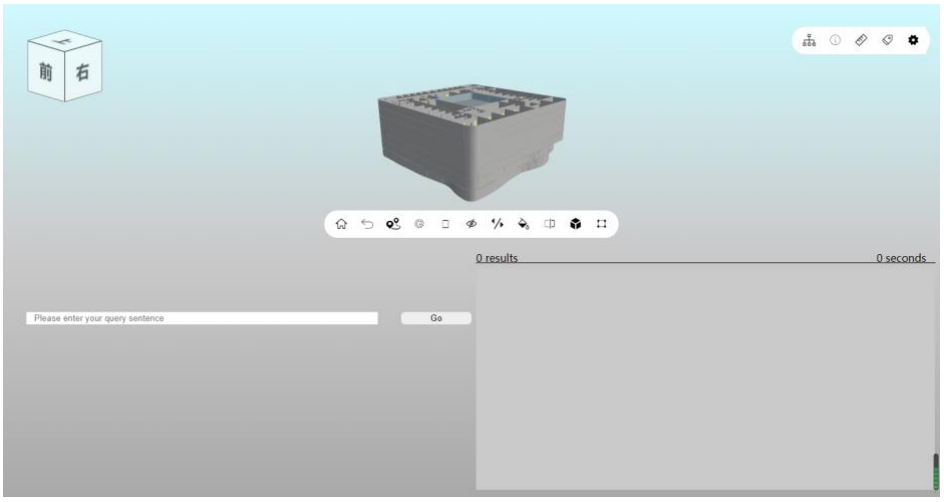
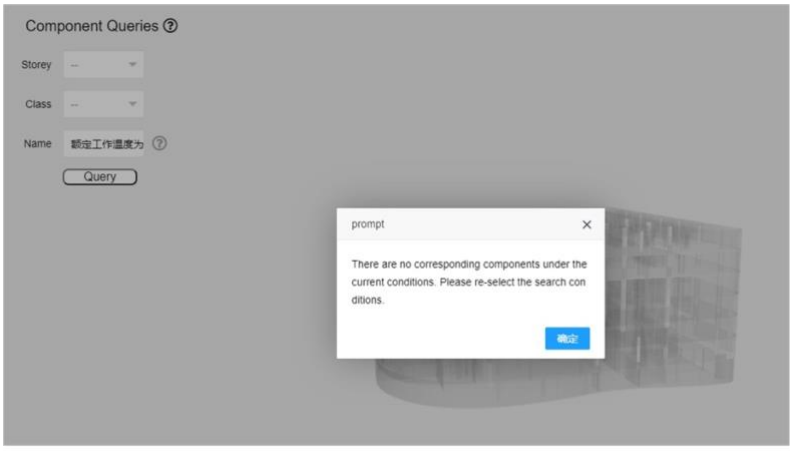
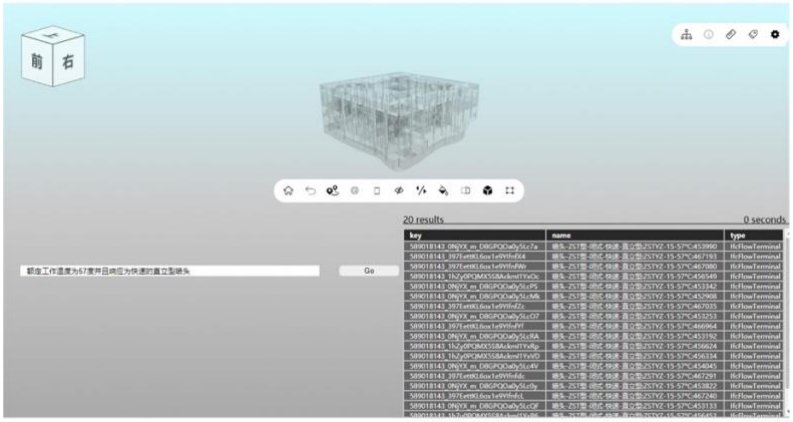


FIG. 11: BIM multi scale information retrieval system



a. Result of BIM system based on keywords for search
sentence "额定工作温度为57度并且响应为快速的直立型喷头"



b. Results in our search engine for search
sentence "额定工作温度为57度并且响应为快速的直立型喷头"

FIG. 12: a. Result of BIM system based on keywords for search sentence"额定工作温度为57度并且响应为快速的直立型喷头". b. Results in our search engine for search sentence"额定工作温度为57度并且响应为快速的直立型喷头"

7. CONCLUSIONS

This study proposed a multi-scale building information retrieval scheme from BIM models using NLP. Firstly, the BIH-Tree representing hierarchical relationship of building information through the IFC standard was presented. Secondly, a novel scheme to unify the queries was developed jointly using NLP and IFD. Thirdly, the BIH-Tree and the queries were transformed into matrix form. Further, a path retrieval algorithm was developed, which can match the parent and child nodes on the BIH-Tree to find the path that meets the retrieval conditions. Finally, through a real engineering application case, the accurate query of multi-scale information such as space, equipment and management data in BIM data is realized. Using IFC Standard to analyze the multi-scale information in BIM model, the multi-scale information of space, equipment and management data in BIM model is transformed into the form of hierarchical tree (BIH-Tree), which can help the retrieval system to better organize, manage and understand the data information in BIM model. This retrieval method can effectively solve the problem that the current retrieval system can not deal with complex query statements, and can ensure that the query results are more in line with the needs of users. Moreover, the proposed multi-scale information retrieval method in BIM model can be applied to the operation and maintenance management system based on BIM, for example, in the field of fire protection, to help managers quickly and accurately obtain fire alarm information and improve the rescue efficiency. In building compliance review, this method helps to obtain multi-scale building information for review in BIM model, and ensures the speed and accuracy of review. In the process of building maintenance inspection, the retrieval method can help staff quickly and accurately find the equipment that needs to be repaired and replaced (such as: installed on the third floor, double tube fluorescent lamp produced by XXX manufacturer), reduce the cost of information screening, and improve the level of data resource utilization and information service ability.

In our ongoing work, the applications of the proposed method will be expanded further. Each BIM model is a large knowledge base. Thus our future work will focus on how to link the knowledge in BIM with the knowledge in AEC according to building topology Ontology. However, our method still has some limitations: 1) The BIM data in our retrieval system cannot completely cover the building information in the entire AEC domain. In addition to the semantics of the IFC data model and IFD dictionary, each item should have its own semantic link to the topological organization. 2) The dependency rules are defined to determine the constraint relationship between words and phrases. More flexible rules should be combined to achieve more accurate query sequence. 3) The relationship in the BIH-Tree is a parent-child relationship, and more relationships should be added in the future work.

REFERNCES

Abualdenien, J., Schneider-Marin, P., Zahedi, A., Harter, H., Exner, H., Steiner, D., ... & König, M. (2020). Consistent management and evaluation of building models in the early design stages. ITcon, 25, 212-232.

Bell, H., Bjørkhaug, L., Bjaaland, A., & Grant, R. (2008). IFD Library White Paper. Available at: www. ifd-library. org/images/IFD_Library_White_Paper_2008-04-10_I. pdf (accessed January 2012).

Chen, K., Chen, W., Li, C. T., & Cheng, J. C. (2019). A BIM-based location aware AR collaborative framework for facility maintenance management. ITcon, 24, 360-380.

Chen, Y. C., Lin, B. Y., & Lin, C. H. (2017). Consistent Roof Geometry Encoding for 3D Building Model Retrieval Using Airborne LiDAR Point Clouds. ISPRS International Journal of Geo-Information, 6(9), 269.

De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In Lrec (Vol. 6, pp. 449-454).

Duddy, K., Beazley, S., Drogemuller, R., & Kiegeland, J. (2013). A platform-independent product library for BIM. In Proceedings of the 30th CIB W78 international conference. WQBook.

Fleming, K., Long, N., & Swindler, A. (2012). Building Component Library: an online repository to facilitate building energy model creation (No. NREL/CP-5500-54710). National Renewable Energy Lab.(NREL), Golden, CO (United States).

Gao, G., Liu, Y. S., Lin, P., Wang, M., Gu, M., & Yong, J. H. (2017). BIMTag: Concept-based automatic semantic annotation of online BIM product resources. Advanced Engineering Informatics, 31, 48-61.

Gao, G., Liu, Y. S., Wang, M., Gu, M., & Yong, J. H. (2015). A query expansion method for retrieving online BIM resources based on Industry Foundation Classes. Automation in construction, 56, 14-25.

Ghaffarianhoseini, A.[Ali]., Tookey, J., Ghaffarianhoseini, A.[Amirhosein]., Naismith, N., Azhar, S., Efimova, O., & Raahemifar, K. (2017). Building Information Modelling (BIM) uptake: Clear benefits, understanding its implementation, risks and challenges. Renewable and Sustainable Energy Reviews, 75, 1046-1053.

Gui, N., Wang, C., Qiu, Z., Gui, W., & Deconinck, G. (2019). IFC-Based Partial Data Model Retrieval for Distributed Collaborative Design. Journal of Computing in Civil Engineering, 33(3), 04019016.

Gunay, H. B., Shen, W., & Yang, C. (2019). Text-mining building maintenance work orders for component fault frequency. Building Research & Information, 47(5), 518-533.

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). Natural Language Processing: Python and NLTK. Packt Publishing Ltd.

Jeon, J., Lee, J., & Ham, Y. (2019). Quantifying the impact of building envelope condition on energy use. Building Research & Information, 47(4), 404-420.

Jones, B. I. (2020). A study of Building Information Modeling (BIM) uptake and proposed evaluation framework. Journal of Information Technology in Construction (ITcon), 25(26), 452-468.

Lin, D. (2003). Dependency-based evaluation of MINIPAR. In Treebanks (pp. 317-329). Springer, Dordrecht.

Lin, J. R., Hu, Z. Z., Zhang, J. P., & Yu, F. Q. (2016). A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM. Computer-Aided Civil and Infrastructure Engineering, 31(1), 18-33.

Liu, X., Wang, X., Wright, G., Cheng, J., Li, X., & Liu, R. (2017). A state-of-the-art review on the integration of Building Information Modeling (BIM) and Geographic Information System (GIS). ISPRS International Journal of Geo-Information, 6(2), 53.

Pasini, D., Caffi, V., Daniotti, B., Spagnolo, S. L., & Pavan, A. (2017). The INNOVance BIM library approach. Innovative Infrastructure Solutions, 2(1), 15.

Preidel, C., & Borrmann, A. (2015). Automated code compliance checking based on a visual language and building information modeling. In ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction (Vol. 32, p. 1). IAARC Publications.

Preidel, C., Daum, S., & Borrmann, A. (2017). Data retrieval from building information models based on visual programming. Visualization in Engineering, 5(1), 18.

Shi, X., Liu, Y. S., Gao, G., Gu, M., & Li, H. (2018). IFCdiff: A content-based automatic comparison approach for IFC files. Automation in Construction, 86, 53-68.

Tang, S., Shelden, D. R., Eastman, C. M., Pishdad-Bozorgi, P., & Gao, X. (2019). A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. Automation in Construction, 101, 127-139.

Wei, G., Zhou, Z., Zhao, X., & Ying, Y. (2010, April). Design of building component library based on IFC and PLIB standard. In 2010 2nd International Conference on Computer Engineering and Technology (Vol. 4, pp. V4-529). IEEE.

Wu, S., Shen, Q., Deng, Y., & Cheng, J. (2019). Natural-language-based intelligent retrieval engine for BIM object database. Computers in Industry, 108, 73-88.

Xie, Q., Zhou, X., Wang, J., Gao, X., Chen, X., & Liu, C. (2019). Matching Real-World Facilities to Building Information Modeling Data Using Natural Language Processing. IEEE Access, 7, 119465-119475.

Yalcinkaya, M., & Singh, V. (2015). Patterns and trends in building information modeling (BIM) research: A latent semantic analysis. Automation in Construction, 59, 68-80.

Zhang, J. (2017). A logic-based representation and tree-based visualization method for building regulatory requirements. Visualization in Engineering, 5(1), 2.

Zhang, J., & El-Gohary, N. M. (2015a). Automated extraction of information from building information models into a semantic logic-based representation. In Computing in Civil Engineering 2015 (pp. 173-180).

Zhang, J., & El-Gohary, N. M. (2015b). Automated information transformation for automated regulatory compliance checking in construction. Journal of Computing in Civil Engineering, 29(4), B4015001.

Zhang, J., & El-Gohary, N. M. (2016). Extending building information models semiautomatically using semantic natural language processing techniques. Journal of Computing in Civil Engineering, 30(5), C4016004.

Zhou, X., Wang, J., Guo, M., & Gao, Z. (2018). Cross-platform online visualization system for open BIM based on WebGL. Multimedia Tools and Applications, 1-16.

Zhou, X., Xie, Q., Guo, M., Zhao, J., & Wang, J. (2020). Accurate and Efficient Indoor Pathfinding Based on Building Information Modelling Data. IEEE Transactions on Industrial Informatics.

Zhou, X., Zhao, J., Wang, J., Su, D., Zhang, H., Guo, M., ... & Li, Z. (2019). OutDet: an algorithm for extracting the outer surfaces of building information models for integration with geographic information systems. International Journal of Geographical Information Science, 33(7), 1444-1470.