Advanced Engineering Informatics xxx (2015) xxx–xxx

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 xxxx

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.

Please cite this article in press as: G. Gao et al., BIMTag: Concept-based automatic semantic annotation of online BIM product resources, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.003