

Enhanced Explicit Semantic Analysis for Product Model Retrieval in Construction Industry

Han Liu, Yu-Shen Liu, Pieter Pauwels, Hongling Guo, Ming Gu

Abstract—With the rapidly growing of online product models in construction industry, there is an urgent need for developing effective domain-specific information retrieval methods. Explicit Semantic Analysis (ESA) is a method that automatically extracts concept-based features from human knowledge repositories for semantic retrieval. This avoids the requirement of constructing and maintaining an explicitly formalized ontology. However, since domain-specific knowledge repositories are relatively small, the available terminologies are insufficient and concepts have coarse granularity. In this paper, we propose an enhanced ESA method for product model retrieval in construction industry. The major enhancements for the original ESA method consist of two parts. Firstly, a novel concept expansion algorithm is proposed to solve the problem caused by insufficient terminologies. Secondly, a re-ranking algorithm is developed to solve the problem caused by coarse granularity of concepts. Experimental results show that our method significantly improves the performance of product model retrieval and outperforms the state-of-the-art methods. Our method is also applicable to product retrieval in other engineering domain if a specific knowledge repository is provided in that domain.

Index Terms—Information Retrieval, Explicit Semantic Analysis (ESA), Domain Knowledge, Building Information Modeling (BIM), Industry Foundation Classes (IFC).

I. INTRODUCTION

BUILDING Information Modeling (BIM) has become the central technology in the AEC (Architecture, Engineering and Construction) industry [1], which also plays an increasingly important role in smart buildings [2], [3] and smart cities [4]. Meanwhile, the amount of BIM product models is growing rapidly on the web. For instance, the well-known Autodesk Seek [5] contains more than 68,000 commercial and residential building products (e.g. various windows, doors and beams) from over 400 manufacturers, and BIMobject¹ provides a large repository of building product models from 670 brands. Other online product model libraries are like the NBS National BIM

Library² and 3D Warehouse³. The product models are usually directly associated with documentation, e.g. specifications and descriptions. This product documentation commonly contains the textual description of product models, including their functions, dimensions, materials, performance, sustainability, manufacturers, and so forth. The product documentation is independent of the file formats of BIM models. Clearly, much information about the product models is embedded in this textual documentation.

The rapid increasing in the volume of online documented product model libraries also increases the difficulty for quickly finding information that is sufficiently close to the user's specific needs. In order to allow quick and accurate online search and retrieval of product models usable in BIM environments, appropriate information retrieval (IR) approaches should be adopted. Currently prevailing IR services in the AEC industry are mostly keyword-based, which is easy to be implemented. However, the accuracy of traditional keyword-based IR has often been problematic because of the semantic ambiguity of (1) the keywords used in search and of (2) the terminologies used in the search space. This problem also exists when applying traditional keyword-based IR methods to BIM product model libraries. One common solution for domain-specific retrieval is using a domain ontology. The natural language statements can be mapped to domain-specific concepts in a domain ontology, hence making the library and the queries semantically unambiguous. However, building a comprehensive domain ontology involves significant effort and complexity, even with the help of domain experts. The Industry Foundation Classes (IFC) [6], [7] is one of the most notable efforts in this regard, as it is proposed as a common neutral data model for the AEC domain that has been developed over more than 20 years of ontology engineering and evaluations.

In this paper, we investigate the usage of Explicit Semantic Analysis (ESA) [8] as an alternative basis for an IR method that successfully uses a domain-specific knowledge repository to enhance IR in the AEC domain. ESA typically makes use of an external document corpus as a knowledge source. This document corpus is analyzed and converted into a vector representation of the concepts. This vector representation can be understood or interpreted as a temporary light-weight ontology that drives and improves IR. The external large-scale knowledge in encyclopedia (e.g., Wikipedia⁴) provides an excellent example of what a document corpus for ESA

Han Liu, Yu-Shen Liu and Ming Gu are with the School of Software, Tsinghua University, Beijing 100084, China; and also with Tsinghua National Laboratory for Information Science and Technology (e-mail: liuhan15@mails.tsinghua.edu.cn; liuyushen@tsinghua.edu.cn; guming@tsinghua.edu.cn). (Corresponding author: Yu-Shen Liu)

Hongling Guo is with the Department of Construction Management, Tsinghua University, Beijing 100084, China (e-mail: hlguo@tsinghua.edu.cn).

Pieter Pauwels is with the Department of Architecture and Urban Planning, Ghent University, Belgium (e-mail: pipauwel.Pauwels@UGent.be).

The research is supported by the National Natural Science Foundation of China (61472202, 61272229). Dr. Pieter Pauwels is supported by the Special Research Fund (BOF) in Ghent University.

Copyright (c) 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

¹<http://bimobject.com>

²<http://www.nationalbimlibrary.com>

³<https://3dwarehouse.sketchup.com/index.html>

⁴<http://www.wikipedia.org>

could look like. In fact, ESA has initially been implemented using articles from Wikipedia [8]. By properly interpreting the natural language articles and definitions in Wikipedia-like encyclopedia, better IR algorithms can be conceived [9], [10], [11].

However, the wide topic range in large encyclopedia can also be a disadvantage for domain-specific retrieval [12]. Namely, domain-irrelevant entries in such large encyclopedia (like Wikipedia) may slow down the speed and cause ambiguity. The more specific the knowledge repository matches the semantics of the searched library, the better will be the results obtained through ESA-based IR methods. In other words, if the domain of application is known (e.g. construction industry), ESA-based IR methods can rely on a knowledge repository that matches this domain in order to obtain higher performance [12].

In this paper, we therefore propose to use the *documentation* of the IFC schema [7] (not just the IFC schema itself) as a domain-specific knowledge repository for more efficiently searching through the targeted existing BIM product model libraries. We particularly use the documentation that is made available for the IFC4 schema. When using ESA, the domain-specific knowledge source (i.e. the IFC4 documentation) can be automatically processed, without any help of domain experts. Because help from domain experts is not required, qualitative search and retrieval can be implemented more quickly and more easily than the case for methods that explicitly rely on more static ontologies.

A. Related Work

1) *General-purpose IR*: Traditional IR methods for textual information are often keyword-based. However, because of the synonymy and polysemy of natural language, the precision and recall rates of keyword-based retrieval are relatively low. On the one hand, one semantic concept can be expressed using different words (synonymy). When the words used by the author and the query sender are not the same, applicable results will not be returned through keyword-based retrieval, which lowers the *recall rate*. On the other hand, one word can have multiple semantic meanings (polysemy). Since keyword-based IR cannot distinguish the meanings of a term well in different contexts, some irrelevant results might be returned, which lowers the *precision rate*.

In order to solve the problems of synonymy and polysemy, several statistics-based methods have been proposed, which mainly include (1) using a thesaurus such as WordNet for query expansion and disambiguation [13], (2) analyzing the whole set of documents and finding potential links between terms (e.g. Latent Semantic Analysis (LSA) [14]), (3) analyzing top-ranked results from an initial retrieval and using feedback information to refine the search (e.g. Local Context Analysis (LCA) [15]). Overall, such statistic-based solutions are effective in general-purpose retrieval tasks. However, the results are not satisfactory in domain-specific retrieval tasks, when more terminologies of a specific domain are used. Terminologies that are specific to a domain may include special meanings and relations which are not available in

general thesauri such as WordNet. Thus the usage of a general thesaurus may result in poor performance for domain-specific retrieval requests. In our case (i.e. product model libraries in the AEC field), the domain is more fixed and more specific, hence it is possible to use such specific domain knowledge and improve the IR performance.

2) *Ontology-based IR in the AEC Field*: Research initiatives for IR in the AEC industry typically focus on the adoption of domain-specific knowledge. Many of these initiatives are ontology-based [16], [17], [18], [19], [20], [21]. Formal ontologies such as RDF and OWL provide good tools for computers to comprehend semantic information. However, building a comprehensive domain ontology involves significant effort and complexity with the help of domain experts. Typical challenges for using ontology-based IR methods in the AEC domain (and for using ontologies in general) are as follows. (1) Instead of using a more widely accepted ontology for the AEC field, researchers build up an isolated ontology on their own, which takes lots of effort and time. In addition, the resulting ontology is a private conceptualization, as opposed to a shared conceptualization of an area of interest. (2) Since a domain-specific ontology is a formal expression of domain knowledge, it must be built with the help of domain experts, where extra work is needed for communication. (3) Existing terms change and new terms emerge over time, especially in the frequently enlarging and changing BIM product model libraries. Static ontologies typically do not suffice, while dynamic ontologies can typically not keep pace with the rapid changes in product model libraries.

3) *ESA: Using an External Knowledge Repository to Enhance IR*: Instead of aiming at adopting domain knowledge for IR using a static ontology, one can also consider to adopt domain knowledge that is implicitly available in an online domain-specific knowledge corpus. This can avoid the requirement of constructing an explicitly formalized ontology. An alternative method for this purpose is ESA [8], [10], which automatically processes encyclopedia-like knowledge (e.g. Wikipedia) to enhance IR. Knowledge repositories like Wikipedia are structured in an “entry-description” form. In ESA, entries are treated as concepts, and a high-dimensional vector space is built up using concepts as dimensions, i.e. the concept space. The core of ESA is called a *semantic interpreter* [8]. Each term appearing in the query and target documents can be represented as a vector of concepts (the original entries) via the semantic interpreter. As a result, any text fragment can be mapped into a vector in the concept space in this way, which can be compared, indexed and retrieved in that space.

4) *Limitations of ESA in Domain-specific Retrieval*: ESA has shown good performance in IR using Wikipedia as a general-domain knowledge source [9], [11]. However, a wide range of topics in Wikipedia can also be a disadvantage for a narrow domain retrieval, since many irrelevant topic expansions will introduce noise and distortions in capturing term correlations [12], and also increase the time cost for calculation. By adopting a domain-specific knowledge repository, ESA would probably offer better results in the target domain retrieval.

However, the domain-specific knowledge repositories are usually much smaller than general-domain knowledge repositories like Wikipedia. For example, the domain-specific knowledge repository used in this paper, namely the IFC4 documentation [7], contains only 906 concepts and 7660 terms, which is very small compared with Wikipedia corpus that contains 1,187,839 concepts. When ESA is simply combined with such a small knowledge repository, there are two limitations as follows.

(1) **Insufficiency of terminologies in a small domain-specific knowledge repository.** In ESA, the terms that do not appear in the knowledge repository cannot be mapped into the concept space. This issue is quite common in our task. For example, queries and documents contains some brand or manufacturer names, which are clearly not included in the IFC4 documentation. In this case, ESA-based retrieval often returns inaccurate and incomplete results.

(2) **Coarse granularity of concepts in a small domain-specific knowledge repository.** Compared with Wikipedia which contains a large number of concepts, there are a relatively small number of concepts in the small knowledge repository. Many of the concepts generated from a knowledge repository are category definitions. Oftentimes, a category includes some subcategories, but when using a small knowledge repository, there may not be stand-alone concepts for the subcategories. As a result, documents describing different subcategories are likely to be indexed by one identical primary concept (with coarse granularity), so it is difficult to distinguish the subcategories if only using ESA. For example, there are many types of lamps (e.g., “LED”, “fluorescent lamp” and “metal-halide lamp”) are associated with the same concept `IfcLampTypeEnum` in IFC4 (see Fig. 1). Therefore, when receiving any type of lamps as a query keyword (e.g. “LED”), all documents associated with the concept `IfcLampTypeEnum` will be returned, but the documents about “LED” cannot be well distinguished from the documents about other lamp types.

5) *The IFC4 Knowledge Repository:* In the AEC industry, there are several well-known knowledge resources (e.g. *OmniClass*⁵, *Uniclass* and *Masterformat*), which have the potential to become domain-specific knowledge repositories. However, most of them are like taxonomies or classification systems, in which the entries (terminologies) are lack of sufficient textual descriptions. As a result, they are not suitable for ESA. In contrast, the documentation of the IFC4 Release (IFC4) [7] contains the entries associated with rich textual descriptions, which is suitable for ESA as the reference domain-specific knowledge repository in our investigation. The IFC4 Release consists of the IFC4 schema (specified by ISO 16739:2013) [6], which can be considered as the ontology used for describing BIM models, and an extensive textual documentation of all concepts used in the IFC4 schema. The IFC4 Release is published by *buildingSMART International* [22], which provides authority and acceptance in the AEC field because of its role as an international industry-supported standardization body. The IFC4 Release can hence be considered as a semantically rich

domain-specific knowledge repository related to BIM models in online product model libraries.

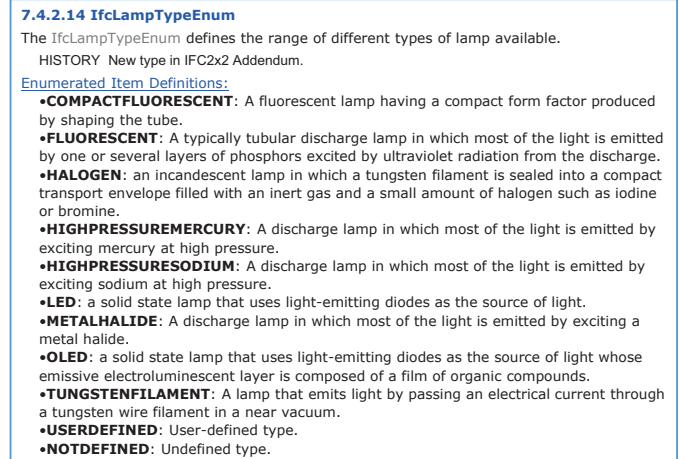


Fig. 1. An example entry in the IFC4 documentation.

Each page in the IFC4 documentation follows the “entry-description” format, similar to Wikipedia. An example entry in the IFC4 documentation is given in Fig. 1. For each concept (entry), the documentation page includes the name (e.g. `IfcLampTypeEnum`)⁶, the description, the sub-categories (e.g., fluorescent, halogen and LED), the involved concept properties, relationships, and so forth. Since some related terms (e.g., “lighting”, “lamp” and “LED”) co-occur in one same entry, ESA is able to build their semantic relatedness. The product model libraries that we consider in this paper typically include specific *building elements* (e.g. windows, doors, beams, columns). Therefore, we will only consider the two chapters in the IFC4 documentation that are related to building elements, namely the 6th chapter “Shared schemas” and the 7th chapter “Domain schemas”. These two chapters form the knowledge repository used throughout the remainder of this paper.

The semantic ambiguity in queries and product documents can be alleviated by using ESA concepts which are derived from the IFC4 documentation instead of from the Wikipedia corpus. For example, consider the short query “Duct Fitting”, which specifically refers to the connection parts between duct segments in the AEC domain. When using the IFC4 documentation as the domain-specific knowledge repository, the query’s top-5 concepts generated by our enhanced ESA are: `IfcDuctSilencerTypeEnum`, `IfcDuctSegment`, `IfcDuctFitting`, `IfcCableFitting` and `IfcDuctSegmentTypeEnum`, which are relevant to the user’s query in the AEC domain. In contrast, when using Wikipedia as the domain-specific knowledge repository, the query’s top-5 concepts generated by the original ESA are: `Salivary ducts`, `Interlobular duct`, `Intercalated duct`, `Major sublingual duct` and `Duct (anatomy)`, which are misinterpreted to the biomedical domain, not the AEC domain. This is because

⁵<http://www.omniclass.org>

⁶<http://www.buildingsmart-tech.org/ifc/IFC4/final/html/schema/ifcelectricaldomain/lexical/ifclamptypenum.htm>

Wikipedia does not cover more detailed knowledge sources in the specific AEC domain, and many Wikipedia articles about “Duct” in the biomedical domain are incorrectly associated with the query “Duct Fitting”.

B. Contributions

Although a domain-specific knowledge repository can make ESA-based retrieval more focused, the available terminologies are insufficient and concepts have coarse granularity in the narrow domain corpus such as the IFC4 documentation. Experimental results have shown that the performance of ESA suffers from a small-scale knowledge repository [12], [23]. Several methods have been developed for improving ESA with large-scale general knowledge repositories like Wikipedia [9], [11], [24], [25], [25], [26]. However, the existing improvements for ESA are not specifically designed for solving the two limitations (as mentioned in Section I-A4), which are caused by the small domain-specific knowledge repositories. To obtain better retrieval results, we present an enhanced ESA method for retrieving online BIM product model libraries using an external domain-specific knowledge repository (the IFC4 documentation [7]). Our main contributions are summarized as follows.

- To solve the problem caused by insufficient terminologies in a small domain-specific knowledge repository, a novel concept expansion algorithm is proposed.
- To solve the problem caused by coarse granularity of concepts in a small domain-specific knowledge repository, a re-ranking algorithm is developed.
- The presented retrieval method is integrated into a retrieval system for demonstrating the utility and effectiveness of our method. The experimental results show that our method significantly improves the performance of BIM product model retrieval and outperforms the state-of-the-art methods.

II. ENHANCED ESA FOR DOMAIN SPECIFIC RETRIEVAL

First, we briefly overview ESA-based retrieval [10], [9], [11]. Then we present the two-step enhancements to the original ESA in Section II-B and Section II-C, respectively.

A. The Overview of ESA-based Retrieval

Upon receiving a query, the original ESA-based retrieval mainly consists of four steps as follows.

(1) Building a semantic interpreter. The semantic interpreter can be regarded as a term-concept matrix, where each column corresponds to a concept and each row denotes a term that occurs in the external corpus of knowledge repository. Each element in the matrix corresponds to the value of *TF-IDF* (short for Term Frequency-Inverse Document Frequency).

(2) Document indexing: Mapping documents into concept vectors and creating the inverted index. In the document indexing stage, each document in the document collection is first represented by a Bag-of-Words (BOW) vector weighted with its Term Frequency (TF) [11]. By multiplying the BOW vector with the matrix of semantic interpreter, each

document is mapped into a concept vector. Once these concept vectors of all documents are generated, an inverted index is created to map back from each concept to its associated documents.

(3) Query processing: Mapping a query into a concept vector. In the query processing stage, a user’s query is also mapped into a concept vector by multiplying the BOW vector of the query with the semantic interpreter matrix.

(4) Fetching the indexed documents. The relevant documents are fetched from the inverted index using the selected query concepts, where computing semantic relatedness between the query and documents is reduced to calculating the cosine similarity between their concept vectors.

Compared with the large-scale general knowledge repositories like Wikipedia, the domain-specific knowledge repositories such as the IFC4 documentation are often much smaller and incomplete. In a domain-specific retrieval task, however, we find that the document collection to be retrieved is much larger than the associated knowledge repository. In order to utilize the large document collection for improving the performance of ESA-based retrieval, we present an enhanced ESA method for retrieving online product model libraries. Inspired by the idea of pseudo-relevance feedback [27], we make use of the top-ranked results from an initial keyword retrieval to help improving the performance of semantic retrieval. The two-step enhancements to the original ESA are as follows.

- **Concept expansion.** In the query processing stage, we propose a new concept expansion algorithm so that the queried terms are less likely to be missed in this step (see Section II-B).
- **Re-ranking.** After fetching the indexed documents, the retrieval results are re-ranked so that the documents most-related to the query are ranked to the top (see Section II-C).

B. Concept Expansion

In the original ESA method, the query is mapped into a concept vector by multiplying the BOW vector of a user’s query with the semantic interpreter matrix. However, because of insufficient terminologies in a small domain-specific knowledge repository, the semantic interpreter of ESA often fails to map some terms that do not appear in the knowledge repository into the concept space. This issue has been discussed in Section I-A4. To solve this problem, we propose a *concept expansion* algorithm to generate the meaningful concept vectors from the top-ranked documents obtained in an initial keyword retrieval.

The semantic interpreter in ESA can be regarded as a term-concept matrix. Assume that this matrix is denoted as a $m \times n$ matrix \mathbf{M} , where m is the number of terms and n is the number of concepts (all articles in the IFC4 documentation). Text fragments can be mapped into the concept space by multiplying their BOW vectors with the semantic interpreter matrix \mathbf{M} .

In the document indexing stage, each document is indexed by the concepts. Let N be the number of all the documents to be retrieved, and therefore the whole document collection can be represented as a $m \times N$ matrix \mathbf{D} , in which each column is

Algorithm 1 Concept Expansion

Input: the BOW vector of the query \mathbf{q} , the document collection \mathbf{D} , the semantic interpreter \mathbf{M} , and the concept index of documents \mathbf{C} ;

Output: the merged concept vector \mathbf{c}_m and the top-ranked documents in an initial retrieval \mathbf{d}_{KW} ;

- 1: get the mapped concept vector \mathbf{c}_q using the semantic interpreter: $\mathbf{c}_q = \mathbf{M}^T \mathbf{q}$;
- 2: get the top-ranked documents from an initial keyword retrieval: $\mathbf{d}_{KW} = \text{top}K(\mathbf{D}^T \mathbf{q})$;
- 3: get the expanded concept vector \mathbf{c}_e in the concept index: $\mathbf{c}_e = \mathbf{C} \mathbf{d}_{KW}$;
- 4: normalize and merge the two concept vectors: $\mathbf{c}_m = \text{top}K(\text{normalize}(\mathbf{c}_q) + \text{normalize}(\mathbf{c}_e))$;

the BOW vector of a document. By mapping all the documents into the same concept space, a concept index matrix for all the documents can be built, which is denoted as a $n \times N$ matrix \mathbf{C} :

$$\mathbf{C} = \mathbf{M}^T \mathbf{D}, \quad (1)$$

where each column in \mathbf{C} is the concept vector of a document. In a similar way, let \mathbf{q} be the m -dimensional BOW vector of a query in the query stage, and the mapped concept vector \mathbf{c}_q can be calculated as

$$\mathbf{c}_q = \mathbf{M}^T \mathbf{q}. \quad (2)$$

It is worth noting that if one term in the query does not appear in the domain-specific knowledge repository, the corresponding row in the semantic interpreter matrix \mathbf{M} would become zero. This leads to that this term would be missed when using the semantic interpreter for mapping the query into the concept space. This problem significantly affects the accuracy of ESA retrieval, since the short queries are usually with only a few terms.

To solve this problem, we propose a novel concept expansion algorithm to get an expanded concept vector from the document collection related to the query. Our idea is inspired by pseudo-relevance feedback [27] which assumes that the top-ranked results in an initial keyword retrieval are more likely to be relevant. Different from textual query expansion, which generates new query terms, our concept expansion method solves the problem of ‘‘insufficient terminologies’’ by directly generating an expanded concept vector. In our method, some potentially relevant concepts can be obtained inversely from the result of an initial keyword retrieval. The result of an initial keyword retrieval is represented as a N -dimensional vector \mathbf{d} ,

$$\mathbf{d} = \mathbf{D}^T \mathbf{q}. \quad (3)$$

In practice, only the top-ranked documents are kept.

Since the concept index matrix \mathbf{C} in Eq. (1) is the mapping between concepts and documents, we can get a new vector \mathbf{c}_e including some ‘‘expanded’’ concepts, as calculated by

$$\mathbf{c}_e = \mathbf{C} \mathbf{d} = \mathbf{M}^T \mathbf{D} \mathbf{D}^T \mathbf{q}. \quad (4)$$

Since the expanded concepts are related to the top-ranked documents, more documents similar to the top-ranked documents can also be retrieved through these expanded concepts.

Algorithm 2 Re-ranking

Input: the query \mathbf{q} , the results in ESA-based retrieval \mathbf{d}_{ESA} , the top-ranked documents in the keyword retrieval \mathbf{d}_{KW} ;

Output: the re-ranked documents \mathbf{d}_{RE} ;

- 1: **for** each term c in the document set \mathbf{d}_{KW} , **do**
- 2: calculate the relatedness $rel_{RR}(c, \mathbf{q})$;
- 3: **end for**
- 4: choose the top-weighted terms $\mathbf{w}_c = \text{top}K(\{rel_{RR}(c, \mathbf{q})\})$;
- 5: **for** each document $d_i \in \mathbf{d}_{ESA}$, **do**
- 6: calculate the document score using Eq. (7);
- 7: **end for**
- 8: sort the documents by the document score: $\mathbf{d}_{RE} = \mathbf{d}_{ESA}.\text{sortBy}(\text{score}(d_i))$;

In the domain-specific retrieval, the number of terms appearing in the document collection is much larger than the number of terms appearing in the small domain-specific knowledge repository. This results in less zero rows in the matrix \mathbf{D} than in \mathbf{M} . Therefore, the query can be mapped into the concept space even if some of the terms do not appear in the small domain-specific knowledge repository.

In our method, we use Eq. (2) and Eq. (4) for generating two concept vectors \mathbf{c}_q and \mathbf{c}_e , respectively. If \mathbf{c}_q misses some terms that do not appear in the semantic interpreter, the retrieval can in any case still go on using \mathbf{c}_e . Next, \mathbf{c}_q and \mathbf{c}_e are normalized and combined into a unified concept vector denoted by \mathbf{c}_m . Finally, only the top-weighted concepts in \mathbf{c}_m are kept, which indicates that certain concepts are closely related to both the initial query and the top-ranked documents. Algorithm 1 shows the detailed algorithm of concept expansion.

C. Re-ranking

The coarse granularity of concepts is another limitation when using ESA in a small domain-specific knowledge repository, as mentioned in Section I-A4. To solve this problem, we propose a re-ranking approach which is based on Local Context Analysis (LCA) [15]. LCA aims to calculate the relatedness between each term in a document and an input query, which was initially introduced for query expansion applications. Instead, we use LCA in this paper for re-ranking the retrieval results to match the initial query well.

The top-ranked documents in the initial keyword retrieval results are assumed most likely to be relevant to the given query. For each term c in the top-ranked documents, LCA calculates the relatedness between c and the query \mathbf{q} by $rel_{LCA}(c, \mathbf{q})$:

$$rel_{LCA}(c, \mathbf{q}) = \prod_{t \in \mathbf{q}} (0.1 + co(t, c)), \quad (5)$$

where t is a term in the query, and $co(t, c)$ is the co-occurrence rate of term c and term t in the top-ranked documents [15].

The value of $rel_{LCA}(c, \mathbf{q})$ indicates that how a term in the top-ranked documents is related to the query. To increase the affect of the terms in the query, we define a new relatedness function as $rel_{RR}(c, \mathbf{q})$:

The screenshot displays the BIMSeek search interface. At the top, the search query 'TOTO Lavatory' is entered. Below the search bar, the results are listed in a table format. Each result includes a title, a brief description of the product specifications, and a small image of the product. The results are ranked based on relevance. On the right side of the interface, there are two panels. The 'Expanded concepts' panel lists various technical terms and their corresponding document counts. The 'Terms selected for re-ranking' panel shows a word cloud of terms that are automatically selected for re-ranking based on the search query.

Fig. 2. The user's interface of the retrieval system for online product model libraries.

$$rel_{RR}(c, \mathbf{q}) = \begin{cases} \alpha TF(c, \mathbf{q}) rel_{LCA}(c, \mathbf{q}) & c \in \mathbf{q} \\ rel_{LCA}(c, \mathbf{q}) & c \notin \mathbf{q} \end{cases} \quad (6)$$

where α is a constant which is chosen as a number larger than the length of the query (we typically select 10), and $TF(c, \mathbf{q})$ counts the term frequency of c in \mathbf{q} . The terms with highest $rel_{RR}(c, \mathbf{q})$ values are selected as a term set \mathbf{w}_c . To re-rank the retrieval results, the score for each document d_i in the retrieval results is calculated by:

$$score(d_i) = \sum_{c \in \mathbf{w}_c} rel_{RR}(c, \mathbf{q}) TF(c, d_i). \quad (7)$$

Algorithm 2 shows the detailed algorithm of re-ranking.

Finally, we demonstrate our two-step enhancements (i.e. Algorithm 1 and Algorithm 2) compared with the original ESA through a retrieval application. Consider the user's query "TOTO lavatory" in Fig. 2, which refers to a lavatory of the manufacturer TOTO or its sub-brand. Since both the terms "TOTO" and "lavatory" are not included in the IFC4 documentation, the original ESA cannot generate the concept vector for this query when using the IFC4 documentation as the knowledge repository. In contrast, using Algorithm 1, the top-4 concepts generated for this query are the following: `IfcBoiler`, `IfcTank`, `SanitaryTerminalTypeSink`, `SanitaryTerminalTypeWashHandBasin`, so that potentially relevant documents can be fetched successfully. Among the fetched documents, the most-related documents are mixed up with some weakly-related ones. In order to re-rank the documents, several terms are selected by Algorithm 2, including "lavatory", "toto", "basin", "sink" and so on. In the re-ranked list, the top-ranking documents match the user's query intent well, as shown in Fig. 2.

III. EXPERIMENTS

A. The Retrieval System and Benchmark

The presented method has been integrated into a retrieval system for demonstrating the utility and effectiveness of our method. In this system, the retrieval service is deployed based on Django and MongoDB. Scrapy crawler⁷ is used to collect online product model documents. In the retrieval system, the keyword-based IR service is provided by Apache Lucene⁸, which is used for pseudo-relevance feedback, and it is also a baseline of performance in our experiments. In this section, all the experiments were run on a 3.60GHz processor with 16GB memory on Windows 10.

The user's interface of the retrieval system is shown in Fig. 2. The user first specifies a search query, and the system returns the ranked results of online product models related to the user's query. In Fig. 2, the top-right panel shows the expanded concepts (i.e. \mathbf{c}_m) computed by Algorithm 1, and the bottom-right panel shows the terms (i.e. the term set \mathbf{w}_c) that are automatically selected for re-ranking using Algorithm 2.

To evaluate the performance of various IR methods, we need to generate a set of test queries. Currently, the document collection used in the retrieval system contains a total number of 17,903 product model documents acquired from Autodesk Seek website[5]. In the document collection, each product document is associated with two types of labels: product category and product manufacturer. In our test, these two types of labels are used as "ground truth" for generating our test queries. By combing different category and manufacturer names, a set of candidate queries is first generated in the form

⁷<https://scrapy.org>

⁸<http://lucene.apache.org>









	keyword-based IR	original ESA	our method
Rank 1	 <i>Cooper Controls</i> Greengate™ Lighting Control	 <i>Prudential Lighting</i> P-46	 <i>Cooper Lighting</i> Invue™ ENC LED
Rank 2	 <i>Cooper Lighting</i> Streetworks™ LED	 <i>Prudential Lighting</i> P-40	 <i>Cooper Lighting</i> Invue™ ENT LED
Rank 3	 <i>Cooper Lighting</i> Invue™ ENT LED	 <i>Prudential Lighting</i> P-43	 <i>Cooper Lighting</i> Invue™ ENV LED

Fig. 3. Comparison of the retrieval results for the same query “Cooper lighting” using three methods (from left to right: Keyword-based IR, the original ESA method and our method). The top 3 results are listed with the thumbnail, manufacturer name and product name.

of “manufacturer name + category name”, such as “TOTO lavatory” and “Cooper lighting”. In order to obtain a more reliable evaluation standard, the queries which cover less than 50 product documents in the document collection are removed from the generated candidate queries. Finally, 63 test queries are kept as the benchmark queries of our experiments.

If some of the terms (e.g. manufacturer names) are not included in the IFC4 documentation, the original ESA method cannot generate the correct query interpretation. For example, considering a user’s query “Cooper lighting”, where “Cooper” is the manufacturer name. Fig. 3 shows the top-3 retrieval results returned by the traditional keyword-based IR method, the original ESA method and our method, respectively. We find that many irrelevant results are returned when using the first two methods. When using the keyword-based IR, the top-one result is a product about the “Cooper controls”, which does not reflect the user’s query intent well. The reason is that the webpage of “Cooper controls” contains many individual terms “lighting” and “Cooper”. When using the original ESA, the products from a different manufacturer “Prudential” are listed in the top-3 retrieved results. Although the top-3 retrieved results are all about “lighting”, the manufacturer name (“Cooper”) is ignored by ESA. This is because the manufacturer name does not appear in the IFC4 documentation when using ESA. In contrast, when using our method, the top-3 results are all about “Cooper lighting”, which match the user’s query intent well.

B. Comparison with Other Related Methods

To compare the performances between our method with other related methods, we adopt several standard evaluations of IR, including mean average precision (meanAP), recall rate, P@10 and failure rate. Recall rate is an indicator about how many relevant documents are found in the retrieval task. Higher recall rate means that more relevant results are returned. MeanAP stands for the averaged precision on each position of the sequence of retrieved documents. Higher meanAP means that more relevant documents are ranked to the front of the results. P@10 is the precision of the top-10 results, which indicates how many results are relevant on the first page of

TABLE I
COMPARISON BETWEEN OUR METHOD AND SOME EXISTING IR METHODS.

	MeanAP	Recall	P@10	Failure rate
KW	0.493	0.736	0.690	1/63
ESA [8]	0.149	0.526	0.143	11/63
ESA+CE	0.221	0.907	0.189	0/63
ESA+RR	0.342	0.526	0.546	11/63
ESA+CE+RR	0.544	0.907	0.671	0/63
ESA+FS [11]	0.153	0.524	0.141	12/63
ESA+NO [24]	0.097	0.821	0.071	5/63
ESA+QE [26]	0.142	0.687	0.121	9/63
IFCQE [16]	0.447	0.745	0.598	2/63
IFCCA [17]	0.277	0.631	0.308	11/63

results. Failure rate is an indicator about the number of cases among the 63 queries, when there is no relevant documents in the results.

In the experiments, the performance of keyword-based IR provided by Lucene (KW) is used as a baseline. First, the two-step enhancements proposed in this paper, including “concept expansion” (ESA+CE), “re-ranking” (ESA+RR) and their combination (ESA+CE+RR), are tested and compared with the original ESA method (ESA). Then the state-of-the-art methods for improving ESA on general-purpose knowledge repositories are compared. In these methods, the first one uses *feature selection* to remove redundant dimensions from a concept vector (ESA+FS) [11]. The second one considers the *non-orthogonality* between concepts (ESA+NO) [24]. The third one performs a *query expansion*, which uses semantic interpreter to generate the expanded query string (ESA+QE) [26]. In addition, two ontology-based methods for domain-specific retrieval are also compared, including a *query expansion* method using an IFC ontology (IFCQE) [16] and a *concept annotation* method using an IFC ontology (IFCCA) [17]. The experimental results are listed in Table I, where the best values are highlighted in bold black. The results show that our method achieves the best results for BIM product model retrieval in the state-of-the-art methods.

C. Experimental Analysis and Discussion

According to the experimental results, the original ESA does not work well when using IFC4 documentation as the knowledge repository. The semantic interpreter of ESA does not generate the correct query interpretation in many cases, and consequently the 11 ones of 63 queries do not fetch any relevant documents. The state-of-the-art methods for improving ESA on general-purpose knowledge repositories cannot yield good performance for the domain-specific retrieval, since they are not specifically designed for solving the problems of ESA-based retrieval on a small-scale knowledge repository. In Table I, the method (ESA+FS) only slightly improves the meanAP value, but it is not helpful in finding more related documents. The method (ESA+NO) and (ESA+QE) are both able to improve the recall rate by returning more potentially related documents, but meanwhile they suffers the loss of accuracy.

In contrast to the performance of the original ESA, the enhancements of “concept expansion” and “re-ranking” are

both effective in improving the performance of domain-specific IR. By comparing the experimental results of **ESA** and **ESA+CE** in Table. I, the results suggest that the enhancement of “concept expansion” can handle the cases that the query terms are not included in the knowledge repository. A query can be interpreted into a better concept vector that matches more relevant documents, which increases the recall rate from 0.526 to 0.907 and decreases the failure rate from 11/63 to 0/63. In addition, by comparing the results of **ESA+CE** and **ESA+CE+RR**, the results show that the enhancement of “re-ranking” increases the meanAP value from 0.221 to 0.544, which indicates that it is effective in ranking most-related documents to the top positions.

Our final solution is the combination of “concept expansion” and “re-ranking” (**ESA+CE+RR**), which outperforms the state-of-the-art methods in meanAP value, recall rate, and failure rate.

IV. CONCLUSION AND DISCUSSION

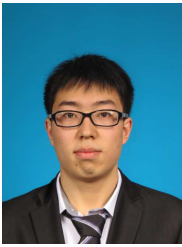
In this paper, we propose an enhanced ESA method for domain-specific information retrieval. Compared with ontology-based retrieval methods, the adoption of ESA allows the automatic generation of semantic information from domain-specific knowledge repositories, which avoids the requirement of constructing and maintaining an explicitly formalized ontology. The proposed two-step enhancements include concept expansion and re-ranking, which are effective in solving the problems of insufficient terminologies and coarse granularity of concepts in a small domain-specific knowledge repository. Using the IFC4 documentation, we build up a retrieval system for online BIM product model libraries. The experimental results show that our method significantly improves the performance of ESA in domain-specific information retrieval tasks.

One limitation of our method is that the single IFC4 documentation cannot fully cover the needs of product retrieval in the AEC field. For example, there are some types of products in OmniClass that are not included in the current version of IFC4 documentation. It is interesting to explore how to combine various AEC knowledge resources (such as ISO 12006-2, Uniclass and OmniClass) with the IFC4 documentation for ESA-based retrieval, which is our future work.

By replacing IFC4 documentation with various domain-specific knowledge repositories, the proposed method is also applicable to other engineering domains. Other than the BIM product documents, it is also interesting to retrieve various kinds of BIM-related documents such as BIM design documents. Another future work is to extend our current method to various BIM-related documents where broader and more diverse domain-specific knowledge repositories and document collections exist.

REFERENCES

- [1] C. Eastman, P. Teicholz, R. Sacks, and K. Liston, *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors*, 2nd ed. Hoboken, NJ: John Wiley and Sons, 2011.
- [2] S. N. Han, G. M. Lee, and N. Crespi, “Semantic context-aware service composition for building automation system,” *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 752–761, 2014.
- [3] J. B. Kim, W. Jeong, M. J. Clayton, J. S. Haberl, and W. Yan, “Developing a physical BIM library for building thermal energy simulation,” *Automation in construction*, vol. 50, pp. 16–28, 2015.
- [4] M. V. Moreno, F. Terroso-Saenz, A. Gonzalez, M. Valdes-Vela, A. F. Skarmeta, M. A. Zamora-Izquierdo, and V. Chang, “Applicability of big data techniques to smart cities deployments,” *IEEE Transactions on Industrial Informatics*, 2016.
- [5] Autodesk, “Autodesk Seek,” 2016, available from: <http://seek.autodesk.com>.
- [6] International Organization for Standardization (ISO), “ISO 16739:2013 – Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries,” 2013, ISO 16739, Geneva.
- [7] buildingSMART International, “Industry Foundation Classes IFC4 official release,” 2013, available from: <http://www.buildingsmart-tech.org/ifc/IFC4/final/html/index.htm>.
- [8] E. Gabrilovich and S. Markovitch, “Computing semantic relatedness using Wikipedia-based explicit semantic analysis,” in *Proceedings of IJCAI*, 2007, pp. 1606–1611.
- [9] O. Egozi, E. Gabrilovich, and S. Markovitch, “Concept-based feature generation and selection for information retrieval,” in *Proceedings of AAAI*, 2008, pp. 1132–1137.
- [10] E. Gabrilovich and S. Markovitch, “Wikipedia-based semantic interpretation for natural language processing,” *Journal of Artificial Intelligence Research*, vol. 34, pp. 443–498, 2009.
- [11] O. Egozi, S. Markovitch, and E. Gabrilovich, “Concept-based information retrieval using explicit semantic analysis,” *ACM Transactions on Information Systems*, vol. 29, no. 2, p. Article 8, 2011.
- [12] T. Gottron, M. Anderka, and B. Stein, “Insights into explicit semantic analysis,” in *Proceedings of CIKM*, 2011, pp. 1961–1964.
- [13] E. M. Voorhees, “Using WordNet to disambiguate word senses for text retrieval,” in *Proceedings of SIGIR*, 1993, pp. 171–180.
- [14] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.
- [15] J. Xu and W. B. Croft, “Query expansion using local and global document analysis,” in *Proceedings of SIGIR*. ACM, 1996, pp. 4–11.
- [16] G. Gao, Y.-S. Liu, M. Wang, M. Gu, and J.-H. Yong, “A query expansion method for retrieving online BIM resources based on Industry Foundation Classes,” *Automation in Construction*, vol. 56, pp. 14–25, 2015.
- [17] G. Gao, Y.-S. Liu, P. Lin, M. Wang, M. Gu, and J.-H. Yong, “BIMTag: Concept-based automatic semantic annotation of online BIM product resources,” *Advanced Engineering Informatics*, vol. 31, pp. 48–61, 2017.
- [18] H.-T. Lin, N.-W. Chi, and S.-H. Hsieh, “A concept-based information retrieval approach for engineering domain-specific technical documents,” *Advanced Engineering Informatics*, vol. 26, no. 2, p. 2012, 349–360.
- [19] G. J. Hahm, M. Y. Yi, J. H. Lee, and H. W. Suh, “A personalized query expansion approach for engineering document retrieval,” *Advanced Engineering Informatics*, vol. 28, no. 4, pp. 344–359, 2014.
- [20] Z. Li, V. Raskin, and K. Ramani, “Developing engineering ontology for information retrieval,” *Journal of Computing and Information Science in Engineering*, vol. 8, no. 1, pp. 737–745, 2008.
- [21] G. Costa and L. Madrazo, “Connecting building component catalogues with BIM models using semantic technologies: an application for precast concrete components,” *Automation in Construction*, vol. 57, pp. 239–248, 2015.
- [22] BuildingSMART International, “BuildingSMART - international home of openBIM,” 2014, available from: <http://www.buildingsmart.org/>.
- [23] M. Anderka and B. Stein, “The ESA retrieval model revisited,” in *Proceedings of SIGIR*, 2009, pp. 670–671.
- [24] N. Aggarwal, K. Asooja, G. Bordea, and P. Buitelaar, “Non-orthogonal explicit semantic analysis,” in *Proceedings of *SEM*, 2015, pp. 92–100.
- [25] T. Polajnar, N. Aggarwal, K. Asooja, and P. Buitelaar, “Improving ESA with document similarity,” in *Proceedings of ECIR*, 2013, pp. 582–593.
- [26] N. Aggarwal, K. Asooja, and P. Buitelaar, “Exploring ESA to improve word relatedness,” in *Proceedings of *SEM*, 2014, pp. 51–56.
- [27] G. Salton and C. Buckley, “Improving retrieval performance by relevance feedback,” *Journal of the American Society for Information Science*, vol. 41, no. 4, pp. 288–297, 1990.



Han Liu received the B.S. degree in Civil Engineering from Tsinghua University, Beijing, China, in 2015. He is currently working on his Ph.D. degree in Software Engineering at Tsinghua University. His research interests include Building Information Modeling (BIM), information retrieval and rule checking.



Yu-Shen Liu is an Associate Professor in School of Software at Tsinghua University, Beijing, China. He received his B.S. degree in Mathematics from Jilin University, China, in 2000. He earned his PhD degree in the Department of Computer Science and Technology at Tsinghua University, China, in 2006. He spent three years as a post doctoral researcher in Purdue University from 2006 to 2009. His research interests include Building Information Modeling (BIM), smart building, information retrieval, semantic search and pattern recognition.



Pieter Pauwels holds a Master (2008) and PhD Degree (2012) in Engineering: Architecture from Ghent University. In his research, he investigates how and to what extent information system support may be provided for architectural design thinking. This research explores the usage of information management techniques, in particular BIM tools, semantic web technologies and linked data. He is now assistant professor at Ghent University, the Department of Architecture and Urban Planning.



Hongling Guo graduated from Harbin Institute of Technology, China, as well as The Hong Kong Polytechnic University, Hong Kong, then worked as a researcher and visiting lecturer at The Hong Kong Polytechnic University for three years. In 2011, he started as an associate professor in the Department of Construction Management at Tsinghua University in China. His research areas include Building Information Modeling (BIM), Virtual Construction/Virtual Prototyping, Construction Safety Management Innovation, and Smart Construction.



Ming Gu received the B.S. degree in computer science from the National University of Defense Technology, Changsha, China, in 1984, and the M.S. degree in computer science from the Chinese Academy of Science, Shengyang, China, in 1986. Since 1993, she has been working as a Professor at Tsinghua University, Beijing, China. Her research interests include formal methods, middleware technology, and distributed applications.