Advanced Engineering Informatics 33 (2017) 154–163

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

A method for clustering unlabeled BIM objects using entropy and TF-IDF with RDF encoding

Mostafa Ali *, Yasser Mohamed

University of Alberta, Canada

ARTICLE INFO

Article history:
Received 31 January 2017
Received in revised form 29 June 2017
Accepted 30 June 2017
Available online 6 July 2017

Keywords:
Cluster
Semantic web
RDF
Entropy
BIM
Tf-idf

ABSTRACT

Oil and gas projects involve different construction disciplines such as mechanical, structural, and electrical. The current practice in these projects involves creating separate building information models for the different disciplines and compiling them into one model to check for collisions or conflicts. Due to intellectual property, contractual requirements, unfinished design, or technical issues during final model compilation, the final merged model lacks essential data for contractors such as the trade of each object in the model. Nonetheless, the model is issued to contractors who utilize it in different pre-construction planning tasks. However, due to data loss, incompleteness, or inconsistency, the model usability can become limited and the contractor has to review the model manually to extract information from it. This is a lengthy and costly task that becomes more challenging in fast-tracked projects that involve periodically issuing updated Building Information Models. One type of information that contractors need for different planning and estimation purposes is the scope of work for different construction trades in different areas of the project. In many cases, models lack explicit attributes of 3D objects that make it possible to perform an automated query of these objects by trade type. This research suggests a state of the art solution to automate the extraction of this information in such cases. In this paper, we describe a method that utilizes Resource Description Framework (RDF) encoding of BIM data together with Term Frequency-Inverse Document Frequency (TF-IDF) and entropy-based algorithms to automatically group 3D objects based on their trade. The proposed methodology is tested using three actual cases of oil and gas projects with more than four million objects in total. The results show that the proposed method can achieve a 91% purity in the generated groups.

1. Introduction

The benefits of building information models have been identified in both Architecture Engineering and Construction (AEC) practice and research [1–3]. Industrial oil and gas projects are among the early adopters of the technology. In these projects, designs are produced by different engineering disciplines (i.e., structural, mechanical, electrical, etc.) and represented as separate BIM models. These models are merged into one model to check for conflicts or clashes. Afterwards, the merged model is issued to one or more contractors for construction. Large industrial projects are typically fast-tracked [4] and as such, BIM models of these projects are not finalized before construction starts. Consequently, they are issued periodically as designs progress in parallel to construction [1].

From a contractor perspective, although models may be incomplete and subject to change, they still represent a rich source of information that can be used for preliminary resource planning and estimation [2]. However, due to legal issues such as intellectual property and contractual requirements [5], and because most models are compilations of many incomplete smaller sub models, some data are lost, missing, or inconsistent. This hinders the ability to automate different tasks in the model [6] and the contractor has to spend a significant number of man hours investigating and extracting relevant data.

A common requirement for a contractor is to classify BIM objects based on their trade (e.g., structural, mechanical, electrical, etc.). Traditionally, this involves reviewing the model manually and identifying custom rules that can be used to filter objects for each trade. A rule is a way to specify an attribute value that is consistent over the domain of one trade such that if the model is filtered based on this rule, only objects from this trade are retrieved. Identifying these rules is a tedious manual task that has to be executed for each project because rules will inevitably

* Corresponding author at: 1-047 NREF Building, University of Alberta, AB T6G 2W2, Canada.

E-mail address: MostafaAli@ualberta.ca (M. Ali).

http://dx.doi.org/10.1016/j.aei.2017.06.005
1474-0346/

change based on different modellers' conventions. Additionally, it is a repetitive task that has to be refined many times to make sure the right rules have been found. This task may also need to be repeated every time a new version of the model is released to the contractor.

This paper proposes a method to automate this task. The proposed method relies on encoding a BIM model using the Resource Description Framework (RDF) [7]. RDF is a modeling data system which explicitly represents every relation between any two data elements. RDF represents information as a set of statements; each statement consists of a subject, a predicate, and an object. RDF has been a recommendation from W3C since 1999 [8,9].

After encoding BIM data in RDF format, we apply an algorithm that utilizes Shannon entropy [10] and Term Frequency Inverse Document Frequency (TF-IDF) [11,12] measures to group the objects into clusters that represent different trades.

In the following sections of this paper we first explain the research objectives, then review literature related to the research problem, outline our proposed solution and, finally, present the results of testing the proposed solution on three real case studies for oil and gas projects that have been successfully completed in Alberta, Canada.

2. Objectives

This study is based on the assumption that the targeted BIM models, i.e. the input for the methodology proposed here, are imprecise, inaccurate and made of several objects that have different attributes and different values for these attributes but none of these attributes explicitly identify the trade for a given object. It is also assumed that the set of attributes and their values change from one project to another. Ideally, a BIM model should contain enough data to sufficiently describe each object in the model. However, when merging models of different engineering systems that have been developed using different software applications into one BIM model (e.g., NavisWorks© [13] models, which are common in oil and gas projects), some data may be missing or inconsistent due to technical issues when merging different models, or incomplete design; or because of intellectual property, or contractual requirements. Data loss or inconsistency can severely limit the usability of the final model by the contractor who has to visually review and inspect objects in the model to extract required information about work items for different construction trades.

The objective of this research is to develop an automated solution to replace or support this visual inspection task. With minimum manual intervention from the user, this solution should assist in identifying distinct groups of objects in the model that belong mostly or fully to the same trade.

To achieve this objective, the following questions will also need to be addressed:

- Are the common attributes between objects from the same trade sufficient to distinguish them in a merged BIM model? A merged BIM model contains data from different sources which differ based on related attributes, authoring software, and modeller preferences. Hence, the proposed method has to find the attributes that are candidates for filtering objects based on a specific trade.
- What is a suitable method to encode and query a model to find these common attributes automatically? A trial and error method can usually be used to find these attributes manually. Instead, we examine the use of a semantic web data model (RDF) and query language (SPARQL) [14] to find these attributes automatically.
- How does one differentiate between common attributes that are used by more than one trade? After finding a common attribute within a trade, the attribute should not automatically be used to filter objects for that trade because the same attribute might be used in another trade as well. Here, Shannon entropy [10] and TF-IDF [11,12] measures will be used to test how much information a given attribute can provide to differentiate between trades.
- What is an appropriate measure of the success of the proposed method? Once clusters are identified, a measure of the purity of these clusters needs to be used to evaluate how much of the objects in a cluster truly belong to the same trade.

The following section reviews relevant solutions used in previous studies to address problems similar to the ones described above. It also provides justification for the choice of solutions used in this study.

3. Previous work and related studies

The following sections discuss previous work related to information migration and merging and how semantic web technology is used as a potential solution that can improve BIM models. The discussion highlights why previously suggested approaches are not suitable for use in this study especially with the assumption of data completeness of BIM models, which is not always true in practice and is a key motivation of this research. Finally, some background about the different clustering and similarity measures used in this study is given.

3.1. Information migration and RDF format

Information is an essential asset for any business and the value of the information increases with the ability to link data from different sources [15]. Unfortunately, linking data from different sources is challenging as it requires not only transferring data and schema, but updating queries and business-logic layer [16].

As an alternative to rigid data schemas, RDF utilizes a flexible format [17]. The RDF statement consists of three parts: subject, predicate, and object, which is known as "triples." Each part of the RDF triple (e.g., subject) is known as a Resource and it should be expressed uniquely using a Uniform Resource Identifier (URI).

In order to add semantics to the stored triples, an ontology should be defined. Ontology is a precise explanation of terms and reasoning in a data domain that allows machines to act as if they understand the meaning and find relations in the domain. This is arguably one of the most powerful features of a semantic web framework. Ontology has been proven to enhance mapping data from different schemas [18].

In addition to its popularity in the web world, the semantic web has many applications in other domains. For example, traditional database structures cannot sufficiently model relations in an object-oriented paradigm. For example, a schema for a steel structure domain that models structural beams and columns cannot specify that the beams and columns have the same superclass (e.g., a structural element); therefore, it will be the responsibility of the application side to capture this relationship.

Alternatively, semantic web stores relationships between objects with the data. This allows a complete decoupling between knowledge representation and the application layer, which means the data will be a stand-alone object with less dependence on software applications, which in turn means fewer problems when transferring data between applications.

In this research, we encoded BIM data into RDF format before applying the clustering techniques. This step serves three