Advanced Engineering Informatics 31 (2017) 48–61

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

# BIMTag: Concept-based automatic semantic annotation of online BIM product resources

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

[a] 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 21 January 2015
Received in revised form 11 August 2015
Accepted 12 October 2015
Available online 23 October 2015

*Keywords:*
Building Information Modeling (BIM)
Industry Foundation Classes (IFC)
Semantic annotation
Latent semantic analysis (LSA)
Information retrieval

ABSTRACT

With the rapid popularity of Building Information Modeling (BIM) technologies, BIM resources such as building product libraries are growing rapidly on the World Wide Web. However, numerous BIM resources are usually from heterogeneous systems or various manufacturers with ambiguous expressions and uncertain categories for product descriptions, which cannot provide effective support for information retrieval and categorization applications. Therefore, there is an increasing need for semantic annotation to reduce the ambiguity and unclearness of natural language in BIM documents. Based on Industry Foundation Classes (IFC) which is a major standard for BIM, this paper presents a concept-based automatic semantic annotation method for the documents of online BIM products. The method mainly consists of the following two stages. Firstly, with reference to the concepts and relationships explicitly defined in IFC, a word-level annotation algorithm is applied to the word-sense disambiguation. Secondly, based on latent semantic analysis technique, a document-level annotation algorithm is proposed to discover the relationships which are not explicitly defined in IFC. Finally, a prototype annotation system, named BIMTag, is developed and combined with a search engine for demonstrating the utility and effectiveness of our method. The BIMTag system is available at http://cgcad.thss.tsinghua.edu.cn/liuyushen/bimtag/.



## 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 technologies 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 search for a large variety of BIM products from manufactures. Currently, it contains 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 containing over 450,000 BIM models with the product data and properties. Other online libraries (e.g. National BIM Library [4] and SmartBIM [5]) and many active online communities (e.g. RevitCity [6]) also have a large amount of information content of BIM-related building products.

The typical libraries of online BIM resources (e.g. [2,4,5]) contain BIM models associated with product documents (e.g. specifications and descriptions of the objective products). The BIM models are normally in their native file format dependent on various 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 are the textual content for describing BIM models including their functions, dimensions, materials, performances, manufacturers, etc. These product documents are independent of the file format of BIM models. In particular, much of knowledge is embedded in textual BIM docu-

* 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), c_loud26@163.com (P. Lin), 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.aei.2015.10.003
1474-0346/

ments generated during design and construction phases [7]. Most of BIM documents are unstructured, in contrast to structured content (e.g. BIM models or database tables) following the strict schema.

However, numerous BIM documents are often obtained from heterogeneous systems or generated by various manufacturers, which are written in unstructured and ungrammatical format possibly with ambiguous expressions and uncertain categories for product descriptions. As a result, this also increases the difficulty for users in retrieving most relevant and accurate information through traditional keyword-based search engines. To overcome this issue, a possible way is to manually annotate the BIM documents to help classify them with specific labels or tags, which is very labor-intensive and subjective. Therefore, there is an increasing need for automatic semantic annotation to reduce the ambiguity and unclearness of natural language in BIM documents.

*Semantic annotation* is about attaching additional information (e.g. names, attributes, comments, descriptions) to a document or to a selected part in the text [8], thereby providing metadata about an existing piece of data. It could help reduce the ambiguity and unclearness of natural language through expressing the notions and their relationships in a more formal language. Many studies have contributed in semantic annotation [9–12], which lower the barrier of linking shared data with the Web resources in various areas. However, the lack of commonly accepted domain-specific formal knowledge still limits the utilization of semantic annotation in the BIM-related area. Therefore, the crucial problem is how to build the BIM-oriented formal knowledge and use the formal knowledge to annotate the Web content of textual BIM documents in different semantic levels.

Based on Industry Foundation Classes (IFC) [13] which is a major standard for BIM, this paper presents a concept-based automatic semantic annotation method for online BIM documents. The method mainly consists of the following two stages. Firstly, with reference to the concepts and relationships explicitly defined in IFC, a word-level annotation algorithm is applied to handle the word-sense disambiguation explicitly. Secondly, by combining the latent semantic analysis technique [14], a document-level annotation algorithm is proposed to discover the relationships that are not explicitly defined in IFC. Finally, a prototype semantic annotation system, named BIMTag, is developed and combined with a search engine for demonstrating the utility and effectiveness of our method. Compared with conventionally manual annotation/tagging approaches, which are time consuming and subjective, our method can automatically derive the intended meaning of terms and their underlying concepts embedded in the content of documents. This also enriches the content of unstructured BIM documents with their contexts which are further linked to the knowledge of BIM-specific domain.

### 1.1. Related work

#### 1.1.1. An overview for semantic annotation of documents

In general, the performance of information retrieval can be improved by two aspects: (1) enhancing semantic annotation of documents and (2) enhancing the user query mechanism. Both aspects are active research areas. This paper focuses on the former, i.e. enhancing semantic annotation of documents. In contrast, our previous paper [15] dealt with the latter, which enhances the user query mechanism for information retrieval without using semantic annotation of documents. The two papers benefit from the preliminary thesaurus of IFC.

Semantic annotation of documents can be performed manually, automatically or semiautomatically [16]. Manual annotation is impractical and unscalable for numerous BIM documents, while automatic annotation tools remain a research challenge. This paper mainly focuses on automatic semantic annotation, leaving manual annotation.

Over the past few decades, automatic semantic annotation has become an increasingly important research topic, which enables many applications such as highlighting, indexing, retrieval, categorization and information extraction [8,12,16]. Semantic annotation aims to formally identify concepts and their relationships in documents. Its implementation consists of two major phases: (1) ontology-based lookup and (2) reference disambiguation [16]. In computer science and information science, an *ontology* is defined as formal, explicit specification of shared conceptualization [17]. The ontology-based lookup is concerned with identifying all candidate mentions of concepts from the ontology. The reference disambiguation then uses contextual information from documents as well as knowledge from the ontology to disambiguate the mentions to the correct ontology concept. Most of existing annotation approaches are based on syntactic matching of ontology concept labels (descriptions) from the content of documents [8,12]. The reader may consult several previous literature (e.g. [8,12,16]) for an overview of current studies. A survey of the state of the art is beyond the scope of this paper. Instead, this section briefly reviews the most related studies associated with our work.

#### 1.1.2. Semantic annotation in engineering document retrieval

Although the main issue discussed in this paper is semantic annotation of online BIM documents, many previous techniques have been developed for annotation, indexing and retrieval of engineering documents. Therefore, reviewing the engineering case will provide a good understanding for our work.

In contrast to general documents, engineering documents are different due to their syntax variations and semantic complexities [18,19]. Syntax variations mainly refer to the usage of synonyms, abbreviations and acronyms, which reflect the domain-specific contents. Semantic complexities occur from the domain-specific relationships among the engineering terms as well as polysemic words. Therefore, a proper disambiguation process is necessary to map the ambiguous terms in engineering documents to standardized concepts. The semantic ambiguity can be alleviated by using a domain ontology, which bridges the gap between query terms and documents. Based on the domain ontology, semantic annotation of documents can be conducted for further information retrieval purpose. In particular, for engineering document retrieval, ontology-based query expansion approaches are a promising direction, since ambiguous terms in user queries and documents can be effectively expanded and interpreted by the domain ontology.

In the last few years, several studies have been devoted to engineering document retrieval with the help of semantic annotation or indexing. For instance, Rezgui [20] used either direct or indirect ontology concept mapping to assist indexing and retrieving construction documents. Li et al. [19] developed an engineering ontology in mechanical design and manufacturing, and applied the ontology to concept tagging and indexing for retrieving unstructured engineering documents and CAD drawings. Weissman et al. [21] proposed a computational framework and a software tool based on this framework for writing, annotating, and searching computer-interpretable product design specifications. Lin et al. [22] presented a passage partitioning approach according to a domain ontology, which provided the ability to generate the concepts in each passage. More recently, Hahm et al. [18] introduced a semantic indexing approach to solve the syntax variations and semantic complexities of engineering documents for information retrieval.

However, the above approaches are still limited to the needs of their particular applications, which cannot provide effective and comprehensive support for our purpose, i.e. automatic semantic annotation of online BIM product documents. On the one hand, most of existing studies focuses on the word-level or passage-level annotation/indexing rather than the document level, which lack the ability to understand the document as a whole. On the other hand, the above ontologies in engineering document retrieval are mostly hand-crafted, which lack the utilization of BIM-specific domain knowledge in semantic annotation.

#### 1.1.3. BIM-related semantic resources and ontologies

Ontology is considered as a key element to enhance domain-specific semantic annotation [17,20,23,24]. It 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 [25]) contain a large number of general concepts, they are not designed for specific domain, 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 the description of properties of each concept [26].

Several well-known semantic resources have been developed for various AEC applications, and they have the potential to be enhanced as domain ontology for annotating online BIM documents [20,27]. The most notable efforts [20,27] include ISO 12006-2, Uniclass, OmniClass, Industrial Foundation Classes [28], etc. Among these existing semantic resources, structured taxonomies deserve particular attention. However, improper taxonomies may have the opposite effects leading to confusion and difficulty to retrieve [20], so a properly structured taxonomy should be carefully selected to meet our needs. As a relatively new field, ontology research for BIM resources is still rare, which needs more exploration.

As the commonly used data exchange standard for BIM, Industry Foundation Classes [28] developed by buildingSMART (formerly the International Alliance for Interoperability, IAI) to facilitate interoperability in the AEC industry. Today, 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 semantic resource, which provides the commonly shared concepts, attributes and relationships of BIM resources.

Recently, several IFC-based ontologies have been developed for particular application needs [29–33]. For instance, Pauwels et al. [32] utilized an IFC ontology to semantic rule checking. Beetz et al. [30] 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 [29] used an IFC ontology to extract partial model from a complete IFC model. In addition, several applications used IFC ontologies for querying spatial information within a building information model [33,34].

Up to now, very little attention is paid to utilize IFC-based ontologies for semantic annotation of online BIM product documents. In contrast, this paper develops a BIM-specific ontology based on the IFC schema, and applies the ontology to annotate and index online BIM documents. To alleviate the insufficiency of hand-crafted ontology, our method combines the word-level annotation (i.e. word-matching technique based on context analysis) with the document-level annotation (i.e. latent semantic analysis technique).

### 1.2. Contributions

Our main contributions can be summarized as follows.

- Based on the IFC ontology, we present a concept-based automatic semantic annotation method for online BIM documents, which can be used for information extraction, indexing and retrieving applications. The presented method mainly consists of the following two stages.
- On the first stage, a word-level annotation algorithm using context analysis is developed for annotating BIM documents to handle the word-sense disambiguation problem. The algorithm aims to derive the intended meaning of terms and their underlying concepts embedded in the content of documents.
- On the second stage, a document-level annotation algorithm based on latent semantic analysis [14] is proposed to discover the relationships that are not explicitly defined in IFC. This can be used to discover the *topic* and some metadata of the documents.
- Our method is evaluated on the document collection acquired from Autodesk Seek. The experimental results show that our method performs better retrieval than both the search without semantic annotation and the search with WordNet-based semantic annotation. Finally, the proposed semantic annotation method is combined with a search engine for retrieving online BIM product resources.

The remainder of the paper is organized as follows. Section 2 describes our concept-based semantic annotation method. Section 3 illustrates our semantic annotation system and demonstrates the experimental results. Finally, Section 4 concludes the paper and discusses the limitations and future work.

## 2. The concept-based semantic annotation method

Online BIM product documents are usually written in a way for human beings to read, but not formally described as the standardized concepts within the IFC specification. Many previous studies [18,19,21,22] have shown that fundamental characteristics of human verbal behavior have greatly limited engineering document retrieval. On the one hand, information providers often use different words for describing the same meaning or concept (synonymy) because of the tremendous variety in the vocabulary. As a result, the relevant concepts in the IFC specification will therefore be missed. On the other hand, since a single word often has more than one meaning (polysemy), the irrelevant concepts in the IFC specification might be returned.

Concept-based semantic annotation is a promising direction, which could overcome some of the above issues by employing word mapping, word-sense disambiguation (WSD), and dimensional reduction techniques. Also, it can help derive the meanings of terms in documents and their underlying concepts, rather than by simply matching characters or strings like keyword matching technologies. Concept-based semantic annotation can be accomplished in a variety of ways according to the granularity of structuring units.

The proposed method includes two levels of semantic annotation: a word-level annotation as the finest granularity, and a document-level method as the coarse granularity. Starting with online BIM documents as input, our method mainly consists of the following four steps:

(1) IFC-based ontology construction (see Section 2.1).
(2) Word-level semantic annotation (see Section 2.2).
(3) Document-level semantic annotation (see Section 2.3).

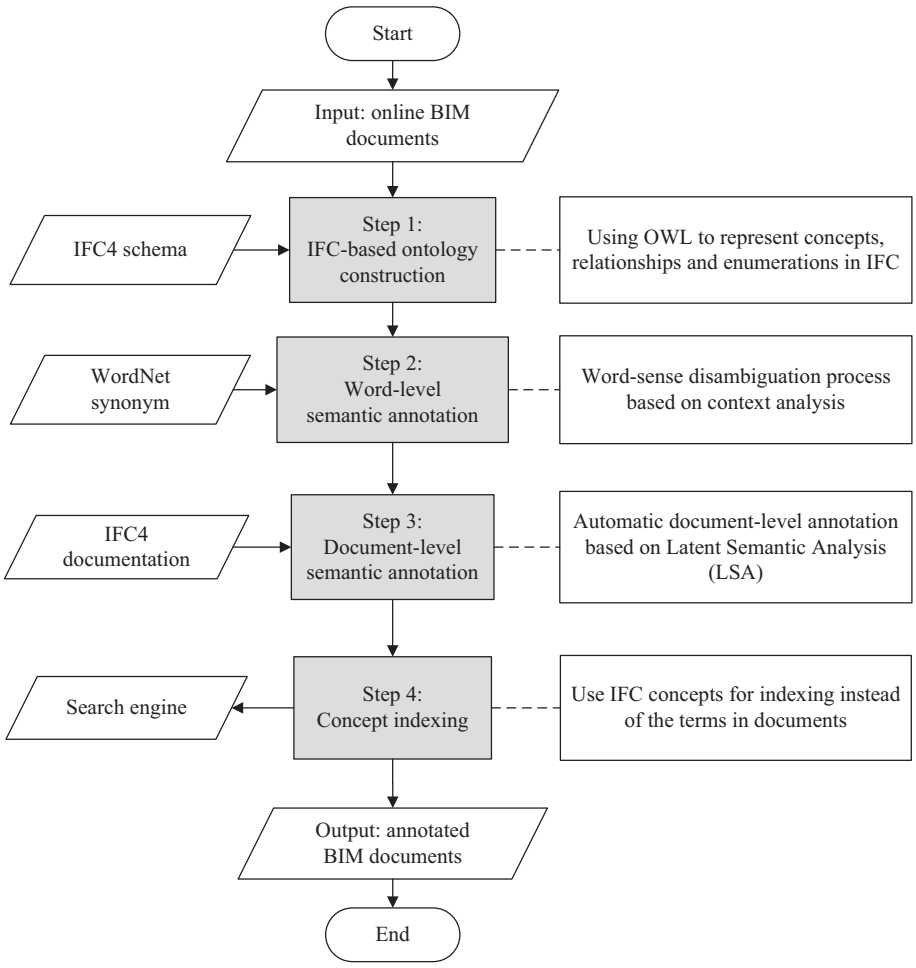


**Fig. 1.** The main procedure of our semantic annotation method.

(4) Concept indexing (see Section 2.4).

In this method, ontology construction and concept indexing are performed offline, while the semantic annotation processes (including word-level and document-level) are conducted automatically. Fig. 1 shows the main procedure of our method. The following will introduce each step in details.

### 2.1. Step 1: IFC-based ontology construction

To achieve automatic semantic annotation of online BIM documents, the principal work relies on two aspects: (1) how to construct an IFC-based ontology for the need of annotation and (2) how to utilize the ontology for annotating and indexing BIM documents. The former is introduced here and the latter will be presented in Sections 2.2–2.4.

The proposed annotation method is based on the latest version IFC4 specification [13]. The IFC4 specification enhances the capability of previous IFC specifications in several areas of building elements, building service elements and structural elements. 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.

#### 2.1.1. The scope of the ontology for semantic annotation

Although the underlying IFC specification has provided a candidate semantic resource for BIM, the scope defined in IFC is too broad for our semantic annotation purpose. This will greatly limit the accuracy and efficiency of annotation/search due to a large number of irrelevant terms included in the original IFC specification. Therefore, the contents in IFC that are relevant to AEC products should be considered for the ontology. In this paper, only the entities inherited from *IfcElement* (e.g. walls, beams and doors) are kept, where *IfcElement* (as a subtype of *IfcProduct*) defines physical objects that make up AEC products. In contrast, other entities not inherited from *IfcElement* (e.g. the entities inherited from *IfcProcess* and *IfcRelationship* define the process and relationships) are not relevant to our annotation purpose, and therefore they are not considered in the scope of the ontology.

In addition, we also consider the property sets related to the entities inherited from *IfcElement*. The official *Property Set Definition* (PSD)[1] is provided in the XML format (i.e. PSD-XML) for defining some additional properties and property sets outside of the IFC specification. The PSD includes numerous alphanumeric attribute definitions attached to building elements, spaces and other components. For example, the properties relating to the entity *IfcDoor* include "Acoustic Rating", "Thermal Transmittance", "Status", "Reference", "Fire Rating", etc.

Furthermore, we also add some user-defined properties (e.g. "classification", "manufacturer", "source" and "format") for further annotation applications (see Section 3.1).

[1] http://www.buildingsmart-tech.org/specifications/pset-releases.

#### 2.1.2. The representation of the ontology

Domain ontologies have a variety of representative forms, among which OWL (Ontology Web Language) [35] is selected as the modeling language in this study. With syntax extensions from the RDF (Resource Description Framework) – a lightweight representation of data and knowledge, OWL has been proposed by W3C as the ontology language of Semantic Web. The IFC specification is represented as a schema in the EXPRESS language defined in ISO 10303-11 [36]. In this section, we aim to convert the IFC EXPRESS schema to OWL ontology by a mapping between EXPRESS and OWL, so that the ontology can be used for annotation and retrieval.

There have been some existing efforts in the conversion of the IFC schema into OWL [30,37,38]. In particular, Beetz et al. [30] introduced a semiautomatic method for converting the IFC EXPRESS schema to OWL, which is a complete and direct mapping between two languages. Terkaj and Šojić [37] presented an enrichment of the EXPRESS to OWL conversion patterns with OWL class expressions that specifically capture certain constraints of the IFC standard. Pauwels and Terkajt [38] developed an OWL DL (Description Logics) profile in order to enable reasoning. Unlike the above efforts, we develop a lightweight ontology in order to support semantic annotation of BIM product documents.

For EXPRESS simple types (including String, Integer, Real, Binary and Boolean), they are equivalent to OWL (XSD types) and therefore are converted directly. For instance, String type is directly mapped into `xsd:string`. For EXPRESS entities, they are translated to `owl:Class`, and their subtypes and supertypes are converted to the OWL inheritances. The OWL data property is used to represent EXPRESS simple attributes, and the OWL object property represents EXPRESS named attributes. The OWL cardinality is used to denote EXPRESS attribute's optional flag. EXPRESS inherited attributes from supertype entities can be renamed according to the user's needs. Regarding EXPRESS inverse attributes which point to the related entities, they are translated to the OWL inverse property. EXPRESS enumeration types and Select types are mapped to the OWL clauses `owl:oneOf`.

#### 2.1.3. Improving human readability

The IFC schema is in a formal machine-readable notation (but not for human readability), which follows a specially 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). If the IFC schema is directly translated into the ontology, the translated terms cannot directly use the plain English words. In fact, the items in most existing IFC ontologies are still following the direct naming convention [30,32]. Therefore, the translated terms should be further refined by the process of segmenting the names of entity, removing the prefix and eliminating redundancy, for our annotation applications. This process is semi-automatically done in our work.

#### 2.1.4. Overview of the ontology for annotation

The constructed ontology for semantic annotation is organized in a tree-like data structure rooted from *IfcElement*, in which each node denotes a concept and each arc represents a relationship between concepts. The relationships include the inheritance and the type enumeration. For example, there is an inheritance relationship between *IfcBuildingElement* and *IfcCovering*, and there is a type enumeration relationship between *IfcCovering* and "Membrane".

Table 1 gives the outline of concepts in the resulting ontology. Fig. 2 shows a portion of class hierarchy of the ontology. In Fig. 2, the *Covering* inherited from the class *Building Element* has a property *PredefinedType* mapped to *Covering Type Enumeration*. The *Element*, as the supertype of *Covering*, is the generalization of all components that make up an AEC product. In the IFC4 specification, the *Element* is the supertype of *Building Element*, *Civil Element*, *Distribution Element*, *Feature Element*, *Furnishing Element*, and *Geographic Element*, *Transport Element*, *Element Component*, *Element Assembly* and *Virtual Element*. The terminal nodes of the ontology are composed of the enumeration types of IFC concepts. For example, the enumeration types of the concept "Covering" include "Membrane", "Cladding", "Insulation", "Roofing", "Modeling", "Ceiling", "Flooring" and "Wrapping", while the enumeration types are also connected to their supertype (i.e. "Covering") to define the class hierarchy in the ontology.

### 2.2. Step 2: Word-level semantic annotation

The second step of our algorithm is to use a word-level annotation algorithm to handle syntax variations and semantic complexities of BIM documents. The algorithm consists of three parts: (1) syntactic preprocessing of BIM documents, (2) candidate concept identification and (3) word-sense disambiguation.

#### 2.2.1. Syntactic preprocessing of BIM documents

Before doing annotation, the document collection should be prepared in the syntactic preprocessing stage. In our system, a well-known Heritrix crawler is used to obtain the data set (i.e. the document collection of BIM product resources on the Web) for annotation. After the documents are crawled from the Web, they are firstly processed on a syntactic level as follows.

(1) Split the tokens at white-space. This makes paragraphs and sentences into the terms that can be processed by the following procedure.
(2) Divide the text at non-letter characters and lowercase, where all the characters should be stored in lowercase for convenience.
(3) Remove the special words called stop words which are assumed to carry very little stand-alone meaning, such as "a", "are", "by", "for", and "on".

**Table 1**
The outline of concepts in the resulting IFC concepts for annotation.

| Taxonomies | #concepts | Examples of concepts | Acquisition resources |
|---|---|---|---|
| Building element | 33 | Door | *IfcBuildingElement* |
| Civil element | 1 | Civil element | *IfcCivilElement* |
| Distribution element | 77 | Pump | *IfcDistributionElement* |
| Element assembly | 1 | Element assembly | *IfcElementAssembly* |
| Element component | 11 | Fastener | *IfcElementComponent* |
| Feature element | 8 | Opening element | *IfcFeatureElement* |
| Furnishing element | 2 | Furniture | *IfcFurnishingElement* |
| Geographic element | 1 | Geographic element | *IfcGeographicElement* |
| Transport element | 1 | Transport element | *IfcTransportElement* |
| Virtual element | 1 | Virtual element | *IfcVirtualElement* |

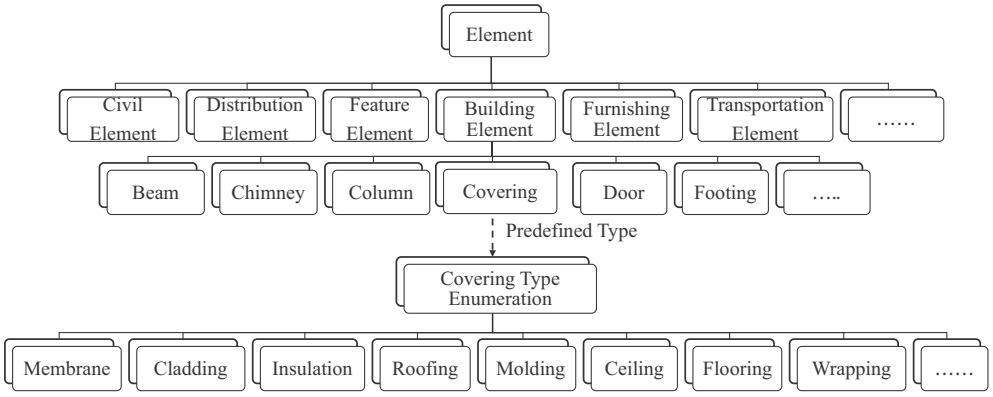


**Fig. 2.** Part of the classes and class hierarchy in IFC IR ontology.

(4) Tokenize the terms based on sophisticated grammar. This aims to recognize the special structures like email addresses, acronyms and alphanumeric.
(5) Stem and transform the terms in order to reduce any of the forms of a word, such as "doors" to its elemental root "door".
(6) Store the terms and their associated frequency and positions in the documents.

*2.2.2. Candidate concept identification*

After the document collection is syntactically processed, the candidate concepts can be identified according to the IFC ontology. The tokens of documents are extracted and looked up in the IFC ontology. If a term is matched with a concept in the IFC ontology, the concept is identified, which will be indexed and linked to the URL in the IFC website. However, the IFC specification only contains the limited terms used in the AEC products, which is not sufficient for the document annotation purpose. Therefore, some general ontologies should be combined with the IFC ontology for a wider term identification.

WordNet [25], as a general ontology, is the most commonly used lexical database for the English language. In this paper, we use WordNet synonym sets (i.e. *synsets*) for identifying the corresponding IFC concepts. For example, the term "covering" in WordNet has several synonyms (including "cover", "screening", "masking", "coating" and "application"), so each of those synonyms appearing in a document is identified as a candidate concept of the IFC concept "Covering" (i.e. *IfcCovering*). Those candidate concepts will be disambiguated in the following step.

*2.2.3. Word-sense disambiguation*

One limitation of simple word-level mapping is that contextual information is not utilized, which may result in incorrect mapping between the terms of documents and concepts of IFC ontology. In WordNet, a term may have multiple meanings, while not all the meanings should be annotated with the corresponding concepts of the term in the IFC ontology. Therefore, a proper disambiguation process for the ambiguous terms becomes necessary. For example, the term "column" in WordNet has several synonyms such as "pillar" and "editorial". However, the synonym "editorial" means an article giving opinions or perspectives in the newspaper, which should not be annotated with the IFC concept "Column" (i.e. *IfcColumn*) in BIM-specific domain.

Here, we use a strategy to disambiguate the word-sense based on local context analysis to reduce incorrect mappings. If a term in a document has the same meaning with an IFC concept, the IFC concept (or its related IFC concepts) is more likely to co-occur with the term in the local context. This strategy is similar to the Local Context Analysis (LCA) method [39,40] which disambiguates the word-sense between the query words and the terms in the local context. Based on the above strategy, we also develop a context-based similarity measure between a term in a document and an IFC concept, which is given below.

$$Sim(t,c) = \frac{\sum_{c_i \in O(c)} TF(c_i) IDF(c_i)}{|O(c)|}, \tag{1}$$

where

$$TF(c_i) = \frac{n_i}{\sum_k n_k} \tag{2}$$

and

$$IDF(c_i) = \log \frac{|D|}{1 + |d \in D : t \in d|}. \tag{3}$$

In Eq. (3), $t$ is a term in the target document, $c$ is a candidate IFC concept corresponding to the term $t$, $O(c)$ is the related concepts of $c$ in the IFC ontology (e.g. subtype, supertype, etc.), $c_i$ is one of concepts in $O(c)$, $n_i$ is the term frequency of $c_i$ in the document, $n_k$ is the number of $k$th term in the document. $|D|$ is the total number of documents in the corpus, $|d \in D : t \in d|$ is the number of documents where the term $t$ appears. In Eq. (3), $TF(\cdot)$ is the term frequency in document, and it rewards the concepts co-occurring frequently with the term $t$. $IDF(\cdot)$ is the inverse document frequency, and it penalizes the concepts occurring frequently in the collection. Finally, $Sim(t,c)$ is compared with a predefined threshold to decide if the term $t$ has strong enough relationship with its candidate concept $c$. In our experiment, the threshold is typically set as 0.01.

*2.3. Step 3: Document-level semantic annotation*

As the most direct way to annotate a corpus, the word-level annotation algorithm introduced in Section 2.2 enables some applications to access every composing unit in the corpus. However, the word-level annotation based on thesaurus or an ontology still lacks the ability to understand the document as a whole. This might be an issue in many semantics-oriented categorization, information extraction and retrieval applications. Instead, the document-level semantic annotation tries to analyze a document as a whole, which can provide more contextual information for some applications such as retrieval, navigation and document classification.

In this paper, we apply latent semantic analysis (LSA) [14] technique to the document-level semantic annotation for BIM documents. The LSA is a technique for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. In the LSA, documents are mapped

to a vector space of reduced dimensionality through singular value decomposition (SVD). In this way, the similarity between documents is kept in the latent semantic space, while terms that are close in meaning will be mapped to the similar positions in the space.

Although the LSA is able to compute the similarity between documents in the latent semantic space, it lacks the ability to interpret the semantics of the documents explicitly. In many specific domains, some commonly used semantic resources (e.g. IFC in the AEC domain) could be used to interpret the semantics of the documents with standardized concepts. In this work, we incorporate the concepts in the IFC ontology with the LSA technique for achieving the document-level semantic annotation process, which is given as follows.

*2.3.1. Constructing the term-document matrix*

As for many vector space information retrieval models, the LSA utilizes the term-document association structure such as a term-document matrix. The term-by-document matrix is constructed in the syntactic preprocessing stage of word-level annotation (see Section 2.2.1). Every document vector represents a document in a very high dimensional vocabulary space. It captures weighted vocabulary distribution patterns of the document. The term-document matrix is defined as

$$A = (f_{ij})_{m\times n}, \tag{4}$$

where $f_{ij}$ represents the frequency of the term $t_i$ in the document $d_j$.

*2.3.2. Dimensionality reduction of the matrix*

The matrix $A$ in Eq. (4) can also be represented by the truncated singular value decomposition (SVD), where the SVD of the original term-document matrix can be written as

$$A = U\Sigma V^T, \tag{5}$$

where $A$ is the $m \times n$ term-document matrix, $U$ is the $m \times m$ orthogonal matrix whose columns define the left singular vectors of $A$, $V$ is the $n \times n$ orthogonal matrix whose columns define the right singular vectors of $A$, and $\Sigma$ is the $m \times n$ diagonal matrix containing the nonnegative singular values of $A$ in descending order along its diagonal. The $k$-dimensional reduced-rank approximation of $A$, denoted by $A_k$, is constructed by setting all but the $k$-largest singular values of $A$ equal to zero so that

$$A_k = U_k\Sigma_k V_k^T, \tag{6}$$

where $U_k$ and $V_k$ comprise the first $k$ columns of $U$ and $V$, and $\Sigma_k$ contains the $k$-largest singular values of $A$. Using the components of $A_k$, all terms and documents can be encoded as vectors in the $k$-dimensional space. For example, the $j$th term vector is $t_j = \Sigma_k U_k^T e_j$, and the $j$th document vector is $d_j = \Sigma_k V_k^T e_j$, where $e_j$ denotes the $j$th canonical vector of appropriate dimension.

*2.3.3. Mapping IFC concepts to the latent semantic space*

The annotation process with the LSA is accomplished by representing each IFC concept by its corresponding IFC4 document (the HTML version) on the buildingSMART official website [13]. As a result, an IFC concept becomes a vector denoted by $C$, which consists of the terms of the document.

In Eq. (6), $U_k$ denotes the $k$-dimensional semantic space of the terms, and $V_k$ denotes the $k$-dimensional semantic space of the documents. By mapping the term vector $C$ into the semantic space using the "fold-in" technique [14], we get the corresponding vector $C'$ in its $k$-dimensional latent semantic space, i.e.

$$C' = C^T U_k \Sigma_k^{-1}. \tag{7}$$

*2.3.4. Identifying the closest concept*

Let $D_j$ denote the $j$th document and $C_i$ the $i$th IFC concept, respectively. The similarity between $D_j$ and $C_i$ can be computed with the cosine of the angle between two vectors, i.e. the document vector $U_j = (u_{1j}, u_{2j} \dots u_{kj})^T$ and the concept vector $C_i' = (c_{1i}', c_{2i}' \dots c_{ki}')^T$, respectively, in the $k$-dimensional latent semantic space. The similarity is defined by

$$Sim(D_j, C_i) = \frac{\sum_{l=1}^{k} c_{li}' u_{lj}}{\sqrt{\sum_{l=1}^{k} (c_{li}')^2}\sqrt{\sum_{l=1}^{k} (u_{lj})^2}}. \tag{8}$$

Consequently, we can obtain the similarity values between each BIM document and all IFC concepts. In our implementation, the concepts with the greatest similarity values are selected to annotate the documents.

### 2.4. Step 4: Concept indexing

Finally, semantic indexing can be further conducted for the annotated BIM documents by the terms and concepts discovered from the documents. The concepts are considered to have the higher semantic importance than the ordinary linguistic terms within the document. There are several open source search engines supporting document indexing. In this work, we adopt Apache Lucene [41] to index documents. The Lucene is a free open source information retrieval software library, which provides high-performance indexing creation and efficient search algorithms. The Lucene uses an inverted index, which is created based on statistical information of document collections, such as term frequency, document frequency and term position.

In our implementation, the IFC concepts obtained by the word-level annotation are indexed in a new field *concept*, besides the existing fields (*id*, *url*, *title* and *content*) in the Lucene. The new *concept* field is assigned with the higher *boost* values which denote the weights of the items in the ranking process. The IFC concepts obtained by the document-level annotation are saved as a new field *classification* in the indexing process. The other user-defined properties are extracted from the web page directly, which are also saved in the new fields (e.g. *manufacturer*, *source* and *format*). These new fields receive the additional field *boost* in the ranking process.

Furthermore, the generated index can be utilized in concept-based information retrieval application for improving the retrieval performance. This application and evaluation will be given in the next section.

## 3. Experimental results and applications

### 3.1. System overview

Based on the proposed method, we have developed a prototype semantic annotation system, named BIMTag, for online BIM product documents. The BIMTag system provides the functions of semantic annotation, indexing and retrieval. The IFC ontology and other general ontologies (e.g. WordNet) are kept in the semantic repositories in the system. The Lucene [41] information retrieval engine has been adopted to index documents and measure the similarity, along with tokens and stems.

Fig. 3 shows a screenshot of the web interface of the BIMtag system. There are two ways to access the annotated BIM documents as follows.

- **Access by querying**. When the user inputs a specific query (e.g. "door"), the system will run a query expansion algorithm [15] for retrieving online BIM documents. Then, the system automatically ranks these documents and displays the search results

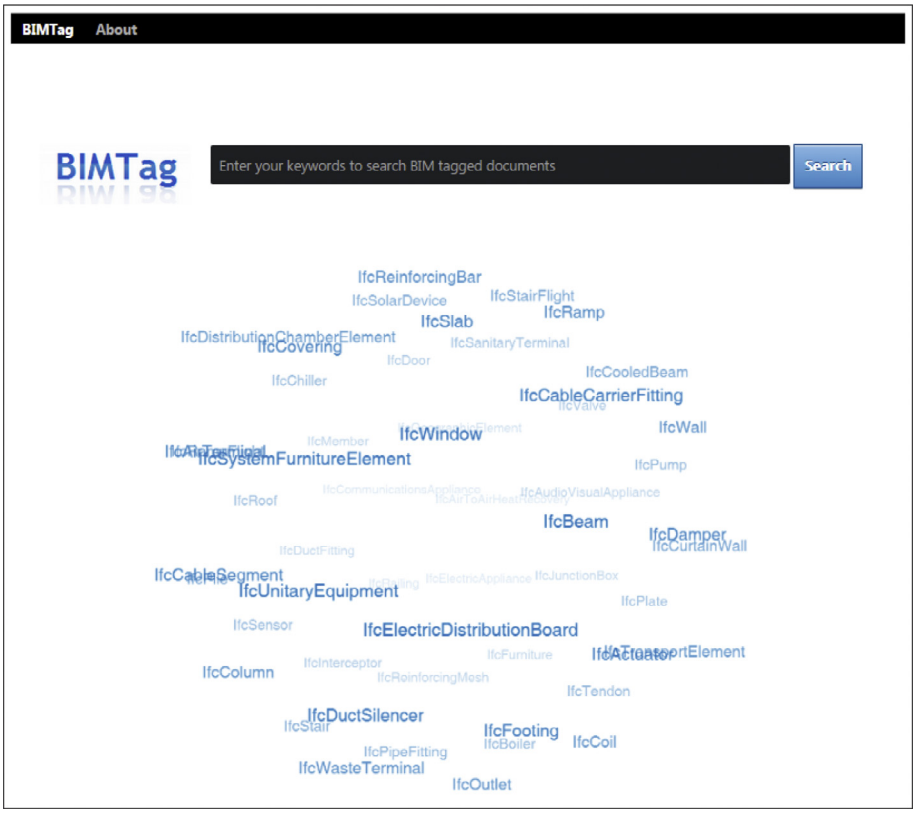


**Fig. 3.** A screenshot of the web interface for the BIMTag system. The user can access the annotated BIM documents by searching with a query or clicking the words in the tag cloud.

(see Fig. 4). By clicking on one of the search results, the system guides the user to the annotated web page (see Fig. 5) rather than the original web page.

- **Access by a tag cloud**. Alternatively, the user can click the words in the tag cloud on the bottom. The tag cloud is a visual representation for text data, which is usually used as website navigation aids. In Fig. 3, each tag is a single term (e.g. *IfcWindow*) in the IFC specification, and it is hyperlinked to the annotated BIM documents associated with this tag.

Fig. 5 shows the annotated web page, which corresponds to the 1st ranked search result (i.e. "Inswing French Door") in Fig. 4. In Fig. 5, the annotated terms are highlighted in the content of the document, and they are hyperlinked to the corresponding concepts of the IFC4 specification. When the mouse cursor is hovered over a highlighted term, a tooltip is popped up to show the IFC concept associated with this term. By clicking on a highlighted term in the document, a popup window displays the external semantic description of the corresponding IFC concept on the buildingSMART official website [13] (see the additional small panel "Linked IFC concept" in Fig. 5).

For example, the highlighted term "door" in the document is associated with the IFC concept "Door" in our ontology, which is hyperlinked to the external specification of *IfcDoor* at http://www.buildingsmart-tech.org/ifc/IFC4/final/html/link/ifcdoor.htm (see the popup window in Fig. 5).

The IFC schema is a data model in a formal machine-readable notation, which is written in EXPRESS with a file extension ".exp" (e.g. "IFC2X3_TC1.exp"). The IFC specification consists of such a schema and associated informal human-readable semantic definitions (e.g. the HTML documentation being available at the buildingSMART). The associated semantic definitions on the buildingSMART are human-readable, which cover the definitions of all IFC entities including the natural language names, entity definition, attribute definitions, formal propositions, entity inheritance, attribute inheritance, and so on. In our work, the linked documents on the buildingSMART are used in the process of document annotation with the LSA, where each IFC concept is represented by its corresponding document (see Section 2.3.3). As a result, an IFC concept becomes a vector, which consists of the terms of the document.

In Fig. 5, the right panel is divided into two tabs, which display the IFC concepts discovered by the word-level annotation and the document-level annotation, respectively. The two tabs are illustrated as follows. The bottom tab (see "Word-level annotation') shows a list of IFC concepts which are recognized in the current document by the word-level annotation. The recognized IFC concepts are sorted by the frequency of appearance in the document. As for this example, the concept *IfcDoor* is ranked in the first place, which appears 5 times in the current document.

The top tab (see "Document-level annotation") shows some tags of user-defined properties (including "classification", "manufacturer", "source" and "format"), which are extracted from the current document using the document-level annotation. The user-defined properties were added in the ontology in Section 2.1.1. In Fig. 5, the tag of "classification" corresponds to the IFC concept (e.g. *IfcDoorStandardCase*), which has the greatest similarity value with the current document using Eq. (8). The other tags (i.e. "manufacturer", "source" and "format") are directly extracted from the current website. As for this example, the tag of "manufacturer" is "Integrity from Marvin", the tag of "source" is "seek.autodesk.com", and the tag of "format" is "RFA" which means that the web page contains BIM models in the Revit file format ("*.RFA").

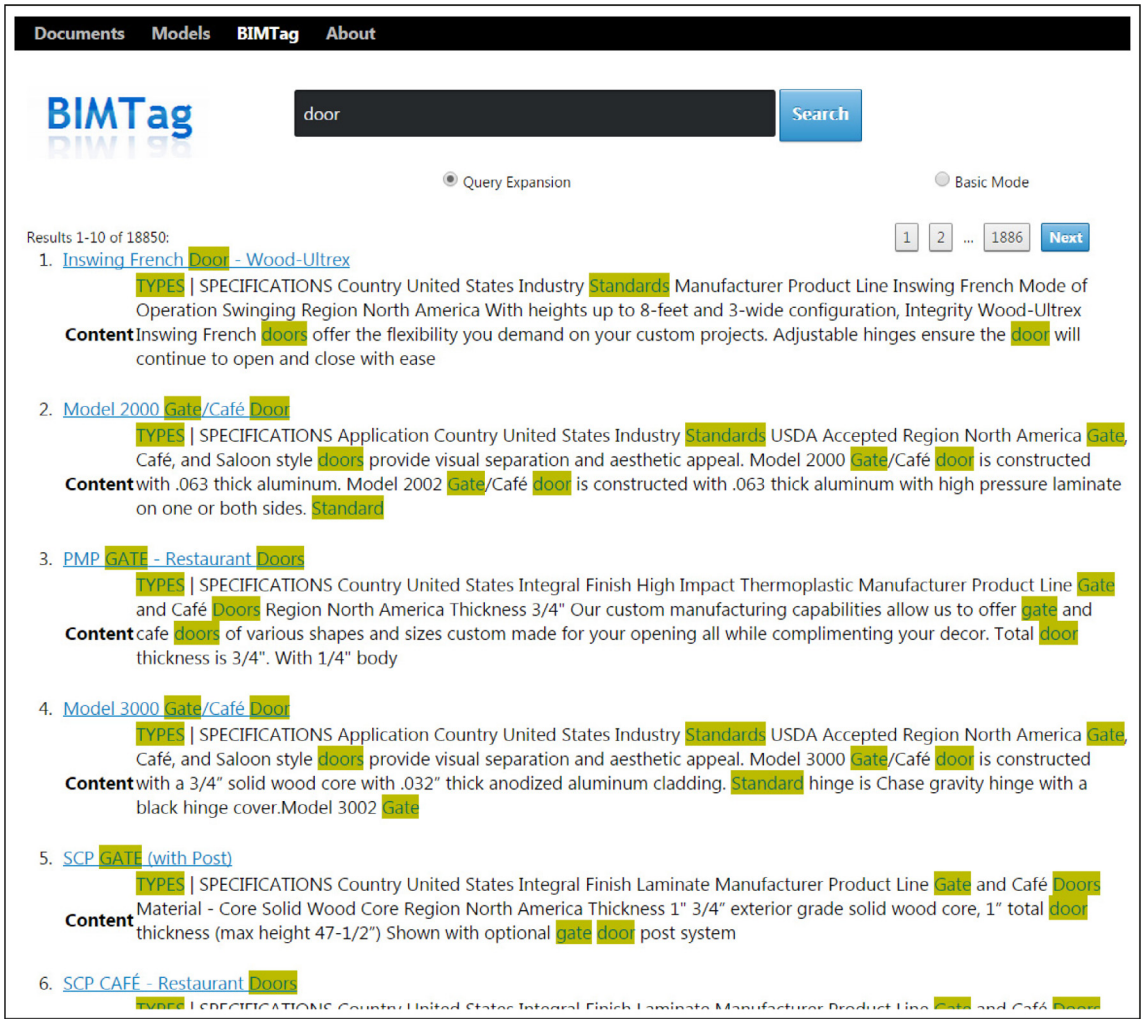


**Fig. 4.** A screenshot of the search results with a user's query (e.g. "door"). By clicking on one of the search results, the system guides the user to the annotated web page (see Fig. 5) rather than the original one.

### 3.2. Evaluation

In order to evaluate the performance of the proposed method, we run an information retrieval experiment on a document collection with or without our semantic annotation. The ranking process uses the Vector Space Model (VSM) [42] to determine how relevant a given document is to the user's query. Here, the 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 experimental system, the toolkit OWLAPI, a semantic system framework, is used for handling the ontology and OWL language. The indexing and ranking are offered by Apache Lucene, and the Heritrix Web Crawler is used to collect online BIM documents. In this section, all the experiments are run on a 2.93 GHz processor with 8 GB memory under Windows 7.

Currently, the document collection used in our experiment contains the 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 online BIM resources. 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". Therefore, we conceive of a test query "Covering", and judge whether each web page in the search results corresponds to "23.35.00.00". There are 30 test queries used in our experiment, as shown in Table 2.

To measure the performance of our method, we adopt the standard evaluation procedure from information retrieval, namely *precision-recall* curves, for evaluating the retrieval results [43,44]. 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 $Relevant \cap Retrieved$. Then the recall is defined as

$$\frac{Relevant \cap Retrieved}{Relevant},$$

and the precision is defined as

$$\frac{Relevant \cap Retrieved}{Retrieved}$$

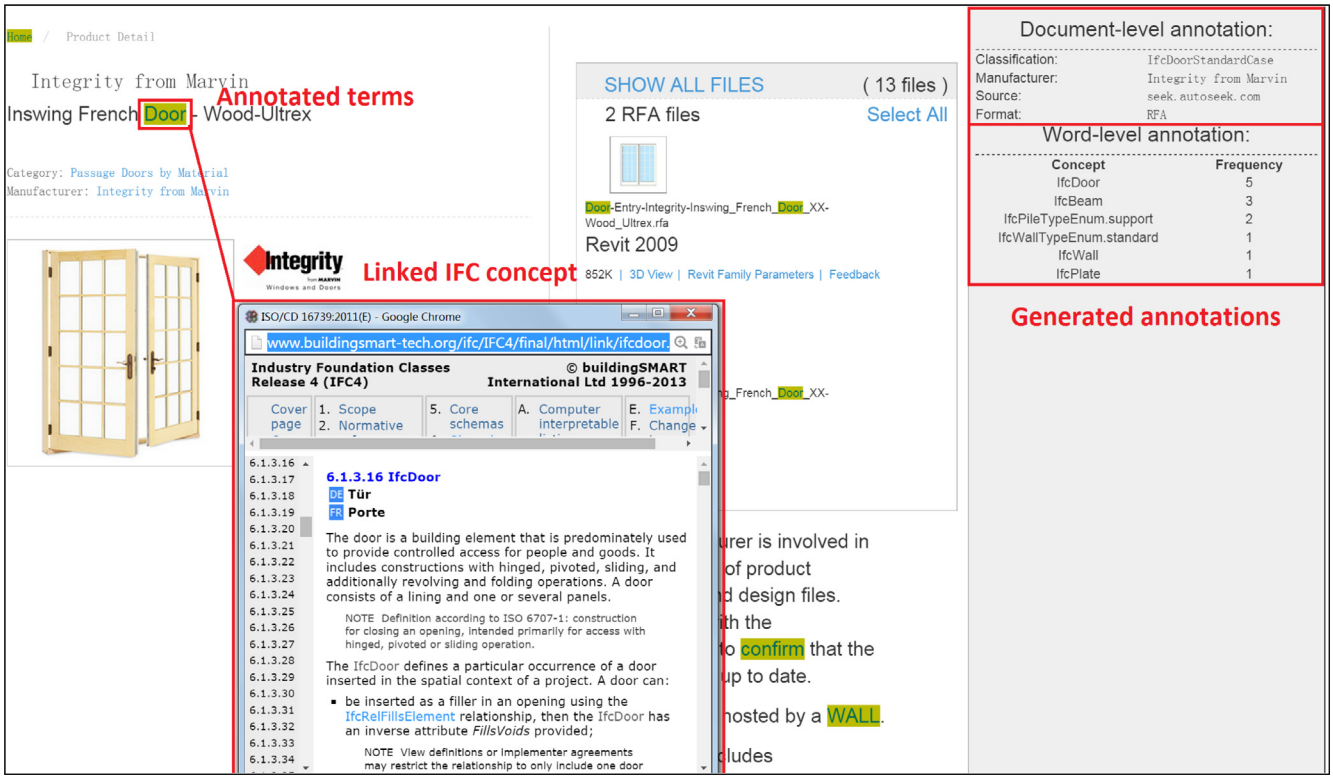


**Fig. 5.** Illustration of the annotated BIM document at the BIMTag system, which provides an easy way to learn semantic annotation to the end user. The annotated terms are highlighted in the content of the document, and they are also hyperlinked to the corresponding concepts of the IFC4 specification on the buildingSMART [13] (see the additional small panel "Linked IFC concept"). Moreover, the right panel shows the concepts discovered by the word-level annotation and the document-level annotation, respectively.

**Table 2**
Test queries used in our experiment.

| | | |
|---|---|---|
| Beam | Curtain Wall | Railing |
| Covering | Column | Door |
| Roof | Slab | Wall |
| Window | Roof | Fan |
| Lamp | Energy Conversion Device | Coil |
| Cooling Tower | Electric Generator | Evaporative Cooler |
| Heat Exchanger | Damper | Fan |
| Air Terminal | Lamp | Outlet |
| Sanitary Terminal | Furnishing Element | Furniture |
| Transport Element | Pipe Fitting | Valve |

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.

In the experiment, we compare our method with several retrieval methods, including the search without semantic annotation (i.e. native Lucene), the search with semantic annotation based on WordNet (synonyms, hyponyms and hypernyms, respectively). Fig. 6 shows the average precision-recall curves. The results show that our method performs better retrieval than both the search without semantic annotation and the search with WordNet-based semantic annotation.

In addition to the precision-recall curves, we also compute the *F*-measure [45] for evaluating the retrieval results. The *F*-measure (also *F*-score) is a measure of test accuracy, and it is the harmonic mean of precision and recall. The *F*-measure is defined as

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

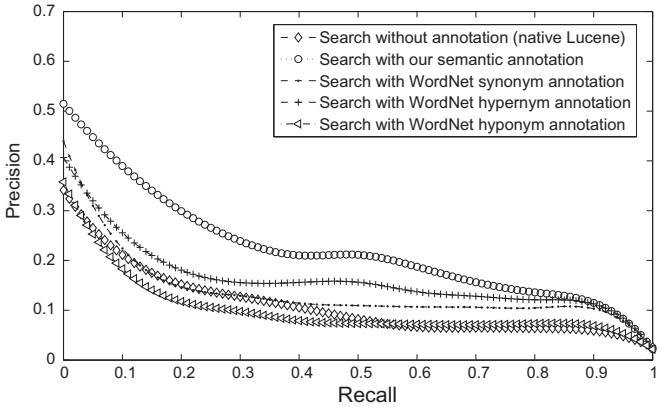


**Fig. 6.** The average precision-recall curves of retrieval results with the proposed annotation method and other annotation strategies.

It measures the effectiveness of retrieval with respect to a user who attaches times as much importance to recall as precision. $\beta$ 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.

Table 3 gives the *F*-measure values of our method and other methods. The results suggest that our method outperforms other methods.

**Table 3**
*F*-measure of our method and others.

| Methods | *F*-measure |
|---|---|
| Search without annotation (native Lucene) | 0.180093695 |
| Search with WordNet synonym annotation | 0.162367191 |
| Search with WordNet hypernym annotation | 0.189367722 |
| Search with WordNet hyponym annotation | 0.124649668 |
| Search with our semantic annotation | 0.2257648 |

### 3.3. Applications

The automatic semantic annotations enable various applications such as highlighting, indexing, categorization, information extraction and retrieval. In this paper, we combine the proposed semantic annotation method with a search engine, named BIMSeek, for retrieving online BIM product resources. The BIMSeek contains two individual modules: *BIM document* search module and *BIM model* search module. The former focuses on the application of retrieving online BIM documents, while the latter deals with the application of searching online BIM models.

#### 3.3.1. Document search module in BIMSeek

Fig. 7 shows the screenshot of document search module in BIMSeek. In this module, we implement the keyword-based search based on annotated BIM documents, which utilizes the word-level annotation algorithm proposed in Section 2.2. Based on the synonyms, the inheritance and enumeration relationships that are explicitly defined in our ontology, the terms in each document are mapped to their corresponding IFC concepts.

As for the example in Fig. 7, the concept "covering" (*IfcCovering*) has some enumeration types (including "membrane", "cladding", "insulation", "roofing", "modeling", "ceiling", "flooring" and "wrapping") in the ontology. Therefore, the document's terms matched to the above enumeration types will be annotated with their corresponding concept "covering". In Fig. 7, the 1st ranking result which include the terms "membrane" and "floor" is returned for the query "covering".

#### 3.3.2. Model search module in BIMSeek

The document-level annotation can be used to extract the global or general information of BIM documents, e.g. extracting some tags of classification, manufacturer, source and format on the website. This can be used to discover the *topic* and some metadata of the documents.

Fig. 8 shows the screenshot of model search module in BIMSeek. In this module, our word-level annotation method is applied to extract the tags of some user-defined properties from the associated product documents. The extracted tags provide a convenient way for users to browse online BIM resources in terms of classification, manufacturer, etc.

As for the example in Fig. 8, the search results with a query "window" are returned, where the extracted tags are highlighted. Especially for the 2nd search result, the tag "classification" is "IfcWindow" which is recognized in the document-level annotation process, the tag "manufacturer" is "CGI Windows and Doors, Inc.", the tag "source" is "www.arcat.com", and the tag "format" is "RFA" which means that the web page contains BIM models in the Revit file format ("*.RFA"). The user can utilize the extracted tags to seek or browse the resources under certain tag conditions.

#### 3.3.3. Connecting BIMSeek to design platform

In order to apply BIMSeek to find appropriate BIM resources for design reuse, we develop a plug-in for the design platform *Autodesk Revit 2014*, as shown in Fig. 9. The developed Revit plug-in allows the user to directly search for online BIM resources in the design stage. When a BIM resource is located in BIMSeek, one

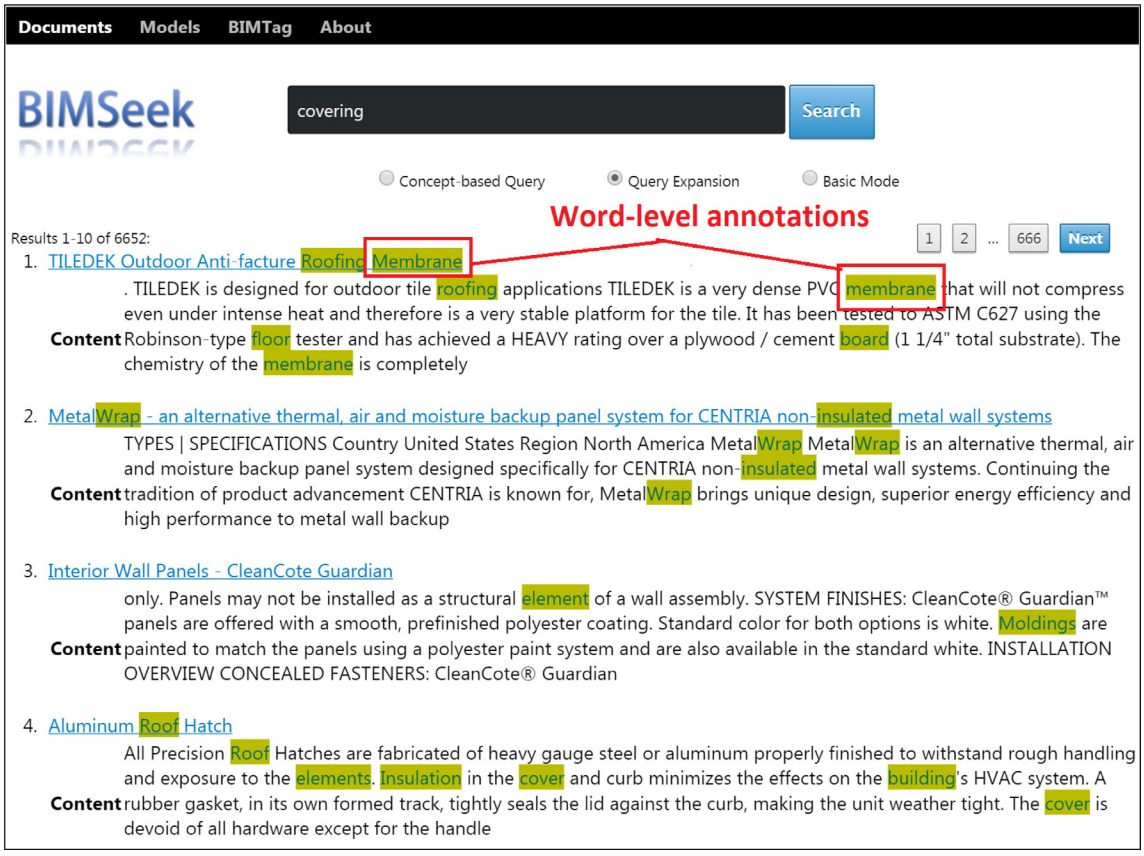


**Fig. 7.** The screenshot of document search module in BIMSeek, where the input query is "covering".

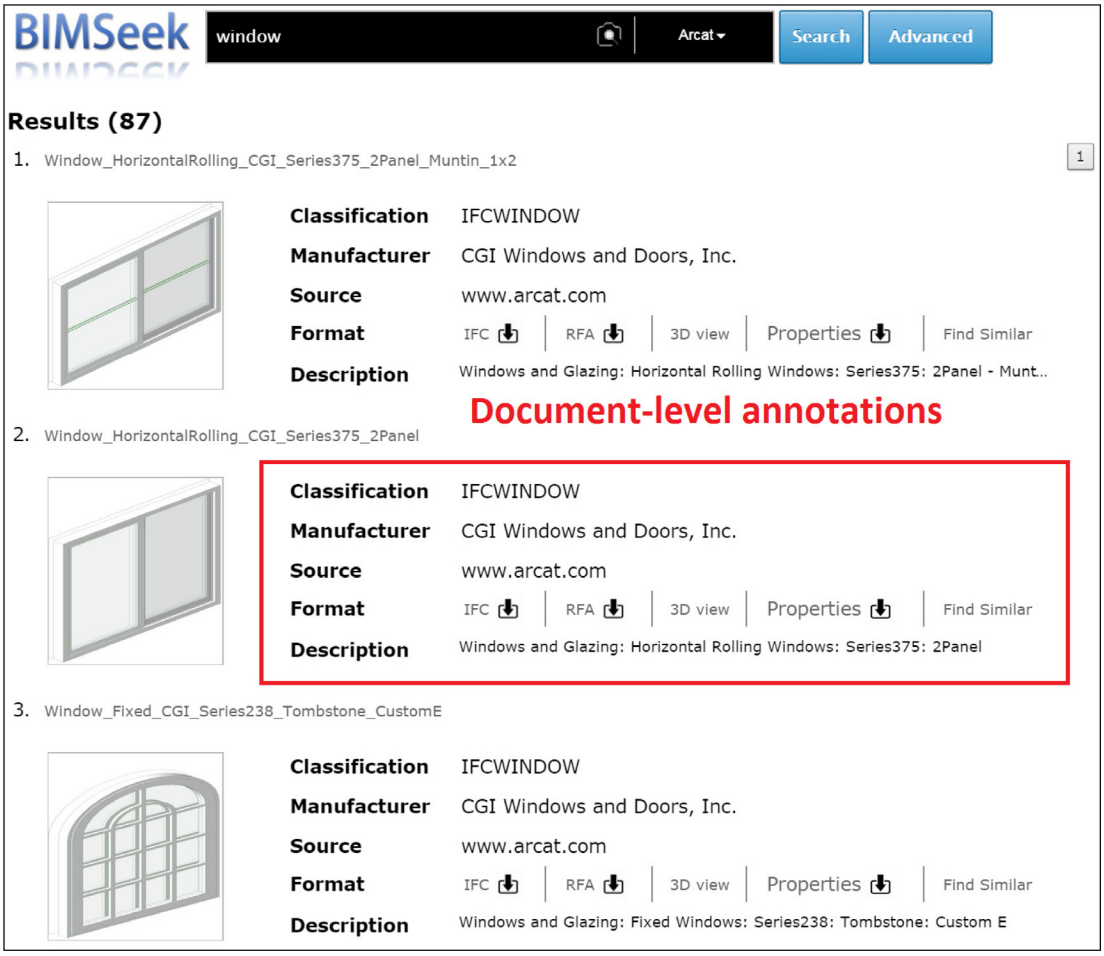


**Fig. 8.** The screenshot of model search module in BIMSeek, where the input query is "window". The tags are extracted from the associated product documents.

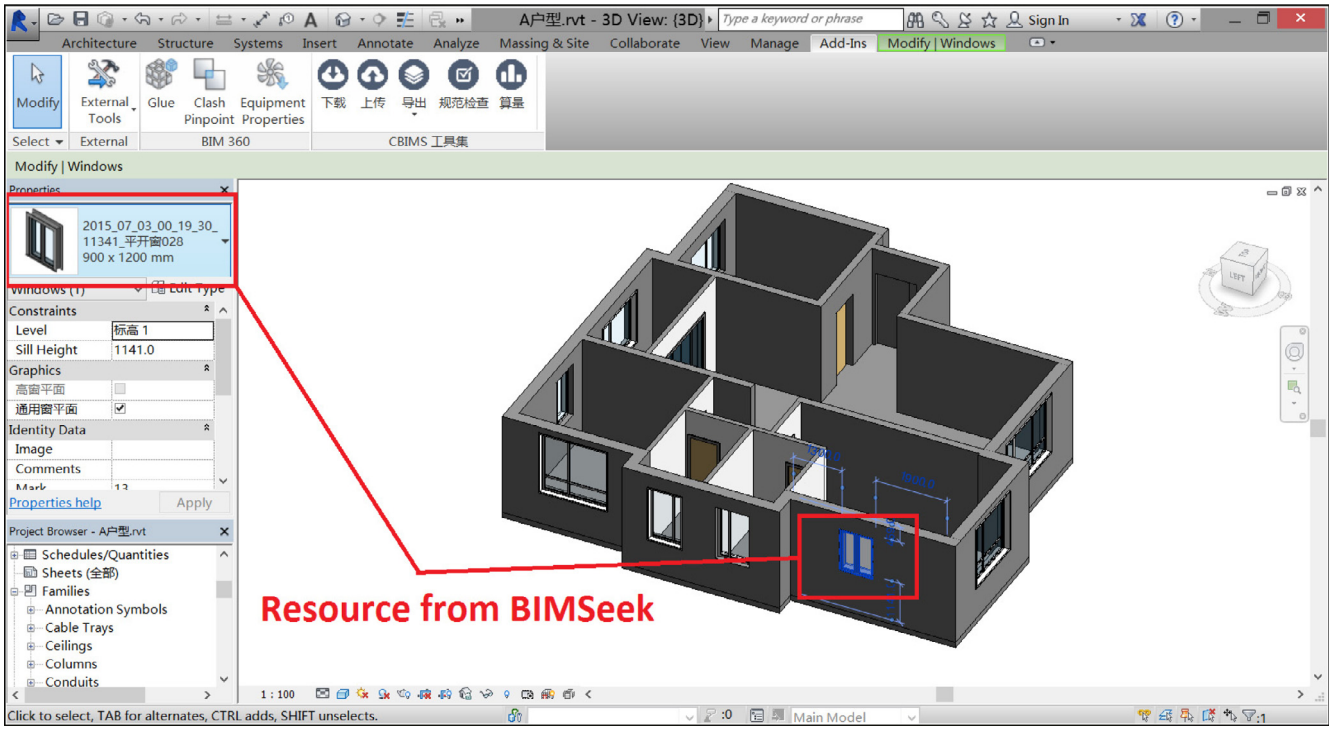


**Fig. 9.** A plug-in for connecting the BIMSeek system to the *Revit* platform.

can directly download the relevant BIM model from the BIMSeek system. Then the plug-in can automatically load the downloaded file into the *Revit* platform. In this way, online BIM resources from various websites can be connected to design platform for design reuse.

In Fig. 9, the highlighted window component is found and downloaded from the BIMSeek system (corresponding to the RFA file of the 2nd search result in Fig. 8), and then the downloaded file is automatically loaded into the *Revit* platform through the plug-in. Finally, the window component is inserted into the architectural model for design reuse.

## 4. Conclusion and discussion

In this section, we conclude this paper, discuss the differences between ifcOWL and our IFC ontology, and give the limitations and future work.

### 4.1. Conclusion

This paper first develops a domain-specific ontology based on the IFC spcification, which encodes the knowledge in the AEC field for semantic annotation. With reference to the built ontology, a concept-based automatic semantic annotation method is proposed for online BIM product documents. The proposed method consists two stages including the word-level annotation and the document-level annotation. In the word-level annotation, the word-sense disambiguation based on context analysis is applied to derive the intended meaning of terms and their underlying concepts. In the document-level annotation, a method based on latent semantic analysis technique is proposed to extract the document-level semantics in whole.

The results suggest that the word-level annotation is able to understand the terms in the documents, while the document-level annotation is able to understand the documents. The proposed semantic annotation method can be used for various applications such as highlighting, indexing, categorization, information extraction and retrieval. Finally, a prototype system of semantic annotation, named BIMTag, is developed and applied to a search engine for demonstrating the effectiveness of our method. The experimental results show that the search engine with our annotation method can achieve the retrieval for BIM product resources better than other annotation methods.

### 4.2. Discussion

Several recent studies [30,37,38] have conducted the conversion of the IFC schema into OWL, where their purpose was to construct the complete and direct mapping between EXPRESS and OWL towards the formalization and standardisation direction. The early conversion proposal by Beetz [30] was named ifcOWL, while the recent effort presented by Pauwels and Terkajt [38] was an OWL DL (Description Logics) profile in order to enable reasoning.

Unlike their purpose, we focus on how to construct a light-weight IFC ontology for our particular application, i.e. semantic annotation and retrieval of online BIM product documents. Therefore, during the IFC ontology construction process, we follow several principles which are different with previous ifcOWL construction.

- Keep the IFC ontology as simple and usable. The IFC ontology is based on the simple constructs of OWL-Lite (except the *oneof* construct). The simplicity and usability are the primary consideration of the ontology rather than the usage of reasoning engines. In this paper, only the entities inherited from *IfcElement* are considered, which are related to AEC products.
- Keep only part of the object inheritance and property definition from the original IFC EXPRESS schema, which makes the IFC ontology as small as possible. In order to wider synonym extension, we combine the IFC ontology with general ontologies (e.g. WordNet [25]).
- In addition to the IFC schema, we also make use of the official Property Set Definition (PSD) for defining some additional properties which are outside of the IFC specification. Alternatively, some user-defined properties can be defined in the ontology for customizing information extraction.

### 4.3. Limitation and future work

One of the limitations of our method is that the single IFC ontology cannot fully cover the needs of semantic annotation in BIM-specific domain because of its multidisciplinary and multi-stakeholder nature. Therefore, there is a need to combine or merge the IFC ontology with some existing AEC ontologies such as ISO 12006-2, Uniclass and OmniClass [46], so that more extensive BIM resources can be used for semantic annotation. In the future, we would like to integrate more AEC ontologies related to BIM into the semantic annotation system.

In the current implementation, only the two relationships of inheritance and enumeration are used in the IFC ontology. In our ongoing work, we are trying to use the properties and restrictions in the IFC specification for ontology construction and semantic annotation. Furthermore, some automatic semantic analysis techniques like explicit semantic analysis (ESA) [47] might be applied to enhance semantic annotation and retrieval, by using massive human knowledge repositories such as Wikipedia. This could be one of future work. Finally, annotation and retrieval of various kinds of BIM-related documents like BIM design documents, product specifications and COBie documents are also very attractive. Thus, future research may extend our current study to various BIM-related documents where broader and more diverse domain ontologies and document collections exist.

## Acknowledgements

The authors appreciate the comments and suggestions of all anonymous reviewers, whose comments significantly improved this paper. The research is supported by the National Science Foundation of China(61472202, 61272229) and the National Technological Support Program for the 12th-Five-Year Plan of China (2012BAJ03B07). The last author is supported by the National Key Technologies R&D Program of China (2015BAF23B03).

## Appendix A. Supplementary material

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

## 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, second ed., John Wiley and Sons, NJ, 2011.
[2] Autodesk, Autodesk Seek, 2014. <http://seek.autodesk.com>.
[3] BIMobject, 2014. <http://bimobject.com/>.
[4] NBS of UK, National BIM Library, 2014. <http://www.nationalbimlibrary.com/>.
[5] SmartBIM LLC, SmartBIM, 2014. <http://www.smartbim.com/>.
[6] Pierced Media LC, RevitCity, 2014. <http://www.revitcity.com>.

[7] P. Demian, P. Balatsoukas, Information retrieval from civil engineering repositories: importance of context and granularity, J. Comput. Civ. Eng. 26 (6) (2012) 727–740.
[8] A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff, Semantic annotation, indexing, and retrieval, J. Web Semantics 11 (2) (2004) 49–79.
[9] F. Bueno, A. Garca-Serrano, J.L. Martnez-Fernndez, Enrichment of text documents using information retrieval techniques in a distributed environment, Expert Syst. Appl. 37 (12) (2010) 8348–8358.
[10] Q. Rajput, S. Haider, BNOSA: a Bayesian network and ontology based semantic annotation framework, J. Web Semantics 9 (2) (2011) 99–112.
[11] M. Mota, C. Medeiros, Introducing shadows: flexible document representation and annotation on the Web, in: 2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW), 2013, pp. 13–18.
[12] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna, Semantic annotation for knowledge management: requirements and a survey of the state of the art, J. Web Semantics 4 (1) (2006) 1570–8268.
[13] BuildingSMART, The IFC4 Standard, 2014. <http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-release>.
[14] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Soc. Inform. Sci. 41 (6) (1990) 391–407.
[15] G. Gao, Y.-S. Liu, M. Wang, M. Gu, J.-H. Yong, A query expansion method for retrieving online BIM resources based on Industry Foundation Classes, Autom. Constr. 56 (2015) 14–25. <http://cgcad.thss.tsinghua.edu.cn/liuyushen/ifcqe/>.
[16] K. Bontcheva, H. Cunningham, Semantic annotations and retrieval: manual, semiautomatic, and automatic generation,;;, in: J. Domingue, D. Fensel, J. Hendler (Eds.), Handbook of Semantic Web Technologies, Springer, Berlin, Heidelberg, 2011, pp. 77–116.
[17] T. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220.
[18] 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.
[19] Z. Li, V. Raskin, K. Ramani, Developing engineering ontology for information retrieval, J. Comput. Inf. Sci. Eng. 8 (1) (2008) 737–745.
[20] Y. Rezgui, Ontology-centered knowledge management using information retrieval techniques, J. Comput. Civ. Eng. 20 (4) (2006) 261–270.
[21] A. Weissman, M. Petrov, S. Gupta, A computational framework for authoring and searching product design specifications, Adv. Eng. Inform. 25 (3) (2011) 516–534.
[22] 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.
[23] Y. Li, Z. He, 3D indoor navigation: a framework of combining BIM with 3D GIS, in: 44th ISOCARP Congress, 2008.
[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] G. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[26] T. Lukasiewicz, U. Straccia, Managing uncertainty and vagueness in description logics for the semantic web, J. Web Semantics 6 (4) (2008) 291–308.
[27] K. Lin, L. Soibelman, Promoting transactions for A/E/C product information, Autom. Constr. 15 (6) (2006) 746–757.
[28] Industry Foundation Classes (IFC), IFC4 Release Candidiate 4, 2014. <http://www.buildingsmart-tech.org/ifc/IFC2x4/rc4/html/index.htm>.
[29] L. Zhang, R. Issa, Ontology based partial building information model extraction, J. Comput. Civ. Eng. 27 (6) (2013) 576–584.
[30] J. Beetz, J. Leeuwenand, B. Vries, IfcOWL: a case of transforming EXPRESS schemas into ontologies, Artif. Intell. Eng. Des. Anal. Manuf. (AI EDAM) 23 (1) (2009) 89–101.
[31] L. Zhang, R. Issa, Development of IFC-based construction industry ontology for information retrieval from IFC models, in: Proceedings of the 2011 EG-ICE Workshop, 2011.
[32] 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.
[33] 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.
[34] 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.
[35] W3C, OWL 2 Web Ontology Language Overview, 2009. <http://www.w3.org/TR/owl-features/>.
[36] ISO10303, Product Data Representation and Exchange – Part 11: Description Methods: The EXPRESS Language Reference Manual, 2003.
[37] W. Terkaj, A. Šojić, Ontology-based representation of IFC EXPRESS rules: an enhancement of the ifcOWL ontology, Autom. Constr. (2015) (in press).
[38] P. Pauwels, W. Terkajt, EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology, Autom. Constr. (2015) (in preparation).
[39] J. Xu, W.B. Croft, Query expansion using local and global document analysis, in: Proceedings of ACM SIGIR'96, 1996, pp. 4–11.
[40] J. Xu, W.B. Croft, Improving the effectiveness of information retrieval with local context analysis, ACM Trans. Inform. Syst. 18 (1) (2000) 79–112.
[41] Lucene, 2014. <http://lucene.apache.org/>.
[42] G. Salton, Developments in automatic text retrieval, Science 253 (5023) (1991) 974–980.
[43] V. Raghavan, P. Bollmann, G.S. Jung, A critical investigation of recall and precision as measures of retrieval system performance, ACM Trans. Inform. Syst. 7 (3) (1989) 205–229.
[44] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
[45] C.J.V. Rijsbergen, Information Retrieval, second ed., Butterworth, London, 1979.
[46] N. El-Gohary, T. El-Diraby, Merging architectural, engineering, and construction ontologies, J. Comput. Civ. Eng. 25 (2) (2011) 109–128.
[47] O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using explicit semantic analysis, ACM Trans. Inform. Syst. 29 (2) (2011) (Article 8).