Automation in Construction 56 (2015) 14–25

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

A query expansion method for retrieving online BIM resources based on Industry Foundation Classes

Ge Gao [a,d], Yu-Shen Liu [a,b,c,*], Meng Wang [a], Ming Gu [a], Jun-Hai Yong [a]

[a] BIM Research Group, School of Software, Tsinghua University, Beijing, China
[b] Key Laboratory for Information System Security, Ministry of Education of China, China
[c] Tsinghua National Laboratory for Information Science and Technology, China
[d] Department of Computer Science and Technology, Tsinghua University, China

ARTICLE INFO

Article history:
Received 2 July 2014
Received in revised form 1 April 2015
Accepted 14 April 2015
Available online xxxx

Keywords:
Building Information Modeling (BIM)
Information retrieval
Industry Foundation Classes (IFC)
Ontology
Query expansion
Local context analysis (LCA)

ABSTRACT

With the rapid popularity of Building Information Modeling (BIM) technology, BIM resources such as building product libraries are growing rapidly on the World Wide Web. As a result, this also increases the difficulty for quickly finding useful BIM resources that are sufficiently close to user's specific needs. Keyword-based search methods have been widely used due to their ease of use, but their search accuracy is often not satisfactory because of the semantic ambiguity of terminologies in BIM-specific documents and queries. To address this issue, we develop a prototype semantic search engine, named BIMSeek, for retrieving online BIM resources. The central work consists of two parts as follows. Firstly, based on Industry Foundation Classes (IFC) which is a major standard for BIM, a domain ontology is constructed for encoding BIM-specific knowledge into the search engine. Using the ontology, term`inologies in BIM documents can be disambiguated and indexed. Secondly, by combining the ontology and local context analysis technique, an automatic query expansion method is presented for improving retrieval performance. Compared with traditional keyword-based methods and WordNet-based query expansion methods, the experimental results demonstrate that our method outperforms them. The search engine is available at http://cgcad.thss.tsinghua.edu.cn/liuyushen/ifcqe/.

1. Introduction

Building Information Modeling (BIM) technology has been receiving an increasing attention in the AEC (Architecture, Engineering and Construction) industry [1]. Compared with the traditional Computer Aided Design (CAD) technology, BIM is capable of restoring both geometric and rich semantic information of building models, as well as their relationships, to support lifecycle data sharing. With the rapid popularity of BIM technology in the AEC field, BIM resources such as building product libraries are growing rapidly on the World Wide Web (WWW). For instance, the well-known Autodesk Seek [2] is an online system, which provides a large repository of building products on its website and allows users to find a wide variety of BIM products from manufacturers. It currently carries more than 65,000 commercial and residential building products from nearly 1000 manufacturers, and is still growing daily. BIMobject [3] is another widely visited website, which contains over 450,000 BIM models with the product data and properties. Other online libraries such as National BIM Library [4], Google 3D Warehouse [5], SmartBIM [6] and many active online communities (e.g., RevitCity [7]) also contain a large number of information contents of building products related to BIM models.

However, the large amount of online BIM resources also increases the difficulty for quickly finding useful information that are sufficiently close to user's specific needs. In the engineering field, some studies have reported that engineers spent a large amount of time in searching for information [8–10]. In the AEC field, the percentage of time spending on search might be higher. To search online BIM documents quickly and accurately, existing information retrieval approaches should be appropriately adopted for possible improvement. The current practice in information retrieval (IR) mostly relies on keyword-based search methods, which can provide an easy way to quickly retrieve documents. However, search accuracy of traditional keyword-based retrieval models, such as Boolean model, vector space model, or probabilistic model, has been often problematic because of the semantic ambiguity of terminologies in BIM documents and queries. The semantic ambiguity in BIM documents can be alleviated by using a domain ontology. Meanwhile, the user's query can be expanded with domain-specific terms to improve the accuracy of the search result.

This paper aims to study the particular problem for retrieving online BIM documents. To achieve this purpose, we develop a prototype semantic search engine, named BIMSeek, by combining the usability of keyword-based interface with automatic query

* Corresponding author at: School of Software, Tsinghua University, Beijing 100084, China. Tel.: +86 10 6279 5455; mobile: +86 159 1083 1178.
E-mail addresses: gg07@mails.tsinghua.edu.cn (G. Gao), liuyushen@tsinghua.edu.cn (Y.-S. Liu), wm0409@gmail.com (M. Wang), guming@tsinghua.edu.cn (M. Gu), yongjh@tsinghua.edu.cn (J.-H. Yong).
URL: http://cgcad.thss.tsinghua.edu.cn/liuyushen/ (Y.-S. Liu).

http://dx.doi.org/10.1016/j.autcon.2015.04.006
0926-5805/

expansion techniques. The central work consists of two parts as follows. Firstly, based on Industry Foundation Classes (IFC) [11] which is a major standard for BIM, a domain ontology named IFC IR Ontology is constructed for encoding BIM-specific knowledge into the search engine. Using the ontology, terminologies in BIM documents can be disambiguated and indexed. Secondly, by combining the domain ontology and local context analysis (LCA) technique [12,13], an automatic query expansion method is presented for improving retrieval performance. Here, LCA is used to select query expansion terms based on co-occurrence with the query terms within the top-ranked documents. Compared with traditional keyword-based methods and WordNet-based query expansion methods, the experimental results demonstrate that our method outperforms them.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 gives the objective of our study. Section 4 introduces the construction process of IFC-based ontology. Section 5 describes our query expansion algorithm based on the ontology. Section 6 illustrates our search engine and demonstrates the experimental results. Finally, Section 7 concludes the paper and discusses the limitations and future work.

2. Related work

Most traditional keyword-based search approaches are based on the vector space model (VSM) [14]. In this model, documents and queries are represented as a vector of term weights and the retrieval is done according to the similarity between these vectors. In essence, such approaches try to derive the meaning of the text from the observable syntactic and statistical behaviors without representing the meaning directly. However, these approaches have the ambiguity problem for retrieving online BIM documents. The main reason for the problem is that BIM documents are different from general documents because of their syntax variations and semantic complexities. For instance, online BIM documents are often organized by different information providers, and it is not easy to formulate the well-designed queries for the end-users. As a result, some documents, which do not contain the query terms, may not be returned to the user even though they are semantically relevant to the given query.

To handle such problem, one possible way is to use semantic-based IR approaches, in which the search does not completely rely on exact term matching. Instead, such approaches usually seek to improve retrieval accuracy by attempting to encode the user's searching intent and the contextual meaning of terms in documents. Semantic-based IR approaches could be roughly categorized as either explicit or implicit, according to whether an explicit concept space is defined and utilized. An implicit technique aims to analyze and explore the hidden relationships between a set of documents and the terms, without requiring the concepts defined explicitly. Existing implicit approaches include local context analysis (LCA) [12,13], latent semantic analysis (LSA) [15], probabilistic LSA (pLSA) [16], latent Dirichlet allocation (LDA) [17], etc. In contrary, an explicit technique tries to extract the explicit semantic definitions between words and terms as well as the relationships between them, such as synonym, hyponym and hypernym. These definitions and relationships are usually represented in some form of linked data such as taxonomy, thesaurus or ontology. Existing explicit approaches such as ontology-based query expansion [18], semantic indexing [19], explicit semantic analysis (ESA) [20] and granularity analysis [21] have been developed for improving the semantic-based retrieval.

In the last few years, some semantic-based IR techniques have been devoted to the study of information retrieval systems for the engineering field [9,10,22–27]. Many of them dealt with semantic representation or domain ontology construction for applying to various engineering fields. Li et al. [10] developed an engineering ontology in mechanical design and manufacturing, and applied the ontology to index and retrieve unstructured engineering documents and CAD drawings. Lin et al. [23] introduced some text-based information retrieval experiments in the AEC industry, and described an AEC online product search engine by incorporating domain knowledge and IR techniques [24]. Rezgui [27] used domain knowledge to formulate an ontology that assists in indexing and retrieving construction documents in e-COGNOS system. Weissman et al. [26] proposed a computational framework for authoring and searching product design specification documents using semantic mapping. Demian and Balatsoukas [22] investigated the effects of search result interfaces (particularly the aspects of granularity and context) in systems for searching archives of construction documents. Lin et al. [25] presented a passage partitioning approach according to domain ontology, and applied the approach to a concept-based IR system for engineering domain-specific technical documents. More recently, Hahm et al. [9] introduced a personalized query expansion approach for engineering document retrieval based on a self-constructed engineering ontology.

Among the IR techniques in the engineering field, it is noteworthy that query expansion approaches [18] which aim to overcome the ambiguity of natural language and also the difficulty in using a single term to represent an abstract concept. It is generally conducted by supplementing original queried terms by morphological variations or semantically related terms, so the performance degradation caused by syntax variation and semantic complexity of BIM documents can be overcome with query expansion. Query expansion can be classified as interactive, manual, or automatic approaches according to their strategies [18], where automatic query expansion is an alternative strategy for domain-specific retrieval by employing taxonomy (e.g., WordNet) or domain-specific ontology. Considering the characteristics of BIM documents, automatic query expansion based on BIM-specific ontology is an appropriate strategy because ambiguous and complicated queried keywords can be disambiguated and interpreted by the ontology. Also, terminologies in BIM documents can be disambiguated and indexed with the ontology. As a result, the user's short queries can be expanded and matched to syntax varied and semantically complicated documents. By combining query expansion with domain ontology, there have been several search engines on information retrieval in the engineering domain [9,10,24,25,27]. However, they are still limited to their ontology sources and the needs of particular applications as follows. Firstly, most of existing studies focus on retrieving the controlled collection of engineering documents, rather than Web-centric BIM resources. Secondly, existing methods are mainly based on explicit semantic-based IR techniques, which are limited to the capability of hand-crafted ontology or thesaurus and cannot fully fulfill the retrieval task for dynamically changed Web-based BIM resources. Thirdly, ontologies of existing IR approaches lack the utilization of BIM-specific domain knowledge for online document retrieval.

As the commonly used data exchange standard for BIM, Industry Foundation Classes (IFC) [11] led by the buildingSMART, formerly known as International Alliance for Interoperability (IAI), plays a crucial role to facilitate interoperability between various software platforms in the AEC industry. To date, the IFC standard has been widely supported by the market-leading BIM software vendors. As the most widely used taxonomy and specification in BIM applications, the underlying IFC specification is therefore our preferred candidate semantic resource, which provides a sharable skeleton on which the BIM-oriented IR ontology can be built. Recently, several IFC-based ontologies have been studied for particular application needs [28–32]. For instance, Pauwels et al. [31] applied an IFC ontology to semantic rule checking. Beetz et al. [29] presented an approach for converting the IFC schema into the OWL format, which is a remarkable effort to lift the IFC specification onto the ontology level. Zhang and Issa [28] used an IFC ontology to extract partial model from a complete IFC model. In addition, several applications also used IFC ontologies for querying spatial information within a BIM model [32,33].

The common strategies mentioned in previous studies [28–31] are to utilize their respective ontologies for extracting specific information from an IFC file itself, whereas the currently available IFC files are relatively few on the WWW. In practice, a large number of available online BIM resources consist of BIM components/models (in their native file formats but without providing IFC files), and these models are associated with their online product documents (e.g., descriptions or specifications). Therefore, merely parsing and retrieving IFC files cannot take full advantage of the abundant online BIM resources. In contrast, this paper develops a BIM-specific IR ontology based on the IFC schema, and applies the ontology to retrieve online BIM documents rather than IFC files. To alleviate the insufficiency of hand-crafted ontology, our method combines the explicit semantic technique (i.e., ontology-based query expansion) with the implicit semantic technique (i.e., local context analysis [13]). Consequently, our search engine can cover more widely online BIM resources for information retrieval.

3. Objective and methodology

Building product webpage on BIM resource libraries (e.g., [2,4,6]) typically contains BIM models associated with product documents (e.g., specifications and descriptions of the objective products), as shown in Fig. 1. The BIM models are normally in their native file format dependent on various BIM software vendors (e.g., *Autodesk Revit*, *Bentley Architecture* and *Graphisoft ArchiCAD*) or in industry-neutral file format (e.g., IFC/ifcXML). The relevant product documents on the Web are the textual contents for describing BIM models including their functions, dimensions, materials, performances, manufacturers, etc. These product documents are independent of BIM model file format. In particular, much information is embedded in textual BIM documents generated during design and construction phases [22]. Most online BIM documents are unstructured, in contrast to structured contents (e.g., BIM models or database tables) following a strict schema.

In order to retrieve online BIM resources, we have developed a prototype search engine, i.e., BIMSeek, which contains two individual modules: *document search module* and *model search module*. The former that will be described in this paper focuses on the problem of retrieving online BIM documents by incorporating the domain ontology and query expansion techniques. The latter that is an ongoing work deals with the problem of searching online BIM models by incorporating the same ontology and structured query techniques. Although the two modules in BIMSeek process different contents (documents or models) of online BIM resources, they are all based on the common ontology, named IFC IR Ontology developed in this paper (see Section 4), for the semantic search engine. In this paper, only the hierarchical and enumeration relationships of the ontology are utilized for retrieving BIM documents. However, in the model search module, the properties and restrictions of the same ontology are also utilized for searching BIM models, where BIM models are converted into OWL (Ontology Web Language) instances and the user queries are transformed into SPARQL. Also, it is interesting to make use of document and model information together for improving search accuracy and performance in BIMSeek. In this

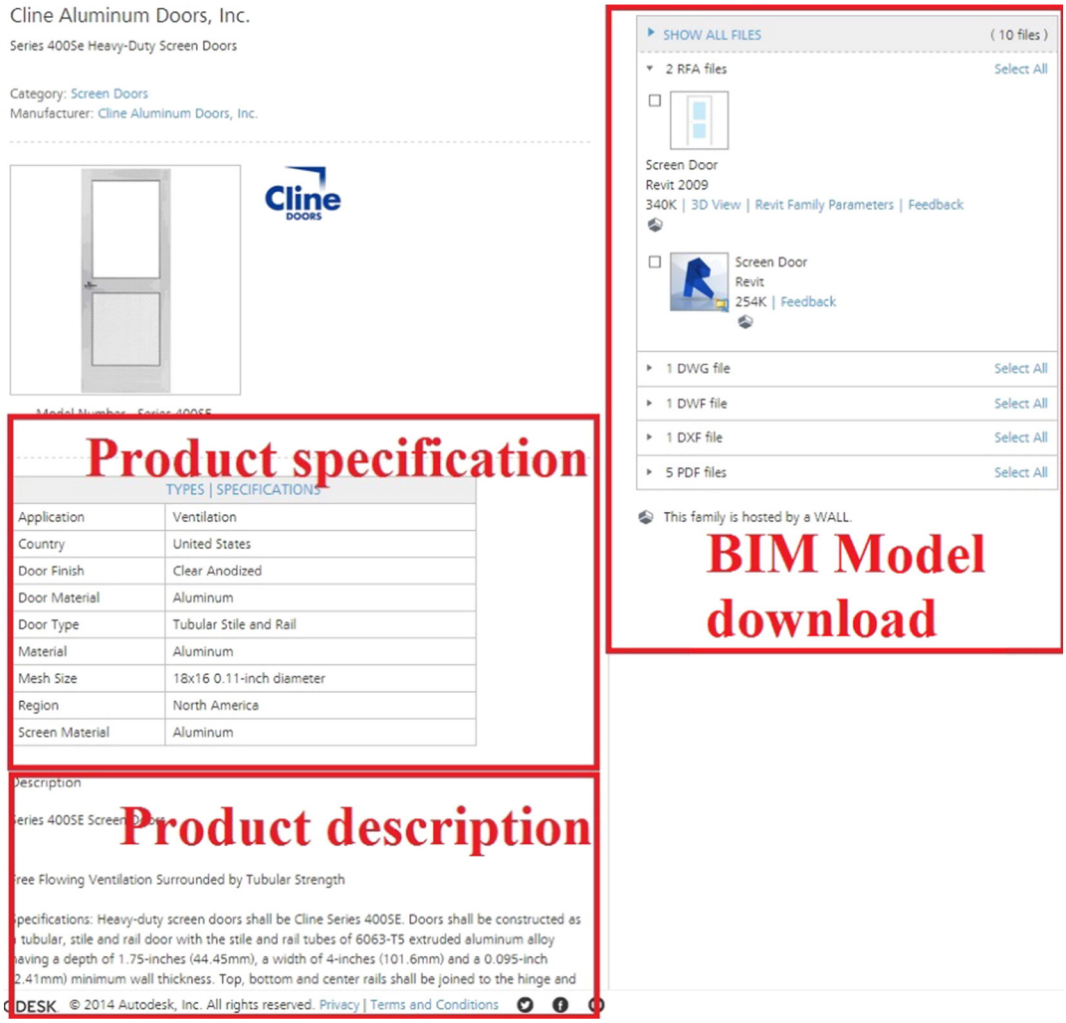


Fig. 1. Illustration of building product webpages that contain BIM models associated with product documents (including product specification and description).

way, more relevant BIM resources might be retrieved for the user, and we leave this study to our future work. In addition, the IFC ontology built in this paper is also used for another module, named BIMTag [34], as part of BIMSeek. Using the ontology, BIMTag achieved semantic annotation for online BIM documents.

The purpose of this study is to use the ontology to process the contextual meaning of terms for retrieving online BIM documents. Based on the ontology, we typically choose automatic query expansion techniques to enhance retrieval performance by reflecting the user's needs. Since automatic query expansion techniques require no effort on the part of the user, they have a significant advantage over manual techniques. Automatic query expansion can be categorized as either global or local [13]. Global techniques rely on analysis of the whole corpus for the source of expansion terms, while local techniques process a small number of top-ranked documents retrieved for a query to expand that query. In information retrieval, query expansion is the first step of semantic search [18,19], which reformulates a given query to increase the likelihood of the term overlap between the query and documents that are likely to be relevant to the user's needs. If used together with an effective word sense disambiguation (WSD) algorithm, query expansion can improve retrieval performance. However, simply using a thesaurus or ontology for automatic query expansion still has some limitations, which could be overcome through using global statistics, such as the document frequency of the query terms, for selecting expansion terms [13].

On the one hand, query expansion without a WSD algorithm or with a poor WSD may cause degradation in retrieval performance, which is mainly caused by the irrelevant terms added to the query. This limitation can be alleviated by combining query expansion with local context analysis (LCA) technique [12,13], where query expansion terms are selected based on co-occurrence with the query terms within the top-ranked documents. With the combination of LCA, query expansion terms generated from thesaurus or ontology undergo a statistical inspection, and consequently some irrelevant expansion terms won't be used for retrieval. On the other hand, ontology-based query expansion may miss some possible expansion terms, which are not defined in thesaurus or ontology but have statistical dependencies with the user's query in corpus. This limitation can be overcome by computing statistical relevance between terms in the most relevant documents, and then these statistical relevant terms are added to the expansion terms for retrieval.

By combining the domain ontology and LCA technique [12,13], this paper introduces an automatic query expansion method to improve retrieval performance for online BIM documents.

4. Development of IFC IR Ontology

To achieve the semantic search engine for online BIM documents, the principal work relies on two aspects: (1) how to construct a BIM-oriented ontology for the needs of IR and (2) how to utilize IR techniques to retrieve BIM documents based on the ontology. The former is introduced in this section and the latter will be presented in Section 5. This section first summarizes some basic concepts of the domain ontology, then argues the IFC specification [11] as a semantic foundation for our purpose, and finally introduces a method for developing the IFC IR Ontology.

In computer science and information science, an *ontology* is defined as "formal, explicit specification of shared conceptualization" [35]. Ontologies can be roughly divided into two categories: *general ontology* and *domain ontology*. The interest of general ontology is the whole world, while domain ontology focuses on specification of particular domain conceptualization. Although some general ontologies (e.g., WordNet [36]) contain a large number of general concepts, they are not designed for domain-specific retrieval, which may lead to inaccurate description of concepts in the AEC domain. In contrast, domain ontology is a representation of semantics in particular domain, which often consists of a hierarchical description of important concepts precisely defined in the domain, along with description of properties of each concept [37]. Domain ontology is considered as a key element to enhance domain-specific IR [35]. It has a variety of representation forms, among which OWL is selected as the modeling language in this study. With the syntax extension from RDF (Resource Description Framework) — a lightweight representation for data and knowledge, OWL has been proposed by W3C as the ontology language of Semantic Web [38].

There are various methods for constructing domain ontologies, such as TOVE, IDEF5, Skeleton, KACTUS, SEN-SUS, METHONTOLOG and Seven-Step methods. Among them the Seven-Step method [39] is regarded as the mature one. It was developed by the School of Medicine in Stanford University, as well as the most widely used ontology editor Protégé [40]. The Seven-Step method is also the only one which strictly conforms to Gruber's five basic principles of ontology construction [35]. Following the general principles of ontology construction and the Seven-Step method, construction of IFC IR Ontology is essentially a process of conceptualizing and formalizing BIM knowledge from the IFC schema. By following the ontology development process of Seven-Step method, we give the main procedure of constructing IFC IR Ontology as follows.

4.1. Step 1: determining the domain and scope of ontology

The first step of Seven-Step method is to define the domain and scope of the ontology. It is an important step for minimizing the amount of data and concepts to be analyzed, especially for the extent and complexity of BIM-related semantics. During the ontology-design process, it may be adjusted if necessary.

The ontology discussed in this work mainly meets the needs of information retrieval, which enables architects, engineers and other design professionals to find useful BIM resources quickly. Therefore, it includes specific concepts related to AEC product information belonging to the product lifecycle phases (design, manufacturing, assembly, etc.). In addition, since online BIM documents are generally organized by different information providers, which often use natural languages for describing the textual contents of documents, our ontology should also include natural language ontologies for capturing terminologies in documents and queries.

4.2. Step 2: selecting and reusing existing resources

The second step of Seven-Step method is to consider reusing existing semantic resources for our particular domain and task. Several well-known semantic resources have been developed for various AEC applications, and they have the potential to be enhanced as domain ontology for retrieving online BIM documents [27,41]. The most notable efforts [27,41] include: ISO 12006-2, Uniclass, OmniClass, Industrial Foundation Classes (IFC) [11], etc. Among these existing semantic resources, structured taxonomies deserve particular attention. However, since improper taxonomies have the opposite effects leading to confusion and difficulty to retrieve [27], a properly structured taxonomy should be carefully selected to meet our needs.

As a relative new field, ontology research for BIM resources is still rare, which needs more exploration. In this study, the IFC4 specification [11] is typically selected as the backbone of IFC IR Ontology. This paper aims to reuse rich semantic contents of newly developed IFC specification without having to understand the complex technical issues of IFC for information seekers. In addition, many general ontologies such as WordNet, EuroWordNet and Cyc are already available in electronic form, so we typically select WordNet [36] as a common lexical database for the English language. In this paper, WordNet will be combined with domain ontology to process terminologies in BIM documents and queries.

4.3. Step 3: enumerating important terms in the ontology

The third step of Seven-Step method is to enumerate important terms in the ontology. In this work, our goal is to recognize the important terms in IFC specification to establish the ontology structure. Initially, it is important to get a comprehensive list of terms without worrying about overlap between concepts. Therefore, the key point lies on how to obtain a comprehensive list of concepts from the IFC specification.

The IFC4 specification is the latest version of IFC standard, and it enhances the capability of previous IFC specifications in several areas of building elements, building service elements and structural elements and accompanying basic definitions. Currently, it contains 766 entities, 206 groups of enumeration types, 408 groups of property sets, 1691 individual properties, and lots of defined types, select types, quantity sets, functions and rules. We confine important terms to *IfcObject* in IFC specification, excluding the others. Currently, we have manually extracted 248 entities, 140 type enumerations, and 583 enumeration items from the IFC4 schema as the important terms of IFC IR Ontology.

However, the IFC schema was designed for computer instead of human readability, which follows a special formalized naming convention. For example, the names of types, entities, rules and functions in IFC start with the prefix "Ifc" and continue with the English words in Camel Case naming convention (no underscore, first letter in word in upper case). Therefore, they should be refined by the process of segmenting the names of entity, removing the prefix and eliminating redundancy, for IR purpose. This process is semi-automatically done in our work.

4.4. Step 4: defining the classes and class hierarchy

The fourth step is to define the classes and the class hierarchy. There are three possible approaches in developing the class hierarchy [42], including (1) the top-down approach, (2) the bottom-up approach, and (3) a combination development process. None of these three methods is inherently better than any of the others. Considering the consistency of our model, we adopt the first one (i.e., the top-down approach) in our work to avoid the duplication of works and the inconsistency of knowledge.

Since the IFC specification has been defined as a hierarchical structure, its structure can be directly adopted in the class hierarchy of IFC IR Ontology. On the top of the ontology, we use the EXPRESS entity relationship directly, that is to say, we can map directly entities as well as their subtypes and supertypes (profiting from OWL classes inheritance). Afterwards, we treat the element's enumeration type as its subclasses. For example, the concept "covering" has the enumeration type (including "membrane", "cladding", "insulation", "roofing", "modeling", "ceiling", "flooring", "wrapping", etc.) according to the IFC schema. And, these types are connected to their superclass (i.e., "covering") to define the class hierarchy in the ontology.

Table 1 gives the outline of concepts in the resulting IFC IR Ontology and Fig. 2 shows a portion of class hierarchy in the ontology. In Fig. 2, the concept *Covering* inherited from the class *Object* has a property *PredefinedType* mapped to *Covering Type Enumeration*. An *Object* is the generalization of any semantically treated thing or process. In IFC4 specification, *Object* is the supertype of *Actor*, *Control*, *Group*, *Process*, *Product*, and *Resource*, which consists of the first level of IFC IR Ontology (see Table 1). In addition, IFC IR Ontology comprises a variety of relationships, for example "*PredefinedType*", "*OperationType*". There are totally 140 enumeration types related by these relationships.

Table 1
The outline of concepts in the resulting IFC IR Ontology.

Taxonomies	Number of concepts	Example of concepts	Acquisition resources
Actor	2	Occupant	*IfcActor*
Control	11	Project order	*IfcControl*
Group	12	Building system	*IfcGroup*
Process	4	Task	*IfcProcess*
Product	211	Covering	*IfcProduct*
Resource	8	Crew resource	*IfcResource*

4.5. Step 5: define the properties of classes

The hierarchical structure is just a frame for the concept system which cannot reflect the varied relations between the concepts. Therefore, we need to describe and define the properties of classes in this step. These properties become slots attached to the classes. Slots are the descriptors used for describing the properties of classes and instances. We define the domain/range of the properties according to the applicable classes and DataType in the XML specification of the property set (Pset) of IFC 4. The OWL datatype property is used to represent the EXPRESS simple attributes, and the OWL object property represents named attributes. As an example of properties in IFC IR Ontology, the properties of class "door" are defined with "Acoustic Rating", "Thermal Transmittance", "Status", "Self Closing", "Smoke Stop", "Durability Rating", "Has Drive", "Is External", "Glazing Area Fraction", "Hydrothermal Rating", "Infiltration", "Fire Exit", "Security Rating", "Handicap Accessible", "Reference" and "Fire Rating".

4.6. Step 6: defining the facets of the slots

The slots of classes may have different facets to describe the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take. The conversion of EXPRESS simple types (String, Integer, Real, Binary, Boolean, Logical) is direct, as they have equivalents in OWL (XSD types). For instance, String type is directly mapped into xsd:string type. And, the value of "Fire Rating" in IFC (in a common property set of "door") is *IfcLabel*, we directly map it to xsd:string type. As a result, "Fire Rating" becomes a slot of class "door", whose type is string in OWL.

Besides mapping the value type of the slots from EXPRESS to OWL, we also represent the allowed value types with OWL rules, and the allowed number of the values with OWL cardinality restriction according to the EXPRESS attribute's optional flag.

4.7. Step 7: creating the instances

The last step of Seven-Step method is to create individual instances of classes in the hierarchy. However, since our goal in this study is information retrieval of unstructured BIM documents with the help of IFC IR Ontology, we don't have to use the instances of IFC files. Consequently, converting an IFC file to its ontology instance is not necessary for our current application. Of course, it is also interesting to explore information retrieval of IFC instance files, which is the future work in BIMSeek development.

5. Automatic query expansion process

With reference to IFC IR Ontology, this section presents an automatic query expansion method for retrieving online BIM documents. Starting with query words as the input in the search engine, our algorithm mainly consists of four steps: generating the candidate concepts, expanding the candidate concepts, query pruning and keyword matching. The main procedure of our algorithm is shown in Fig. 3. The following will introduce each step in detail.

5.1. Step 1: generate the candidate concepts

The users may not always formulate their search queries using the standardized concepts in IFC specification, so the original query terms are not suitable for direct searching. To execute the semantic search, we should identify the "meaning" of every term in the user's query, so that a matching concept in the IFC IR Ontology can be found.

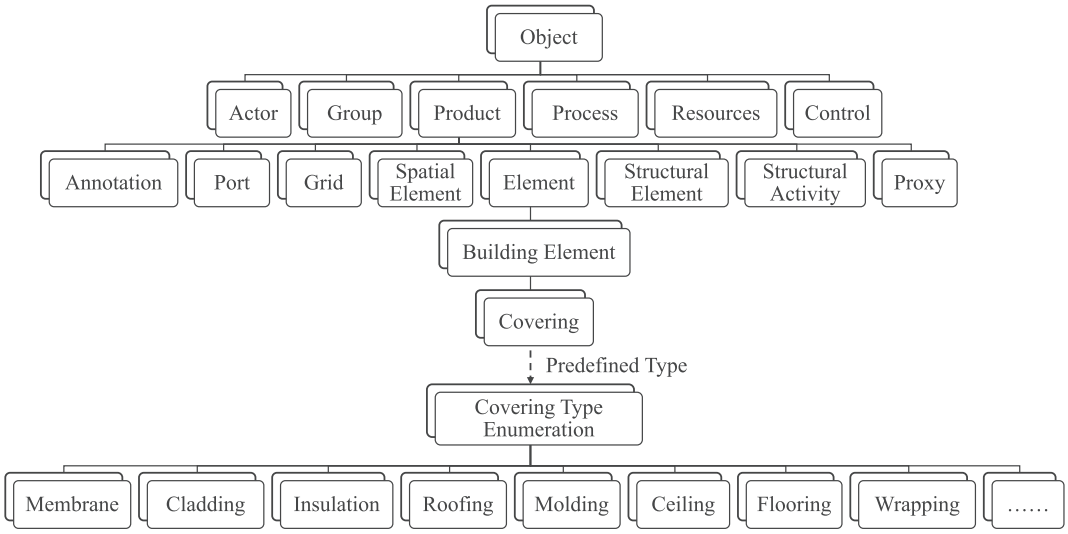


Fig. 2. Part of the classes and class hierarchy in IFC IR Ontology.

This is a necessary step to semantic retrieval, through which semantics of the terms appearing in the user's query will be automatically replaced by the standardized concepts in the ontology.

The first step of our algorithm preprocesses the simple noun phrases in the user's query, and generates the candidate concepts corresponding to IFC IR Ontology. In our implementation, we first use WordNet to find the synonyms for each query term, and then generate the candidate concepts of the query term by matching the domain ontology. WordNet [36] is a large lexical database of English, in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). With reference to a generic lexicon in WordNet, a list of candidate concepts can be produced for further retrieval. For instance, the terms “cover”, “screening”, “masking”, “coating” and “application” are the synonyms of “covering” in WordNet, while only “covering” is the standardized concept in IFC IR ontology. When some synonyms words of “covering” appear to the user's query, the “covering” will be located as a candidate concept.

However, WordNet, as the generic lexicon rather than domain ontology, often contains a large number of concepts that are not relevant to the target domain. As a result, these irrelevant concepts may lead to an inaccurate estimation of concept specificity and information retrieval. For example, “application” should not be regarded as the candidate concept of query “covering” in BIM-specific domain. To address this issue, a context-based method for concept identification can reduce the incorrect mapping, which will be illustrated in Section 5.3.

5.2. Step 2: expand the candidate concepts

Based on WordNet, Step 1 only generates some candidate concepts for the user's query in a level of word-concept matching. In information retrieval, the user-entered query is usually simple and short, and it is hard for a search engine to completely express the user's information needs. In order to represent the information seeker's needs in BIM-specific domain, the candidate concepts can be further expanded by utilizing the relationships between concepts in IFC IR Ontology. The expanded concepts will be added to the query to improve the accuracy and coverage in retrieval.

In the second step of our algorithm, a concept expansion algorithm is proposed based on the relationships between concepts in IFC IR Ontology. The key task lies on how to compute the relatedness between two concepts in the ontology. In our implementation, each candidate concept is expanded through its neighborhood over the ontology to yield more concepts close to the candidate concept semantically. The extent of the neighborhood is governed by the relatedness function, as will be shown in Eq. (1). Especially, each expanded concept in the neighborhood is associated with a relatedness value according to its distance to the candidate concept. There are various relationships (including subclass, superclass, type enumeration etc.) between concepts in IFC specification. Therefore, the concepts, which have the above IFC relationships with each candidate concept, will be added to the new query terms, i.e., expanded concepts. Here, a semantic relatedness function between the candidate concept and its expanded concept is addressed for measuring the expansion range. The expansion process will be terminated when the value of semantic relatedness is less than a predefined threshold.

There are several measures of semantic relatedness, such as Leacock–Chodorow measure [43], Jiang–Conrath measure [44], Lin measure [45], Resnik measure [46] and Wu–Palmer measure [47]. The measures of Lin, Resnik and Jiang–Conrath are based on the notion of information content. In contrast, the Leacock–Chodorow measure is based on path-length, which was found more effective than information content-based relatedness measures [46]. In this section, we choose the Leacock–Chodorow measure [43] for building the semantic relatedness function, where the semantic relatedness between two concepts is estimated based on the conceptual links (i.e., the distance) between these concepts with reference to IFC IR Ontology. The smaller the distance between two concepts, the higher the semantic relatedness between the two concepts will be. Therefore, the semantic relatedness is inversely proportional to the semantic distance in the ontology. Our *semantic relatedness function* is defined by

$$Rel_{IFC}(c, c_i) = -\log\frac{len(c, c_i)}{2Depth}, \tag{1}$$

where c is a query concept in IFC IR Ontology, c_i is the i-th concept related to c in its neighborhood, $Depth$ is the maximal tree depth of the ontology. Especially, $len(c, c_i)$ is the shortest path between c and c_i in IFC IR Ontology, which is calculated as the edge number on the shortest path (in practice, plus 1 to avoid that the value is zero). In Eq. (1), $Rel_{IFC}(c, c_i)$ is the semantic relatedness function which takes into account the shortest path $len(c, c_i)$ between two concepts c and c_i defined in IFC IR Ontology. The factor of 2 in the denominator is used to represent the possible maximum length between two concepts. The distance between two concepts reaches the maximum length when they are both leaf nodes and the only common ancestor of them is the root node. This definition guarantees that in any circumstances, the relevance value between two concepts won't be less than 0.

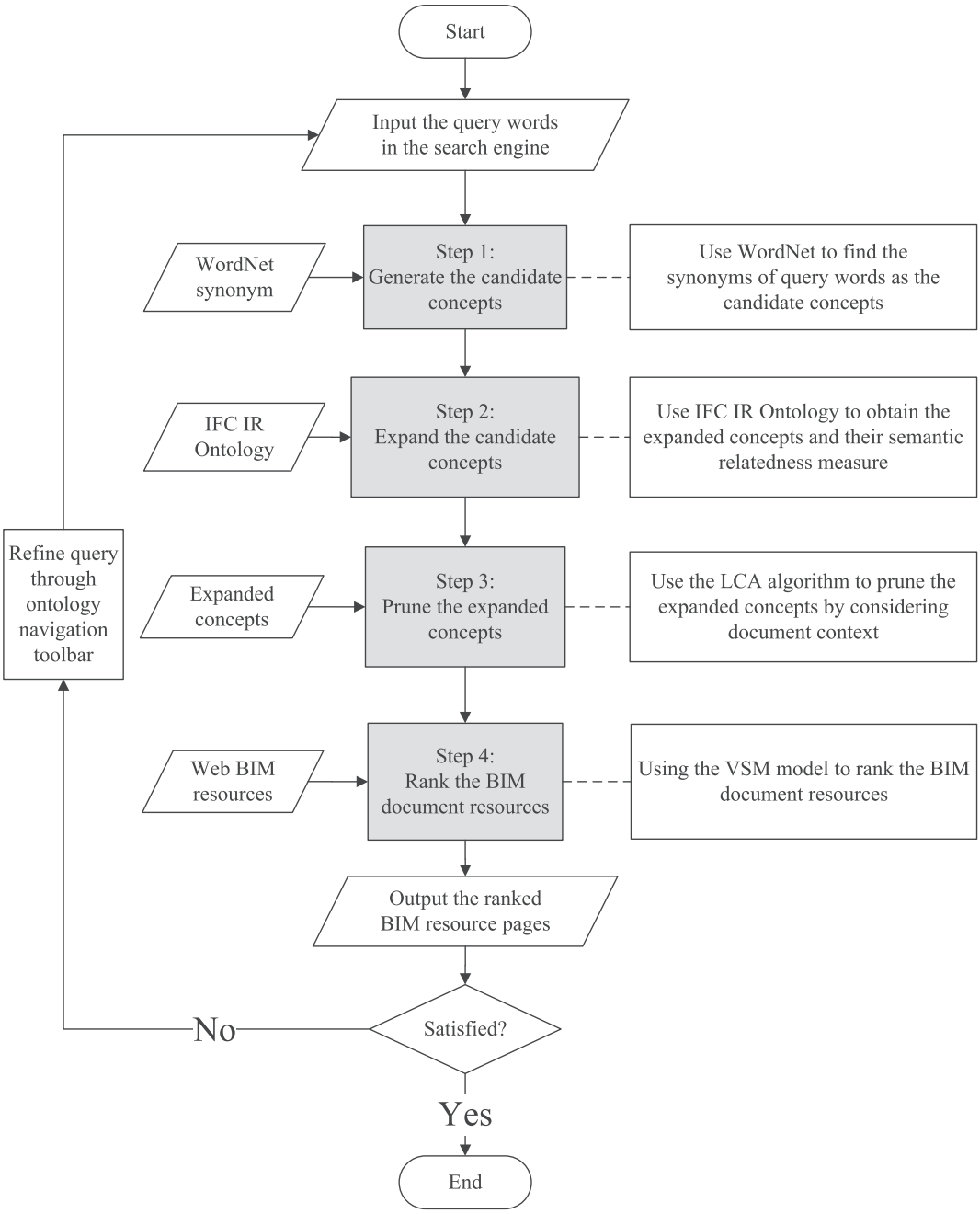


Fig. 3. The main procedure of our algorithm.

As an example in Fig. 2, the concept "covering" can be expanded to some subclasses ("membrane", "cladding", "insulation", "roofing", "modeling", "ceiling", "flooring", "wrapping", etc.) in the 1-level hyponym expansion mode. Considering the concepts "covering" and its subclass "membrane", the *Depth* is equal to 6 and $len(\cdot)$ is equal to 2. Accordingly, the IFC relatedness between "covering" and "membrane" is:

$$Rel_{IFC}\left(\text{“covering”}, \text{“membrane”}\right) = -log\frac{2}{2\times 6} = 0.778.$$

At the end of Step 2, we obtain the expanded concepts and their semantic relatedness values, where all the expanded concepts are saved as a set denoted by *ExpansionSet*.

5.3. Step 3: prune the irrelevant concepts

As discussed in Step 1 and Step 2, the original query terms are mapped into the IFC concepts, and then these concepts are expanded to a set of related concepts to formulate the appropriate domain-specific query. However, simple concept mapping and expansion may have some shortages as follows.

1 The "incorrect mapping" problem. As discussed in Section 5.1, WordNet, as a generic lexicon rather than domain ontology, often contains a large number of synonyms that are irrelevant to BIM-specific domain. As a result, these irrelevant concepts may lead to an inaccurate estimation of concept specificity and information retrieval.
2 The "overexpansion" problem. The expansion of concepts is based on the relationships defined in the IFC specification, which are usually

made by some professionals. These relationships are independent of the document corpus in the domain. In practice, some expanded concepts that are irrelevant to the original query will be harmful to the retrieval, resulting in topic dilution of documents.

To solve the above problems, we use a statistical method to examine the relatedness between concepts. In this step, we utilize local context analysis (LCA) [12,13] for improving IFC-based concept expansion. LCA acquires statistical relatedness between concepts and query terms according to co-occurrence analysis of the concepts with the query terms in top n passages retrieved by the query.

In our implementation, we first use a standard IR system to retrieve the top n ranked passages in the corpus. Then these top-ranked passages are used as the local "context" of the query terms, and the "context" is used to analyze the co-occurrence of the related concepts and the query terms. The concepts, expanded in Section 5.2, are accepted according to a function of the frequencies of occurrence of the query terms and co-occurrence concepts in the retrieved passages, and the inverse passage frequencies for the entire collection of those terms and concepts. Only the concepts that have strong relationship are kept for final retrieval. Since the textual contents of online BIM documents are usually short and each document usually has one single topic rather than multiple topics, the passage in our study is defined as the whole document in a webpage.

Given a user's query Q, we adopt LCA [12,13] to compute the *statistic relatedness* measure between an IFC concept c and Q, as defined by

$$Rel_{LCA}(c, Q) = \prod_{t_{i \in Q}} (\delta + co(c, t_i)), \tag{2}$$

where $co(c, t_i)$ denotes the co-occurrence degree between the concept c and each term t_i in the query Q, which is computed using the term frequency (tf) in the top-ranked documents and inverse document frequency (idf) in the corpus [12,13]. δ is a small positive constant (e.g., $\delta = 0.1$) to avoid that the value of Rel_{LCA} is zero. Multiplication in Eq. (2) is used to emphasize co-occurrence with all query terms.

- Against the "incorrect mapping" problem. For a concept c in the expanded concepts *ExpansionSet*, generated in Section 5.2, if the average of Rel_{LCA} values of concepts expanded from c is very small, it suggests that these expanded concepts are statistically irrelevant to the user's query through analyzing the document corpus. For example, as stated in Section 5.1, the terms "cover", "screening", "masking", "coating" and "application" are initially mapped to the IFC concept "covering". However, the retrieval documents using query term "application" barely include any other related concepts of "covering". Hence the mapping from term "application" to concept "covering" is regarded as an incorrect mapping. Then all the concepts expanded from "covering" in *ExpansionSet*, including "membrane", "wrapping", etc., should be removed.

- Against the "overexpansion" problem. The concepts in *ExpansionSet*, which have the small values of Rel_{LCA} (e.g., less than a predefined threshold 0.3), are removed from the *ExpansionSet*. This means that the expanded concepts with small Rel_{LCA} values are not quite relevant to the original query. After pruning the irrelevant concepts, the remaining part in *ExpansionSet* is added to the final query terms. For example, the mapping from the query term "masking" to IFC concept "covering" can be accepted according to the above principle. However, some of the expanded concepts, like "skirting board" has the relatively small Rel_{LCA} values, hence are abandoned from *ExpansionSet* in our algorithm.

After pruning all irrelevant concepts from the *ExpansionSet*, the final relatedness weight of each concept in *ExpansionSet* can be calculated by combining Eq. (1) and Eq. (2). Given a query concept c and its expanded c_i in *ExpansionSet*, the relatedness weight between c_i and the user's query Q is defined by

$$\omega\left(c_{i,Q}\right) = \frac{\alpha \cdot Rel_{IFC}(c_i, c) \cdot Rel_{LCA}\left(c_{i,Q}\right)}{\alpha \cdot Rel_{IFC}(c_i, c) + Rel_{LCA}\left(c_{i,Q}\right)}, \tag{3}$$

where α is the importance factor (we typically set $\alpha = 1.0$ in our implementation).

5.4. Step 4: rank the documents of BIM resources

After the user's query is expanded by IFC IR Ontology and LCA, document ranking can be implemented based on each document's score against the query. The ranking process uses the vector space model (VSM) [14] to determine how relevant a given document is to the user's query. In the process, a Boolean model is first used to narrow down the documents that need to be scored based on the use of Boolean logic in query specification. Then, these resultant documents are ranked based on their matching scores between the query vector and their document vectors using the VSM. In the VSM implemented in this research, the expanded concepts are associated with the weights computed in Eq. (3). The score between the user's query Q and a document D in corpus is denoted by

$$score(Q, D) = \frac{1}{\left\|\overrightarrow{V_Q}\right\|} \sum_{t \in Q} \left[tf(t, D) \cdot idf(t)^2 \cdot \omega(t, Q) \cdot \frac{1}{\left\|\overrightarrow{V_D}\right\|} \right] \tag{4}$$

where t is a term in the query Q, $\overrightarrow{V_Q}$ is the query vector of user's query Q, $\overrightarrow{V_D}$ is the document's vector of D. $tf(\cdot)$ is the term frequency in document D, which measures how often a term appears in the document. $idf(\cdot)$ is the inverse document frequency, which measures how often the term appears in document corpus. $\omega(t, Q)$ is the weight, calculated by Eq. (3), measuring the importance of each query term t in the query Q.

6. Experimental results

6.1. System overview

By combining the usability of keyword-based interface with automatic query expansion techniques, we have developed a semantic search engine, named BIMSeek, for retrieving online BIM documents. The presented search engine is a *query-driven* information retrieval system, which automatically expands the user's query based on the ontology and then ranks BIM document repositories by analyzing the semantic details carried by such documents. For intuitively showing the ontology's concepts associated with the user's query, an *ontology navigation toolbar* is also developed. The toolbar provides a dynamic navigation way, which allows the user to manually refine the query through exploring the related concepts.

Fig. 4 shows a snapshot view of the main interface of our search engine, which consists of four parts. First, the user inputs a specific query on the top-left (see "Query Area"). Then, the system automatically ranks online BIM documents and displays the search results produced by our algorithm on the bottom-left (see "Result Area"). The ontology navigation toolbar on the top-right allows the user to manually refine the query through clicking on the related concepts (see "Navigation Bar"). Alternatively, one can adjust the algorithm parameters for refining search through the control panel on the bottom-right (see "Parameter Panel").

In the process of building IFC IR Ontology in Section 4, we use OntoSTEP, a Protégé EXPRESS tool [40], to semi-automatically construct a raw IFC IR Ontology, and this raw ontology is used for further ontology

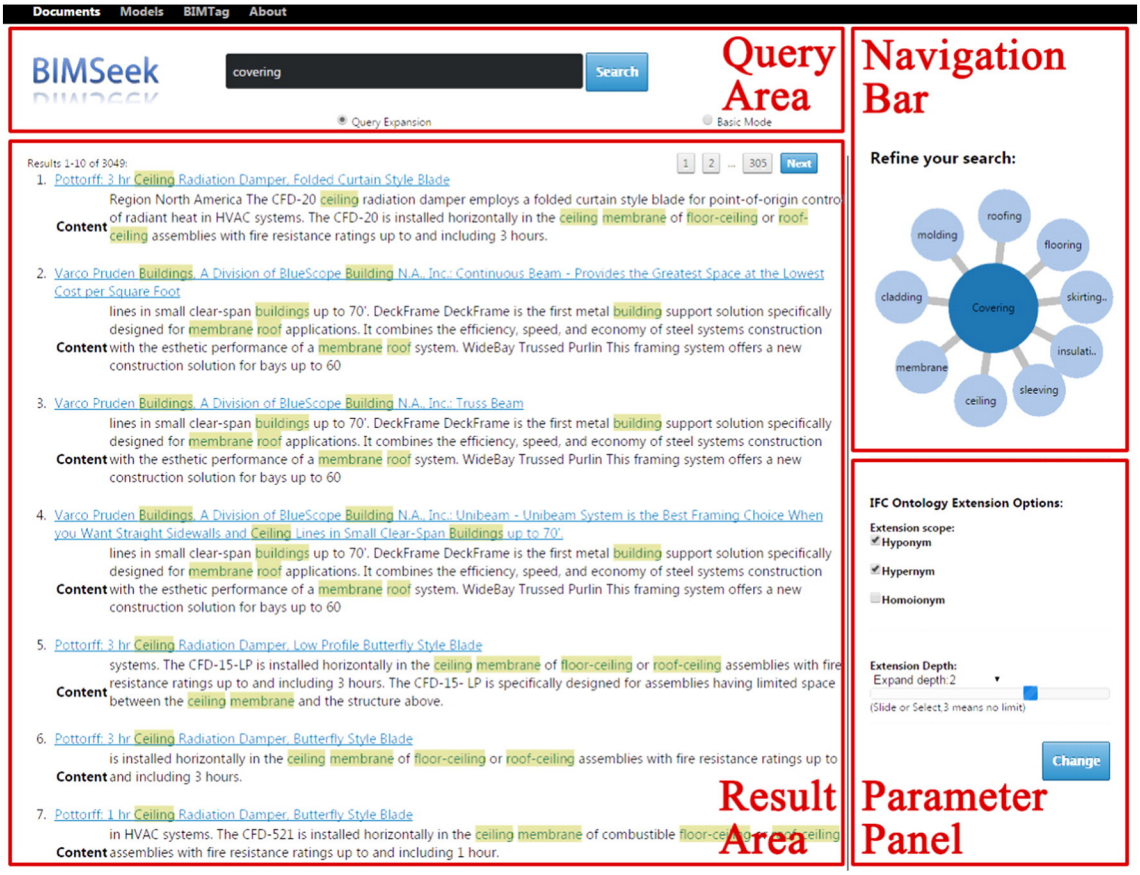
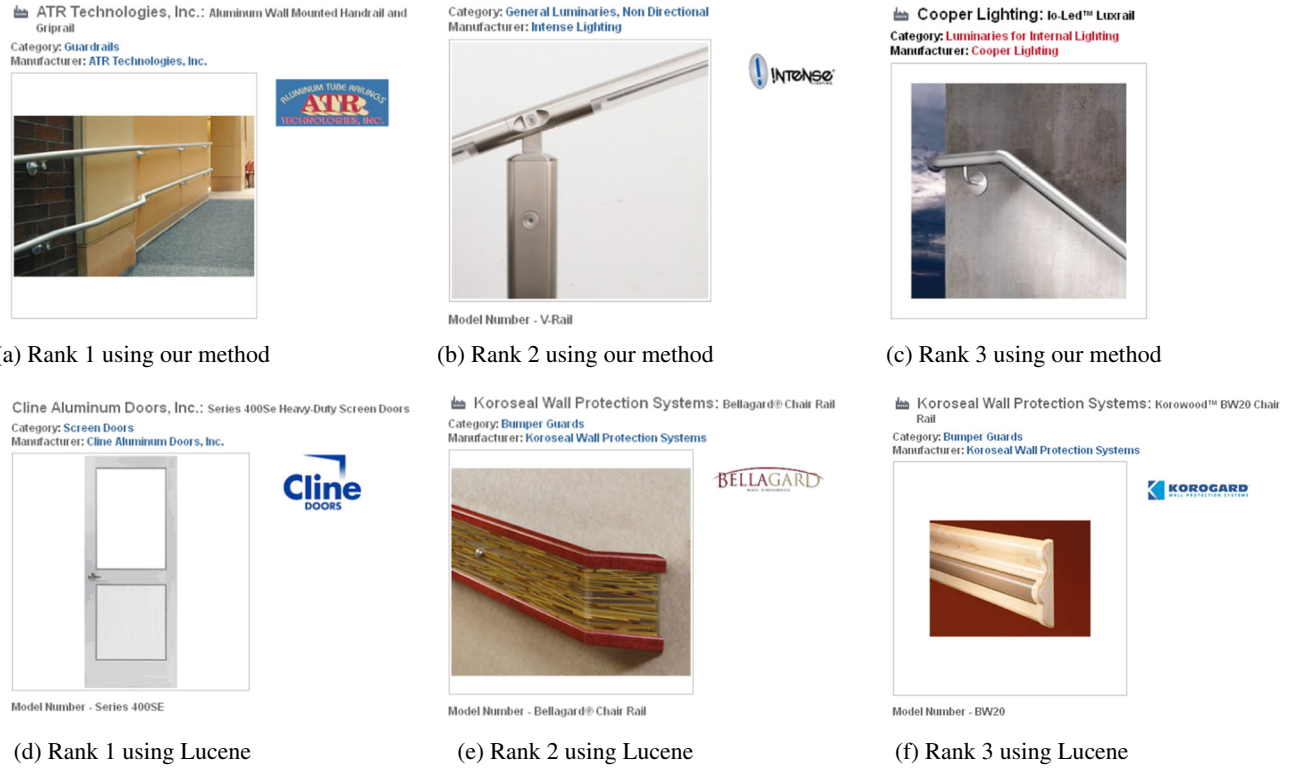


Fig. 4. The screenshot of our search engine. It allows a user to specify a search query using keywords. Then it returns the ranked results related to online BIM resources. The user may refine the search through clicking on "Navigation Bar" for accessing the query-related concepts. The algorithm parameters can be alternatively adjusted through the "Parameter Panel".



(a) Rank 1 using our method (b) Rank 2 using our method (c) Rank 3 using our method

(d) Rank 1 using Lucene (e) Rank 2 using Lucene (f) Rank 3 using Lucene

Fig. 5. Comparison of search results between our method and traditional keyword-based search (Lucene [48]). For the same test query "railing", the webpages corresponding to top 3 query results (from left to right) are typically selected for displaying, where the top row shows our ranking results and the bottom row shows the ones using Lucene system.

refinement. Then the built ontology is automatically converted into OWL structured form. In the process of semantic retrieval in Section 5, the toolkit OWLAPI, a semantic system framework, is used for handling ontology and OWL language. The indexing and ranking are offered by Apache Lucene, and Apache Nutch Web Crawler is used to collect online BIM documents.

The search system developed in this work has been designed to be a flexible tool to test different retrieval methods. The system also provides a friendly interface for the user, in which the user can modify several simple parameters for adjusting the search results. In this section, all the experiments are run on a 2.80 GHz processor with 4 GB memory under Windows 7.

6.2. Evaluation

In the experiments, the comparison between our method and keyword-based search has been conducted. To measure the performance of the keyword-based search, Lucene [48] with the default function was used. Lucene is frequently used as a baseline keyword-based system in IR. The Lucene [48] scoring uses a combination of the VSM and the Boolean model to determine how relevant a given document is to a user's query.

Fig. 5 shows an example of search results with the test query "railing" applied to the benchmark collection of online BIM documents. We typically display the webpages corresponding to top three query results (from left to right), where the top row is our ranking results and the bottom row is the ones using the keyword-based search. For the keyword-based search, only the webpages that contain the keywords "railing" or "rail" are returned. However, some top search results, such as the webpage "Heavy-Duty Screen Doors" in Fig. 4 that just has a rail material description, don't sufficiently reflect user's search intent ("railing"). In contrast, using our method, some ontology's concepts (e.g., "balustrade", "handrail" and "guardrail") associated with the input query "railing" are expanded as new query terms. Consequently, the most relevant webpages, which are not only similar to the input query word (i.e., "railing") but also semantically satisfying the expanded concepts (e.g., "guardrail" or "balustrade"), tend to be ranked at the top in our system. For example, rank 1 in Fig. 4 is a webpage about "aluminum Wall Mounted Handrails", rank 2 in Fig. 4 is the one about "stainless steel handrail", and rank 3 in Fig. 4 is the one about "an indoor/outdoor LED-based handrail". In this way, BIM documents, which do not only hit the query keywords but also semantically satisfy an information seeker's needs, can be found. The example suggests that our retrieval method has an obvious advantage over the keyword-based search in BIM-specific domain.

In order to evaluate the effectiveness of the proposed method, we run an experiment of information retrieval on a document collection. Currently, the document collection contains a total number of 15,176 BIM documents acquired from Autodesk Seek [2]. Autodesk Seek provides three industry-standard classifications (including MasterFormat, OmniClass and UniFormat) for browsing BIM model catalogue. In this test, we typically select OmniClass as the baseline classification of BIM documents for our retrieval evaluation. The OmniClass number of each BIM document is obtained through crawling the categories from the website, and the OmniClass number is used for "ground truth" for each test query. For instance, the OmniClass number "23.35.00.00" has its name "Covering, Cladding, and Finishes". Then, we conceive of a test query "Covering, Cladding, and Finishes", and judge whether each webpage in the search results belongs to the corresponding OmniClass classification (i.e., "23.35.00.00") or its subclass (e.g., "23.35.20.21"). There are 30 test queries used in our test.

To measure the performance of our method, we adopt the standard evaluation procedure from information retrieval, namely *precision–recall* curves, which is used for evaluating the retrieval results [49,50]. The precision–recall curves describe the relationship between precision and recall for an information retrieval method. In the precision–recall curve, the number of relevant documents for each query is denoted as *Relevant*, the number of documents retrieved for the query is denoted as *Retrieved*, and the number of relevant documents correctly retrieved is denoted as *Relevant* ∩ *Retrieved*. Then the recall is defined as $\frac{Relevant \cap Retrieved}{Relevant}$, and the precision is defined as $\frac{Relevant \cap Retrieved}{Retrieved}$ for each experiment. It is desirable to achieve both high precision and recall, but unfortunately this is rather difficult to achieve, especially for the text-based retrieval problem.

We compare our method with several retrieval methods, which include the keyword-based search (Lucene), query expansion based on WordNet (synonyms, hyponyms and hypernyms, respectively). Fig. 6 shows the average precision–recall curves. The results show that our method performs better than both the keyword-based search and WordNet-based query expansion methods.

In addition to the precision–recall curves, we also evaluate other quantitative statistics for evaluation of retrieved results. Specifically, we compute E-measure and F-measure [51]. F-measure (also F-score) is a measure of a test's accuracy, and it is the harmonic mean of precision and recall. The F-measure is defined as

$$F_{\beta=(1+\beta^2)\cdot}\frac{Precision\cdot Recall}{\beta^2\cdot Precision + Recall}.$$

It measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision. β is a non-negative real value denoting the times as much importance to recall as precision. Since the recall and precision are both important in BIM resource retrieval, we set $\beta = 1$ to let the recall and precision rate evenly weighted.

Another score combining precision and recall is E-measure, which means the effectiveness measure:

$$E = 1 - \frac{1}{\alpha\frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}}.$$

In our test, α is set as 0.5 when β is 1. Table 2 gives F-measure and E-measure of our method and other methods. By examining the results, we can see that our method is higher in performance compared with other methods.

7. Conclusions and future work

Based on IFC IR Ontology and local context analysis (LCA), this paper has introduced an automatic query expansion method for retrieving online BIM resources. IFC IR Ontology can be used for the disambiguation of terms on online BIM documents. Ontology-based query

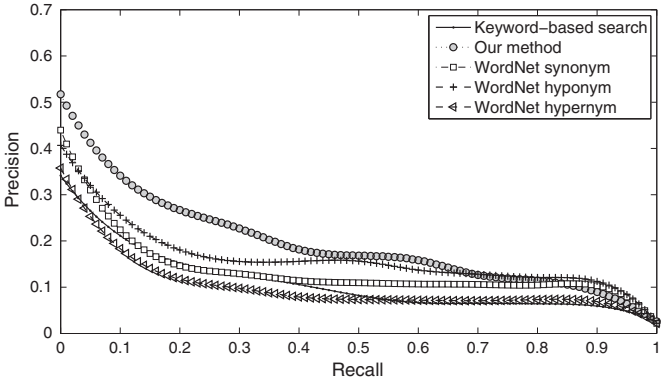


Fig. 6. The precision–recall curves of retrieval results using our method and other methods.

Table 2
F-measure and E-measure of our method and others.

Method	F-measure	E-measure
Keyword-based search	0.180093695	0.819906305
WordNet synonym	0.162367191	0.875350332
WordNet hyponym	0.189367722	0.837632809
WordNet hypernym	0.124649668	0.810632278
Our method	0.24816183	0.75183817

expansion enables the addition of standardized domain terms for search query, while LCA, through BIM document analysis, offers the broader scope of terms related to user information needs. The experimental results show that our method can achieve better efficiency for retrieving BIM documents, than traditional keyword-based search and WordNet-based query expansion methods.

In the current implementation, only the inheritance relationship and the type enumeration in the IFC ontology are used for expanding and matching BIM documents. In our ongoing work, we are trying to use the properties and restrictions in the ontology for searching BIM models. One alternative way is to extract and convert each BIM model into in RDF or OWL format, and query with SPARQL.

One of the limitations of our method is that the single IFC ontology cannot fully cover the IR needs of BIM-specific domain because of its multidisciplinary and multistakeholder nature. Therefore, there is a need to combine or merge the IFC ontology with some existing AEC ontologies such as Uniclass or OmniClass [52], so that more extensive BIM resources can be covered. In the future, we would like to integrate more AEC ontologies (e.g., OmniClass) in the search engine. On the other hand, the number of expansion terms using LCA is arguable. Too few expansion terms may have no impact, and too many will cause a query drift. A more effective query expansion method is another future direction of research.

The experiment described in Section 6.2 is essentially a system-oriented evaluation [21], where the benchmark collection of BIM documents is acquired from Autodesk Seek [2] and OmniClass is used as the baseline classification of BIM documents for retrieval evaluation. Although such a system-oriented experiment can help evaluating the effectiveness of particular document ranking, it cannot examine the full operational characteristics of the proposed search engine. Therefore, a user-centered evaluation [53] might be conducted to supplement the system-oriented experiment by multiple humans independently. In particular, the inter-rater agreement (or inter-rater reliability) of document ranking can be assessed by some statistical measures such as Fleiss' kappa [54]. The user-centered evaluation studies will be left as part of our future work.

Acknowledgments

The authors appreciate the comments and suggestions of all reviewers, whose comments significantly improved this paper. The authors would like to thank Dr. Pieter Pauwels at Ghent University for discussing the problem of IFC-to-RDF conversion. The research is supported by the National Science Foundation of China (61472202, 61272229, 61003095) and the National Technological Support Program for the 12th-Five-Year Plan of China (2012BAJ03B07). The fourth author is supported by the Chinese 973 Program (2010CB328003) and the last author is supported by the Chinese 863 Program (2012AA040902).

Supplementary material

The proposed system and its demonstration can be accessed at: http://cgcad.thss.tsinghua.edu.cn/liuyushen/ifcqe/.

References

[1] C. Eastman, P. Teicholz, R. Sacks, K. Liston, BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors, 2nd Edition John Wiley and Sons, NJ, 2011.
[2] Autodesk, AutodeskSeek, http://seek.autodesk.com 2014.
[3] BIMobject, http://bimobject.com/ 2014.
[4] NBS of UK, National BIM Library, http://www.nationalbimlibrary.com/ 2014.
[5] Google, Google 3D Warehouse, http://sketchup.google.com/3dwarehouse/ 2014.
[6] SmartBIM LLC, SmartBIM, http://www.smartbim.com/ 2014.
[7] Pierced Media LC, RevitCity, http://www.revitcity.com 2014.
[8] A. Court, D. Ullman, S. Culley, A comparison between the provision of information to engineering designers in the UK and the USA, Int. J. Inf. Manag. 18 (6) (1998) 409–425.
[9] G.J. Hahm, M.Y. Yi, J.H. Lee, H.W. Suh, A personalized query expansion approach for engineering document retrieval, Adv. Eng. Inform. 28 (4) (2014) 344–359.
[10] Z. Li, V. Raskin, K. Ramani, Developing engineering ontology for information retrieval, J. Comput. Inf. Sci. Eng. 8 (1) (2008) 737–745.
[11] Industry Foundation Classes (IFC), IFC4 Release Candidate 4, http://www.buildingsmart-tech.org/ifc/IFC2x4/rc4/html/index.htm 2014.
[12] J. Xu, W.B. Croft, Query expansion using local and global document analysis, Proceedings of ACM SIGIR'96 1996, pp. 4–11.
[13] J. Xu, W.B. Croft, Improving the effectiveness of information retrieval with local context analysis, ACM Trans. Inf. Syst. 18 (1) (2000) 79–112.
[14] G. Salton, Developments in automatic text retrieval, Science 253 (5023) (1991) 974–980.
[15] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[16] T. Hofmann, Probabilistic latent semantic indexing, Proceedings of ACM SIGIR'99 1999, pp. 50–57.
[17] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (4–5) (2003) 993–1022.
[18] J. Bhogal, A. Macfarlane, P. Smith, A review of ontology based query expansion, Inf. Process. Manag. 43 (4) (2007) 866–886.
[19] S. Kara, O. Alan, O. Sabuncu, S. Akpinar, N. Cicekli, F. Alpaslan, An ontology-based retrieval system using semantic indexing, Inf. Syst. 37 (4) (2012) 294–305.
[20] O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using explicit semantic analysis, ACM Trans. Inf. Syst. 29 (2) (2011) (Article 8).
[21] X. Yan, R. Lau, D. Song, X. Li, J. Ma, Toward a semantic granularity model for domain-specific information retrieval, ACM Trans. Inf. Syst. 29 (3) (2011) (Article 15).
[22] P. Demian, P. Balatsoukas, Information retrieval from civil engineering repositories: importance of context and granularity, J. Comput. Civ. Eng. 26 (6) (2012) 727–740.
[23] K. Lin, S. Hsieh, H. Tserng, K. Chou, H. Lin, C. Huang, K. Tzeng, Enabling the creation of domain-specific reference collections to support text-based information retrieval experiments in the architecture, engineering and construction industries, Adv. Eng. Inform. 22 (3) (2008) 350–361.
[24] K. Lin, L. Soibelman, Incorporating domain knowledge and information retrieval techniques to develop an architectural/engineering/construction online product search engine, J. Comput. Civ. Eng. 23 (4) (2009) 201–210.
[25] H.-T. Lin, N.-W. Chi, S.-H. Hsieh, A concept-based information retrieval approach for engineering domain-specific technical documents, Adv. Eng. Inform. 26 (2) (2012) 349–360.
[26] A. Weissman, M. Petrov, S. Gupta, A computational framework for authoring and searching product design specifications, Adv. Eng. Inform. 25 (3) (2011) 516–534.
[27] Y. Rezgui, Ontology-centered knowledge management using information retrieval techniques, J. Comput. Civ. Eng. 20 (4) (2006) 261–270.
[28] L. Zhang, R. Issa, Ontology based partial building information model extraction, J. Comput. Civ. Eng. 27 (6) (2013) 576–584.
[29] J. Beetz, J. Leeuwenand, B. Vries, IfcOWL: a case of transforming EXPRESS schemas into ontologies, Artip. Intell. Eng. Des. Anal. Manuf. (AI EDAM) 23 (1) (2009) 89–101.
[30] L. Zhang, R. Issa, Development of IFC-based construction industry ontology for information retrieval from IFC models, Proceedings of the 2011 EG-ICE Workshop, 2011.
[31] P. Pauwels, D.V. Deursen, R. Verstraeten, J.D. Roo, R.D. Meyer, R.V. de Walle, J.V. Campenhout, A semantic rule checking environment for building performance checking, Autom. Constr. 20 (5) (2011) 506–518.
[32] M.P. Nepal, S. Staub-French, R. Pottinger, A. Webster, Querying a building information model for construction-specific spatial information, Adv. Eng. Inform. 26 (4) (2012) 904–923.
[33] A. Borrmann, E. Rank, Specification and implementation of directional operators in a 3D spatial query language for building information models, Adv. Eng. Inform. 23 (1) (2009) 32–44.
[34] G. Gao, Y.-S. Liu, M. Wang, X.-G. Han, BIMTag: semantic annotation of web BIM product resources based on IFC ontology, 21st International Workshop: Intelligent Computing in Engineering 2014 (EG-ICE), 2014.
[35] T. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220.
[36] G. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[37] T. Lukasiewicz, U. Straccia, Managing uncertainty and vagueness in description logics for the semantic web, J. Web Semant. 6 (4) (2008) 291–308.
[38] W3C, OWL 2 Web Ontology Language Overview, http://www.w3.org/TR/owl-features/ 2009.

[39] N.F. Noy, D.L. McGuinness, Ontology development 101: a guide to creating your first ontology, http://protege.stanford.edu/publications/ontology_development/ontology1 01.html 2014.
[40] Protégé, An ontology and knowledge base editor, http://protege.stanford.edu/ 2014.
[41] K. Lin, L. Soibelman, Promoting transactions for A/E/C product information, Autom. Constr. 15 (6) (2006) 746–757.
[42] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, Knowl. Eng. Rev. 11 (2) (1996) 93–136.
[43] C. Leacock, M. Chodorow, Combining Local Context and WordNet Similarity for Word Sense Identification, 1st EditionMIT Press, 1998 (Ch. 11).
[44] J. Jiang, D. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, Proceedings of International Conference Research on Computational Linguistics 1997, pp. 112–123.
[45] D. Lin, An information-theoretic definition of similarity, Proceeding of 15th International Conference on Machine Learning 1998, pp. 296–304.
[46] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, International Joint Conference on Artificial Intelligence 1995 1995, pp. 448–453.
[47] Z. Wu, M. Palmer, Verb semantics and lexical selection, Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics 1994, pp. 448–453.
[48] Lucene, http://lucene.apache.org/ 2014.
[49] V. Raghavan, P. Bollmann, G.S. Jung, A critical investigation of recall and precision as measures of retrieval system performance, ACM Trans. Inf. Syst. 7 (3) (1989) 205–229.
[50] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
[51] C.J.V. Rijsbergen, Information Retrieval, 2nd Edition Butterworths, London, 1979.
[52] N. El-Gohary, T. El-Diraby, Merging architectural, engineering, and construction ontologies, J. Comput. Civ. Eng. 25 (2) (2011) 109–128.
[53] T. Saracevic, Evaluation of evaluation in information retrieval, Proceedings of ACM SIGIR'95 1995, pp. 138–146.
[54] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378–382.