INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 5)
Available online at: www.ijariit.com

A text mining based automated BIM model data checking system

Rong Wen
davidemail.wen@gmail.com
Singapore Institute of Manufacturing Technology (SIMTech), Singapore

Long Xiao
xiao_long@simtech.a-star.edu.sg
Singapore Institute of Manufacturing Technology (SIMTech), Singapore

ABSTRACT

Building Information Modeling (BIM) has been widely used in the construction industry describing geometry of the building, spatial relationships and geographic information, as well as the quantities and properties of BIM objects and their properties. With large amounts of data exported from building information models, BIM data checking usually costs large amounts of man-hours to guarantee BIM model quality. In order to improve accuracy and efficiency for model data checking, applying machine learning techniques is highly expected. The machine learning algorithms can be used by combining data from many buildings, the characteristics, and location of flats to automatically detect flaws of building design, or even the likelihood of construction delays. Anomaly detection is useful to pinpoint modeling errors. In this project, semantic natural language processing technologies embedded with a machine leaning engine is developed to realize (1) automated object and property information retrieve from the massive BIM mode data, (2) similar objects recognition and naming consistence checking, and (3) automated BIM model data completeness and accuracy checking. The machine learning engine is wrapped by a dashboard-based user interface to facilitate manual object/property checking or batch data checking. The specially designed modules of the machine learning engine are allowed to be customized to support different construction regulatory requirements.

Keywords: *BIM, Text Mining, BIM Data Checking, Machine Learning And Smart Decision-Making*

1. INTRODUCTION

Building Information Modeling (BIM) is an intelligent 3D model-based process that gives architecture, engineering, and construction (AEC) professionals the insight and tools to more efficiently plan, design, construct, and manage buildings and infrastructure. BIM includes the geometry of the building, its spatial relationships and geographic information, as well as objects and properties of BIM models' components. With large amounts of data generated from BIM models, the manual process of construction regulatory requirement compliance checking could be time consuming, costly and error-prone [1]. Based on the survey [2], large amounts of money are spent on BIM model data compliance checking every year. Failure to comply with building regulations may also result in fines, penalties, and even criminal court summons and prosecutions [1] [3].

A way to partially avoid bottlenecks of the model data checking on the manually entered information is to implement data-centric workflows [4], which relies on automatic or semi-automatic solutions to data mining, analysis and explore the insights of data in BIM models. To process the large amounts of data exported from building information models, applying machine learning techniques for automated model data check is highly expected. Classification algorithms, anomaly detection, and even time series analysis can be used with BIM. The machine learning algorithms can be used by combining data from many buildings, the characteristics, and location of the flats to find flaws of building design, or even the likelihood of construction delays [5]. On the other hand, anomaly detection is very useful to pinpoint modeling errors, which can be used for maintenance predictions. In this project, a machine leaning engine, wrapped by a dashboard-based user interface, is to be developed for automated model data checking. The specially designed modules of the machine learning engine are allowed to be customized to support different types of construction requirements and rules checking.

2. SCOPE AND OBJECTIVES

The aim of this study is to develop a machine learning engine, using the semantic natural language processing (NLP) technologies, to realize (1) automated object and property information retrieve from the massive BIM mode data, (2) similar objects recognition and naming consistence checking, and (3) automated BIM model data completeness and accuracy checking. The scope of R&D work includes the followings:

- Develop BIM model data import and preprocessing modules.
- Develop a machine learning engine where five key modules including text processing module, named entity recognition, regexes pattern checking, lexical semantic relation and unsupervised lexicon-based checking.
- Develop a machine-user interface for data visualization of BIM model data checking and interactive data checking.

3. COMPARISON WITH SOLUTIONS IN THE MARKET

a) Revit Model Checker allows users to customize logic checking on model elements via XML. However, it has limited capability in checking information completeness and accuracy. There is lack of semantic checking on the object/property entities.

b) Dynamo allows users to programmatically customize model properties and features. There is lack of smart way of semantic checking on the completeness and accuracy. Currently, Revit python shell still doesn't support 3rd party package installation and implementation. They only support their own pyRevit API, thus the customizing python code with data analytics packages is unable to be implemented in the Revit python shell by IronPython. In addition, Dyanmo may hog up huge amounts of RAM and CPU usage, therefore only used for small amount of data processing.

c) BimSens is a tool to create apps and dashboards with read and write interactions between data sources & 3D model viewers. It's able to establish link between BIM models and database, visualizing objects and their properties, and allow for object filtering. However, it is mainly used for data visualization, lack of object/property naming consistence checking.

Comparing with manual compliance checking, the proposed solution is able to significantly reduce the time, costs, and errors of the compliance checking process. Generally, the state-of-the-art BIM model data compliance checking cannot achieve full automation because of relying on the use of hard-coded, proprietary rules, which requires major manual effort in extracting object/property information from BIM model data and coding this information into a rule format. The Fig.1 shows the bottle neck of the regulatory requirements/rules checking with BIM model data. We can see that digitalization and automatic interpretation of requirements are challenging. The use of hard-coded rules by users are effort-intensive and time-consuming, because of the large number of codes and regulations and their frequent revisions/updates.

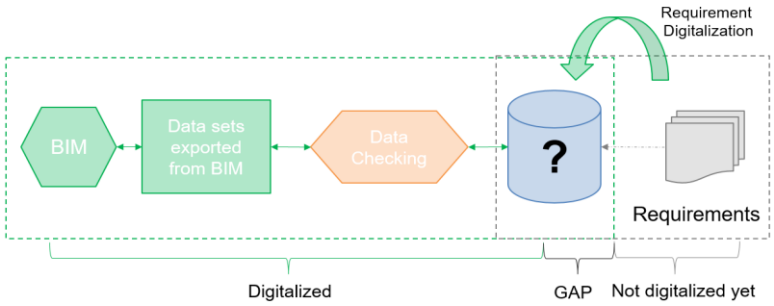

Fig. 1: Bottle neck of the regulatory requirements/rules checking with BIM model data.

4. METHODS AND MATERIALS

An overview of the proposed solution is shown in the Fig.1. BIM modes are repositories of a large number of objects, e.g., doors, columns, walls etc., where massive properties, materials, as well as their quantities and geo-reference information are included. In order to automatically process BIM model data, structured (e.g., quantity, geometry) and unstructured data (e.g., text, phrases) processing methods are required. Automated mode data checking on the object/property information will involve large amounts of text data searching, processing, feature extraction, similarity analysis, pattern checking, supervised or unsupervised semantic analysis. To this end, we propose a machine learning engine built on semantic natural language processing technologies for automated BIM model data completeness, accuracy and consistence checking. As shown in Fig.2, the machine learning engine takes input data exported from BIM models. The outputs are the results computed from BIM model data completeness checking, BIM model data accuracy checking, and BIM model data consistency and compliance checking. The input data can be directly exported from the BIM models or converted from Industry Foundation Classes (IFC) data. The machine learning engine consists of the following five key modules.

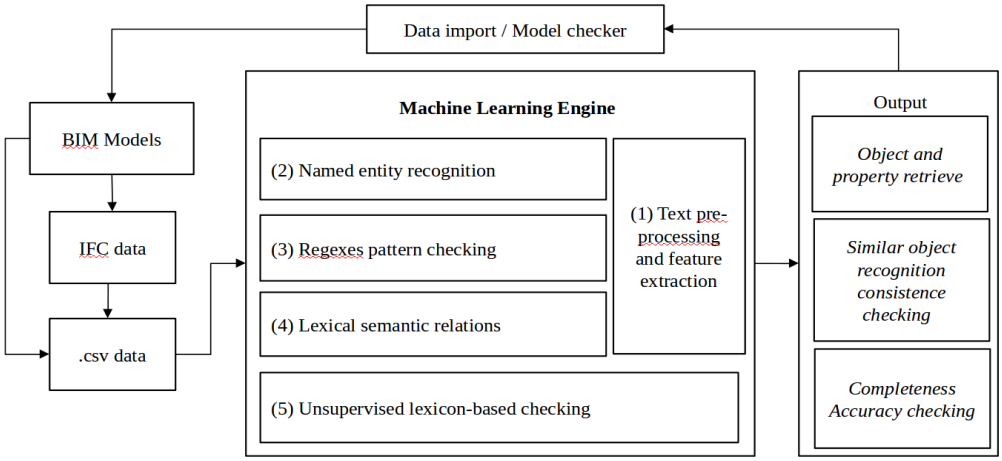

Fig. 2: Overview of machine learning engine for BIM model data checking.

4.1 Text processing module

This modules targets obtain textual components including words, phrases, and tokens which are basic building blocks of input that are fed into next stages of text mining. Word tokenization is implemented to split and segment phrase into constituent words. Word tokenization is a pre-processing step in cleaning and normalizing word based on its respective stems and lemma. Text normalization is a process that consists of a series of steps to wrangle, clean, and standardize textual data into a form that could be consumed by analytical process. One of the main challenges faced in text normalization is there might be presence of incorrect words in the text. The input model data may include spelling mistakes as well as words with several letters repeated. The word correcting process is also included in the text processing module. Other potential components that could be used in the text processing pipeline include parts of speech (POS) tagging, dependency-based parsing and constituency-based parsing.

4.2 Named entity recognition (NER) module

This module is used to automatically extract target entities (e.g., BIM models' objects and their properties) and classify them into certain categories for subsequent processing and analysis. Custom entities can be trained and established for classifying BIM model objects. In this way, the objects can be assigned labels to groups of tokens by which a wide range of named or numerical entities can recognize. Apart from the default entities, spaCy, an open-source software library for advanced natural language processing, can be used to add arbitrary classes to the NER model, by training the model to update it with newer trained examples. With named entity recognition module, objects and their property information in the BIM model data can be automatically retrieved, classified and statistically analyzed.

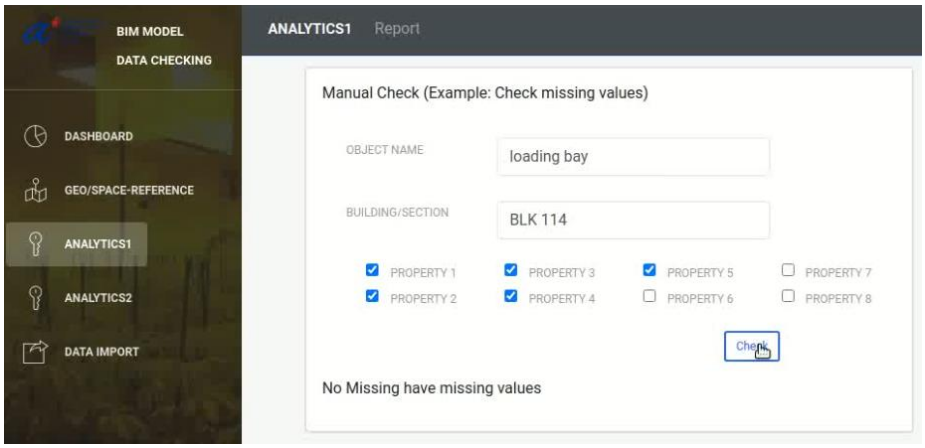

Fig. 3: Missing value checking on objects and their properties.

4.3 Regexes pattern checking

This is textual data matching engine where regular expressions (regexes) are used to create string patterns for search and find specific pattern matches in textual data. The patterns match the target strings based on the rules. The rules can be created and customized by users with flags to change the way the pattern matches can be executed. For example, match patterns ignoring case sensitivity, match any character including new lines, or matching Unicode-based characters (Fig.3). Regular expressions can also be compiled into pattern methods and then used with for pattern search and substitution in strings.

4.4 Lexical semantic relation

The lexical semantic analysis focuses on meaning and context exploration based on standard (e.g., Brown Corpus, WordNet, Collins Corpus and The Corpus of Contemporary American English (COCA)), or custom text corpora resource. Synonyms generated from this module could build relationship among the object/properties of BIM models with similar meaning and context, e.g., wash room, water closet, lavatory and toilette. Custom hyponyms and hypernyms are used to explore related BIM model objects by navigating through a hierarchy of subclass and superclass. Homonym and meronym analysis are used to find entities that contain a specific entity of users' interest. The semantic similarity analysis measures semantic similarity among the target objects and connect similar entities based on their semantic relationships.

4.5 Unsupervised lexicon-based checking

This module is used to improve performance of the consistence and compliance using an unsupervised learning method for words/phrase clustering. The clustering is based on established knowledgebases, ontologies, databases and lexicons that have detailed information. Most of these lexicons have a list of polar words with certain labels or categories associated with them. By applying various techniques like the position of words, surrounding words, and word/phrase distance calculation, scores are assigned to the different labels and categories. After aggregating these scores, the engine can suggest the most similar object/property names. More advanced analyses can also be done by grammatical analysis.

The module (1) prepare the data for the module 4.2, 4.3 and 4.4. The module 4.2 and 4.3 are used for object and property information retrieve, the module 4.3, 4.4 and 4.5 for similar object recognition and consistence checking, and the module 4.1, 4.3, 4.4 and 4.5 for model data completeness and accuracy checking. Depending on software/tools used in applications, the engine's output data can be imported by BIM modelling software, e.g. Revit, Revit Model Checker, to automatically correct or update objects information. Statistical summary of data checking results are displayed on the dashboard of the automated BIM model data system (Fig.4).

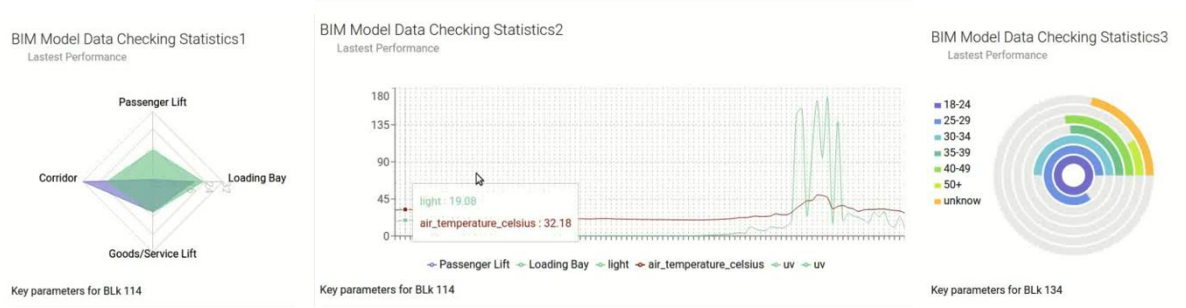

Fig. 4: Statistical summary of checking results.

5. CONCLUSION

In order to realize automated BIM model data compliance checking with construction requirements/rules, we propose and develop a machine learning engine incorporated with text mining processing modules, analysis modules and graph database. The NPL technologies are used to efficiently retrieve target values of objects/properties. Regexes and semantic text classification can be used to find synonyms for object/property consistency checking. A graph database is used to digitalize construction requirements describing relationship among objects/properties. In this way, an automated data checking close-loop can be completed from BIM model data export to requirement compliance checking.

6. REFERENCES

[1] J. Zhang and N. M.El-Gohary, "Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking Automation in Construction", Vol. 73, pp. 45-57, Jan. 2017,

[2] Department of Buildings, Hearing on the Fiscal 2016 Preliminary Budget and the Fiscal 2015 Preliminary Mayor's Management Report, 2015 (bhttp://council.nyc.gov/html/budget/2016/Pre/dob.pdfN (Sept. 4, 2015)).

[3] L.A. Times, Public & Legal Notices, 2015 (bhttp://classifieds.latimes.com/classifieds?category=public_noticeN (Sept. 4, 2015)).

[4] T. Berners-Lee, Status: An Early Draft of a Semi-formal Semantics of the N3 Logical Properties, 2005 (bhttps://www.w3.org/DesignIssues/N3LogicN (Jun. 9, 2016))

[5] J. Zhang, N. El-Gohary, Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking, J. Comput. Civ. Eng. (2013), http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000346, 04015014.

