ITcon

www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

MAPPING BIM SCHEMA AND 3D GIS SCHEMA SEMI-AUTOMATICALLY UTILIZING LINGUISTIC AND TEXT MINING TECHNIQUES

SUBMITTED: July 2014
REVISED: January 2015
PUBLISHED: January 2015 at http://www.itcon.org/2015/14
EDITOR: Turk Ž.

Jack C.P. Cheng, Assistant Professor
The Hong Kong University of Science and Technology
cejcheng@ust.hk

Yichuan Deng, PhD Candidate
The Hong Kong University of Science and Technology
ycdeng@ust.hk

Chimay Anumba, Professor
The Pennsylvania State University
anumba@engr.psu.edu

SUMMARY: *The interoperability between BIM (Building Information Modeling) and 3D GIS (Geographic Information System) can enhance the functionality of both domains. BIM can serve as an information source for 3D GIS, while 3D GIS could provide neighboring information for BIM to perform view analysis, sustainable design and simulations. Data mapping is critical for seamless information sharing between BIM and GIS models. However, given the complexity of today's BIM schemas and GIS schemas, the manual mapping between them is always time consuming and error prone. This paper presents a semi-automatic framework that we have developed to facilitate schema mapping between BIM schemas and GIS schemas using linguistic and text-mining techniques. Industry Foundation Classes (IFC) in the BIM domain and City Geography Markup Language (CityGML) in the GIS domain were used in this paper. Entity names and definitions from both schemas were used as the knowledge corpus, and text-mining techniques such as Cosine Similarity, Market Basket Model, Jaccard Coefficient, term frequency and inverse document frequency were applied to generate mapping candidates. Instance-based manual mapping between IFC and CityGML were used to evaluate the results from the linguistic-based mapping. The results show that our proposed name-to-definition comparison could achieve a high precision and recall. Results using different similarity measures were also compared and discussed. The framework proposed in this paper could serve as a semi-automatic way for schema mapping of other schemas and domains.*

KEYWORDS: *Building information modeling (BIM), City Geography Markup Language (CityGML), Industry Foundation Classes (IFC), Interoperability, Schema mapping*

REFERENCE: *Jack C.P. Cheng, Yichuan Deng, Chimay Anumba (2015). Mapping BIM schema and 3D GIS schema semi-automatically utilizing linguistic and text mining techniques. Journal of Information Technology in Construction (ITcon), Vol. 20, pg. 193-212, http://www.itcon.org/2015/14*

1. INTRODUCTION

Building Information Modeling (BIM) leverages 3D object-based digital models to store and exchange building information. BIM can facilitate information exchange processes and allow people to better collaborate and save cost (Eastman et al., 2008). Users can easily access, modify, and create information in a BIM model for construction projects, supporting convenient and seamless collaboration during the process. On the other hand, 3D Geographic Information Systems (GIS) are able to model city objects in high Level of Detail (LoD), providing a platform which is rich in data and easy for collaboration. For example, the City Geography Markup Language (CityGML) schema is able to store semantic information in addition to geometric and geographic information (Gröger et al., 2012). Both BIM and 3D GIS models store 3D geometry data and semantic data of buildings. Therefore, it is possible to link them and convert them to each other.

BIM and GIS models can benefit from each other. It has been shown that construction activities require data from GIS models to perform operations such as automatic site layout planning (Su et al., 2012), construction activities tracking (Cheng and Chen, 2002), and waste management (Robinson and Kapo, 2004). A comprehensive review about the application of GIS in construction activities was shown in (Bansal, 2007). Meanwhile, BIM models and CAD data are valuable data sources for reconstruction of 3D GIS scenes (Nagel et al., 2009, Benner et al., 2005). While there have been considerable amount of research efforts on the integration between BIM models of different data schemas (Wang et al., 2008, Wang et al., 2007, Garrett et al., 2004), little attention has been given to the integration between BIM models and GIS models.

Data integration between BIM schemas and GIS schemas is challenging because the schemas in the two domains are designed and created for completely different purposes. Particularly, the mapping between semantic information in BIM schemas and GIS schemas is essential for seamless data transformation, yet the complexity of these schemas make the mapping process time consuming and error-prone. Although both BIM and GIS are related to the built environment, BIM often focuses on the detailed building components and project information such as cost and schedule, whereas GIS often focuses on the geographical information and shape of buildings and building components. Therefore, BIM schemas and GIS schemas may use different perspective and terminology to represent the same entities. In addition, the same entities may be represented in different levels of detail in the two domains. To deal with these challenges, schemas from the BIM and GIS domains should be studied and a mapping framework between the schemas should be developed. The data standard chosen to represent BIM and GIS in this paper are Industry Foundation Classes (IFC) and CityGML as they are the representative data standard in the BIM and GIS domains, respectively.

There have been several attempts to map IFC with CityGML. EI-Mekawy et al. (2010) merged IFC and CityGML into a schematic model called Unified Building Model (UBM), in which all entity definitions are extended according to the entity definitions in the two schemas (Isikdag and Zlatanova, 2009). Nagel et al. (2009) proposed a two-process approach to mapping between GIS models and IFC using CityGML as a medium. In the proposed approach, GIS models were transformed into CityGML, and then the mapping between CityGML and IFC was developed (Nagel et al., 2009). The challenging part of the mapping, as reported, was that the process would include a 1-to-n mapping between the two schemas. The transformation from Boundary Representation (BRep) in CityGML to Constructive Solid Geometry (CSG) in IFC could also be difficult and requires a component recognition pattern. Converting from BRep, Swept Solid and CSG in IFC to BRep is also challenging. Wu and Hsieh (2007) introduced a transformation algorithm using a coordinate system transformation matrix (Wu and Hsieh, 2007). However, the semantic information mapping has not been solved in these approaches. None of these research efforts fully utilized the rich entity definitions in the two schemas, which are useful for data mapping. By comparing definitions of entities from IFC and CityGML, the similarity of entities could be discovered, thus providing a new way for mapping discovery of entities between the two schemas.

In this paper, we propose a linguistic-based methodology framework for semi-automatically mapping BIM and 3D GIS schemas. The linguistic-based method uses text-mining techniques to discover the relatedness of entities through their names and definitions. Component-based manual inspection generates the results for validation of the linguistic-based method. By comparing the results of the linguistic-based method and the component-based manual inspection, we can discover the portion of mappings that could be generated from linguistic-based method and proof the effectiveness of the proposed method. This paper is structured as follows: Section 2 introduces the two data standards and the challenges to map between them. Section 3 presents our proposed methodology framework using linguistic-based approach. The results are discussed in Section 4, which is followed by a number of conclusions in Section 5.

2. RESEARCH BACKGROUND AND RELATED WORK

2.1. Introduction to IFC and CityGML

IFC is a major data exchange standard in BIM. Initiated by buildingSMART (formerly International Alliance for Interoperability, IAI) in 1994, IFC has now become a formally registered international standard as ISO/PAS 16739. IFC can satisfy the information creation, storage and exchange needs for different stakeholders by means of well-structured entities covering almost all aspects of construction activities. It supports object-oriented three-dimensional models which are also rich in semantics. IFC is now supported by most commercial BIM software and supports various geometric representations of building parts. The representation may be one or a combination of CSG, Swept Solid, and BRep. In addition, the rich semantic information in IFC may help in efficient collaboration and decision making. BIM models based on IFC could be used not only in the construction phase, but also in the pre-construction phase and operation and maintenance phase, such as in feasibility studies, tendering (Ma et al., 2011), code checking (Pauwels et al., 2011) and operation management (Hassanain et al., 2001).

CityGML, on the other hand, is a relatively new data standard in the GIS domain for 3D GIS models. It was developed in 2002 by the Special Interest Group 3D (SIG 3D) of the initiative Geodata Infrastructure North Rhine-Westphalia in Germany. The SIG 3D is an open group consisting of 70 companies, municipalities and research institutions working on the development and commercial exploitation of interoperable 3D models and geo-visualization. CityGML is now a standard formally accepted in the Open Geospatial Consortium (OGC), an international standards organization which develops and promotes open standards for GIS and geospatial content and services. CityGML has been designed as a semantic model language which represents not only 3D geometry, but also semantic information (e.g. name, address and construction time). It is based on Geography Markup Language (GML), which gives the data standard the potential to easily integrate with other modeling languages in the GIS domain. CityGML supports five Levels of Details (LoDs) that vary from LoD0, which is basically a regional 2D map, to LoD4, which models the inside details of buildings. CityGML also supports Application Domain Extensions (ADE) in which users can create their own extensions for their particular applications. Different LoDs and ADEs could broaden the application area of CityGML.

2.2. Challenges in mapping between IFC and CityGML

IFC and CityGML have different terminology, entities and data representation approaches. Since IFC and CityGML are from different domains, they use different sets of terminologies to represent concepts. For example, the "*room*" concept in CityGML has a corresponding "*IfcSpace*" entity in IFC, which uses a different term. Different terminologies in IFC and CityGML make it difficult to perform a direct name-to-name mapping. Moreover, the entity definitions in IFC and CityGML are different in terms of content and scope. Both IFC and CityGML contain geometric information and semantic information, but IFC has a much richer scope and more entities. IFC 2x3 covers nine domains in the AEC industry, such as structural analysis domain and construction management domain, while CityGML focuses on representing the shape and some relevant information of objects. Besides, although IFC can be expressed in IfcXML format, IFC is fundamentally an EXPRESS based schema, in which every entity refers to or is referenced by other entities, while CityGML is an XML (Extensible Markup Language) based schema. The different schema data structures make it even harder to have a direct one-to-one mapping. The mapping of IFC and CityGML contains two major parts: (1) the mapping of geometric information and (2) the mapping of semantic information. Three issues must be addressed while the integration is being performed: (1) the mapping of data structure, (2) the mapping of values and representations, and (3) the mapping of entities.

Firstly, the mapping of EXPRESS language and XSD (XML Schema Definition) must be developed to resolve the difference in data structure. IFC employs an EXPRESS based modeling language, in which entities are linked to each other. One entity could be referred by various entities and have different meanings in different scenarios. However, the schema of CityGML is defined in XSD, in which a hierarchy structure of entities and elements could be generated. It is thus hard to perform a direct mapping of entities. When mapping entities, the context of this entity should also be considered in the mapping process.

Secondly, the mapping of values and representations involves the transformation of different coordinate systems and the transformation between BRep and CSG/Swept Solid. CityGML employs a world coordinate system to represent objects, in which all the coordinate values are absolute and do not refer to other objects. However, in IFC, each object has a local placement system relevant to other objects. For example, the local placement system of a window may be referenced to a wall placement system, while the wall placement system may be referenced to a building storey placement system. In addition, CityGML utilizes BRep for the object representations, while in IFC, BRep, CSG and Swept solid can be used to represent objects. The mapping between CityGML and IFC contains the transformation of CSG or Swept Solid to BRep, and vice versa. Given the CSG or Swept Solid representation, the boundary of the surfaces of the object will be calculated, and then the boundary is transformed into BRep in CityGML. Transforming from CityGML BRep to CSG or Swept Solid, however, is even more challenging.

The transformation of coordinate systems and the transformation between BRep and CSG/Swept Solid deal with the mapping of geometry only, and have been attempted in various efforts. On the other hand, mapping of entities considers both the geometrical and semantic information and is rarely tackled completely. Since the number of entities to be compared and inspected is large, computer-aided semi-automatic ways would be helpful. There are 1008 IFC entities and 608 CityGML entities defined in the schemas. If we simply inspect existing instances, the mapping between IFC and CityGML might not be complete. Moreover, some entity mappings cannot be discovered by simply looking at their entity names. For example, both the "*IfcPolyLoop*" entity in IFC and the "*LinearRing*" entity in CityGML are the entity to capture a set of coordinate points for a polygon representation, and therefore should be mapped to each other (i.e. a true match). However, if we simply compare their names, this mapping cannot be discovered due to the use of different terminology. The two schemas also utilize diverse vocabularies and different definitions to define the same entities. The complexity of the data schemas requires the development of ways to perform semi-automatic mapping considering the context of the entities. This paper will mainly focus on this issue.

2.3. Semi-automatic ways to perform schema mapping

Automatic data transformation is desirable in the integration between BIM and GIS. In order to achieve automatic data transformation, the schemas of BIM and GIS should be compared and mapped. According to the extent of the computer-aided techniques, mapping discovery could be performed by either manual or semi-automatic methods. No completely automatic schema mapping case has been reported so far since schemas are designed for different purposes and use diverse terminology. Different schemas may use the same terms to represent different meanings, or use different terms to represent the same concept. One manual instance-based mapping discovery between IFC and CIMsteel Integration Standards (CIS/2) was reported in (Lipman, 2009). The authors inspected the entities inside the two schemas and found that some of the mappings could be discovered according to their entity names. For instance, the IFC entity "*IfcCartesianPoint"* and the CIS/2 entity "*Cartesian_point"* are related (Lipman, 2009). The mapping discovery according to entity names is therefore one approach to finding related entities in heterogeneous data schemas.

Although the manual method performed by domain experts can guarantee accuracy, it is time consuming. Computer-aided methods have thus been introduced to find related entities in other schemas. Wang et al. (2008) found related entities between IFC and Automated Equipment information eXchange (AEX) by employing domain constraints, which were expanded definitions and explanations of entities. For instance, the constraints "has casing", "works on air", and "has rotation" were assigned to the entity "fan". Wang et al. (2007) also introduced a semi-automatic approach for mapping between different IFC versions. They looked into the structure and attributes of the entities in order to find the version differences. However, the comparison of the structure and attributes can only be applied for schemas with sufficient structure similarities. This technique is thus not applicable for mapping between CityGML and IFC, since they have very diverse structures. Lawrence et al. (2010) and Lawrence et al. (2014) developed mappings of cost information to BIM by coordinating data from heterogeneous domains using constraints in schemas of cost and BIM schema.

Rahm and Bernstein (2001) reviewed the approaches for automatic schema mapping and concluded that linguistic methods were common approaches in the schema-level mapping. As shown in FIG. 1, linguistic-based mapping may involve the evaluation of entity name similarity, the evaluation of entity description (definition) similarity, and the use of information retrieval (IR) techniques such as word frequencies. None of the semi-automatic methods mentioned above for BIM and GIS schemas considered entity definitions or IR techniques for mapping discovery. Pan et al. (2008) and Cheng et al. (2008) attempted to use text mining techniques for schema matching among IFC, CIS/2 and AEX by employing a domain-specific document corpus for the consideration and similarity comparison of entity linguistic semantics (Cheng et al., 2008, Pan et al., 2008). However, these efforts did not consider a direct comparison and leverage of the entity definitions. Both CityGML and IFC have rich definitions and annotations for their entities in the schemas. These entity definitions constrain the entity, thus providing an opportunity for automatic linguistic-based mapping discovery with the aid of IR and text mining techniques.

The same entity terms like “window” may be defined in different expressions in heterogeneous schemas, but their definitions are still similar and share many wordings. Therefore, entities with more similar definitions are more likely to be identical or related. This similarity will lead to a semi-automatic way of mapping discovery between IFC entities and CityGML entities. In this study, a linguistic based schema mapping framework is developed to prove the relationship between definition similarity and entity mapping and to provide schema mapping suggestions semi-automatically. Name similarity, description similarity and IR techniques were used for semi-automatic schema mapping between IFC and CityGML. The results could limit the search space of the schema mapping between IFC and CityGML. First, entity definition was extracted and compared using text mining techniques. IR techniques of Term Frequency and Inverse Document Frequency were then used. Finally, entity name was used to compare the name-name similarity and name-definition similarity. The results from the framework are verified by true match results for IFC and CityGML generated from manual mapping. The details of the methodology will be presented in the following section.

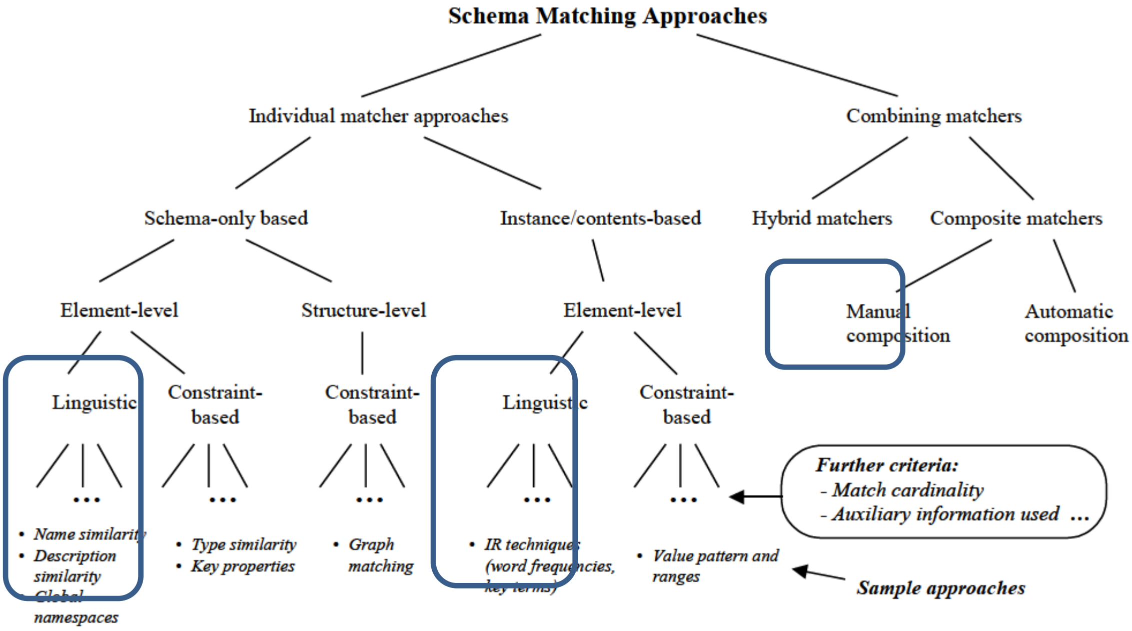


Fig. 1 Classification of schema matching approaches (Rahm and Bernstein, 2001) (Bold rounded rectangles show the methods used in this paper)

3. THE PROPOSED LINGUISTIC-BASED SCHEMA MAPPING FRAMEWORK

This section presents the linguistic-based approach that we developed for semi-automatic mapping of the entities of heterogeneous schemas. The instance-based approach, which involves manual inspection of schemas and instances, are used to evaluate the proposed linguistic-based approach and will be discussed in Section 4. The framework and workflow are illustrated in FIG. 2.

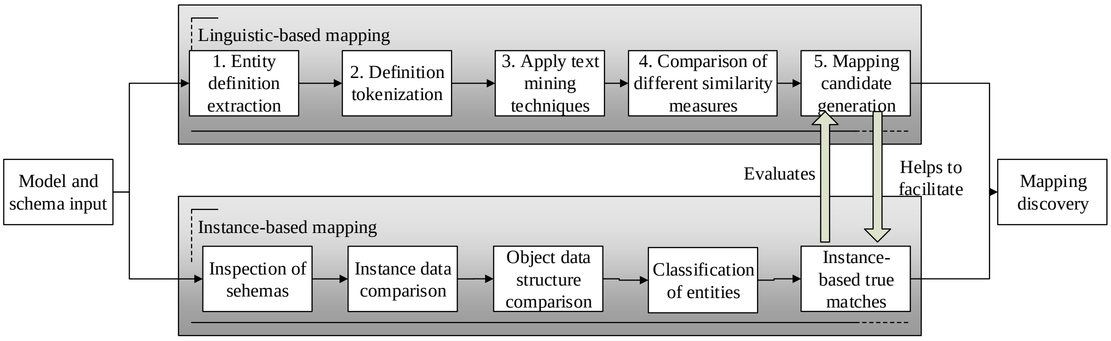

Linguistic-based mapping
1. Entity definition extraction
2. Definition tokenization
3. Apply text mining techniques
4. Comparison of different similarity measures
5. Mapping candidate generation
Model and schema input
Evaluates
Helps to facilitate
Mapping discovery
Instance-based mapping
Inspection of sehemas
Instance data comparison
Object data structure comparison
Classification of entities
Instance-based true matches

Fig. 2 Workflow of the linguistic-based and instance-based methods and their relationship

3.1. Procedures for the linguistic-based method

The linguistic-based method uses the results of text mining techniques to perform relatedness analysis and facilitate the mapping discovery process. The entity definitions in schema documents were extracted and compared to entity definitions in the other schema. Pairs with higher similarity results are likely to be identical or related. To evaluate the similarity of the entity definitions, Cosine Similarity, Jaccard Similarity Coefficient, and Market Basket Model were used.

3.1.1. Entity definition extraction

The first step for the calculation is entity definitions extraction. There are 1008 entities in the IFC schema, and all of them have descriptions and definitions. For the entities in the CityGML schema, considering their referencing to the GML schemas, 607 entities with definitions were found in the documentation of the CityGML and GML schemas. All the 607 entities have descriptions and definitions, which could be extracted directly from these schemas because the CityGML schema is represented in XSD (XML Schema Definition) format.

3.1.2. Definition tokenization

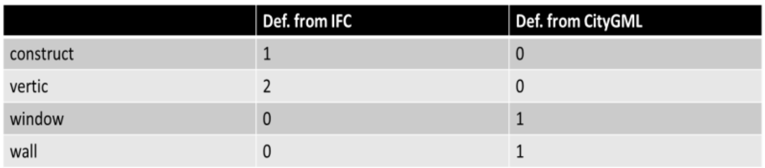
The second step is the tokenization of the entity definitions. All the stop words in the entity definitions were removed and the remaining text was stemmed for further calculation. Stop words are those that occur so frequently that they may not be as relevant to the query as the query to the whole document (Strzalka et al., 2011). Some commonly seen stop words are “is”, “for” and “to”. The stop words interfere with the accuracy of similarity calculation results as they appear too often and may raise the score of some calculations. A list of 450 common stop words was used and all the stop words in the definitions were removed. The remaining text in the definitions was then stemmed. Stemming is the process to change words to their stem or base form (Willett, 2006). For example, “definitions” is changed to “defin” whereas “coordinated” is changed to “coordin”. The stem of a word is not similar to the morphological form of the word, but it ensures all related words have the same stem. The Porter Stemming algorithm was adopted in this study to find the stems (Willett, 2006).

3.1.3. Application of text mining techniques

The third step is to formalize these stemmed definitions into feature vectors for further analysis. The feature vectors were generated as follows: if *concept n* (e.g. window) appears *m* times in the definition, the n-th value of the feature vector of definition would be *m*. An example is shown in FIG. 3.

Text *i* (Definition of *IfcWindow* in IFC): Construction for closing a vertical or near vertical opening …

Text *j* (Definition of *windowtype* in CityGML): Type for windows in walls ….



	Def. from IFC	Def. from CityGML
construct	1	0
vertic	2	0
window	0	1
wall	0	1

$$v_i = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix} \quad v_j = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Fig. 3 Illustration of the feature vector

After all these preparation processes, the entity definitions were ready for comparison. A one-to-one comparison between entities from IFC and those from CityGML was performed by calculating the Cosine Similarity, Jaccard Similarity Coefficient and Market Basket Model scores. Cosine Similarity is a measurement of similarity in the field of text mining. It is a non-Euclidean distance measure between two vectors. If two vectors v_i and v_j are given, the Cosine Similarity of the two vectors can be expressed as:

$$Sim(i, j) = \frac{v_i \cdot v_j}{|v_i| \times |v_j|} \quad (1)$$

By its nature, the Cosine Similarity has a range of [0, 1]. The maximum score of 1 indicates that the two concepts *i* and *j* have almost identical features and have the highest similarity.

The Jaccard Similarity Coefficient is a measurement of the overlap between the feature vectors v_i and v_j of two concepts. Suppose N_{11} refers to the number of features so that both feature vectors contain non-zero values, N_{10} refers to the number of features that appear in v_i and do not appear in v_j, and N_{01} refers to the number of features that appear in v_j and not in v_i. The Jaccard Similarity Coefficient score between the concepts *i* and *j* can be calculated with the following equation:

$$Sim(i, j) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}} \quad (2)$$

The Jaccard Similarity Coefficient score cannot exceed 1, which indicates that the two feature vectors are the same. The Jaccard Similarity Coefficient score is popular for the measurement of relatedness of a term-to-term pair. It can return the overlap of two terms, so it is an efficient way to compare two terms (Cheng et al., 2008).

The Market Basket Model is another data-mining technique to calculate the similarity of two concepts (Pan et al., 2008). Let N_{11} , N_{10} , and N_{01} have the same definitions as those for Jaccard Similarity Coefficient. Given two feature vectors v_i and v_j, the associate rule *i* to *j* without absolute notation could be calculated as:

$$Sim(i, j) = \frac{N_{11}}{N_{11} + N_{10}} - \frac{N_{11} + N_{01}}{n} \quad (3)$$

where n is the number of all features. The Market Basket Model returns a result in the range of (-1, 1). The value of -1 means all the features appearing in concept *j* do not appear in concept *i*. The value of 1 is another boundary value since the value of N_{11} +N_{01} could not be 0 (Cheng et al., 2008).

3.2. Consideration of entity name features

Besides the consideration of entity definitions, entity names were also considered for the comparison. According to (Lipman, 2009), some mappings can be discovered directly by comparing their entity names. Certain entity definitions do not repeat the entity names, so name comparison should also be considered. All the entity names were split into phrases and the “ifc” and “gml” prefixes in entity names were removed. For example, the entity name “*IfcWallStandardCase*” was split into “wall standard case”, and then tokenized and stemmed as described in Section 3.1.2. The entity names were compared to other entity names as well as entity definitions and different weights were assigned to different comparisons, as illustrated in FIG. 4.

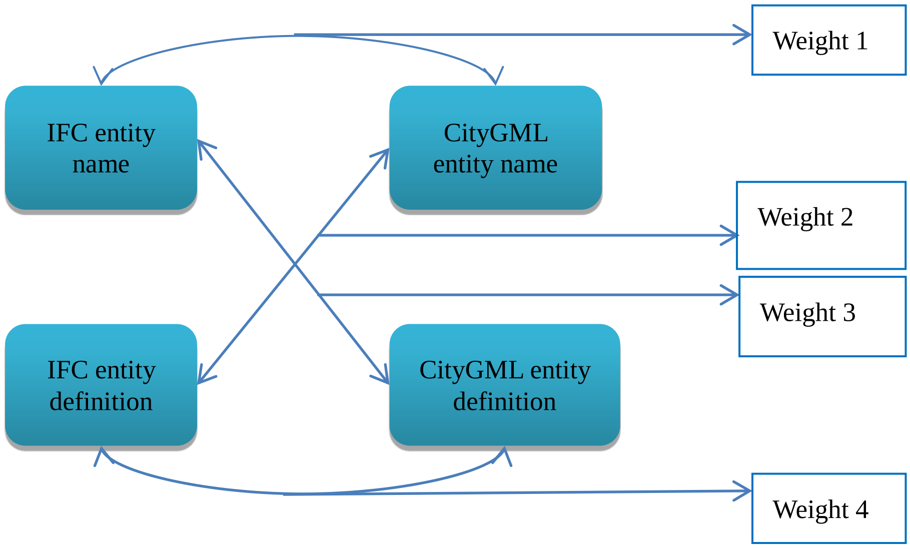


Fig. 4 Comparison of entity names and entity definitions between IFC and CityGML

3.3. Consideration of Term Frequency (tf) and Inverse Document Frequency (idf)

After the similarity scores were calculated, similar entities could be discovered by ranking the similarity analysis scores and setting the threshold for them. To improve the accuracy, Term Frequency (*tf*) and Inverse Document Frequency (*idf*) were introduced in the calculation. Term Frequency *(tf)* and Inverse Document Frequency *(idf)* improve comparison results by considering the frequency of a word in a single document while at the same time considering the inverse proportion of the word in all documents. This is represented as *tf-idf* and it is the weight of a word in the definitions. Words which are rarely seen in all documents will have a higher weight, while those commonly seen words are assigned a smaller weight. Given a single document *d* which belongs to a document pool *D*, the general formula to calculate *tf-idf* for a single word *w* is as follows.

$$(tf*idf)_{w,d} = f_{w,d} \times \log\left(\left|D\right| / f_{w,D}\right) \qquad (4)$$

where $f_{w,d}$ is the frequency of *w* in the document *d*, $\left|D\right|$ is the size of the document pool and $f_{w,D}$ is the number of documents in which *w* appears.

3.4. Metrics for results evaluation of the linguistic-based method

The results of the linguistic-based method were evaluated and compared using the precision and recall metrics. The true matches came from the manual instance-based method by domain experts. The precision and recall of this linguistic-based method can be calculated as:

$$\text{Precision} = \frac{|True\ Matches \cap Predicted\ Matches|}{|Predicted\ Matches|} \quad (5)$$

$$\text{Recall} = \frac{|True\ Matches \cap Predicted\ Matches|}{|True\ Matches|} \quad (6)$$

where the predicted matches are the results of the linguistic-based method while the true matches are given by the manual instance-based method. The higher the precision, the more likely are the candidates in predicted matches to be the true match. The higher the recall, the more likely this method can find the true matches.

3.5. Framework implementation

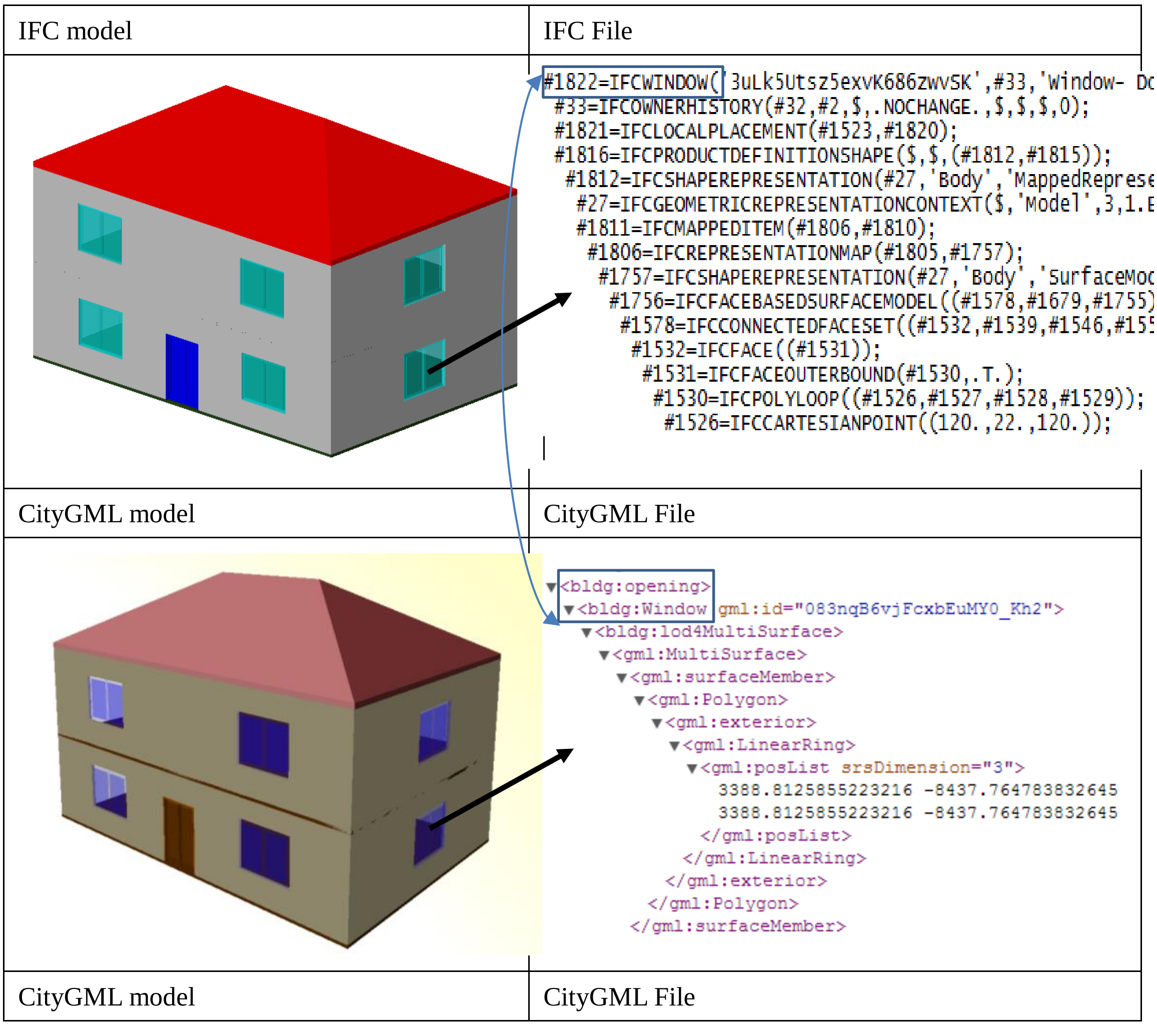
The proposed framework was implemented on a platform we developed using Java. The platform mainly consists of three parts: (1) parsers for XSD (XML Schema Definition) files of the CityGML schema and HTML files of the IFC schema, (2) a similarity comparison engine that calculates different similarity scores, and (3) a program that reports the comparison results to a spreadsheet and generates mapping candidates. XSD files are represented in the XML format and therefore could be parsed by standard XML parsers. In this study, the open-sourced JDOM 1.1.2 (Hunter and Lear, 2012) was used for developing the parser for XSD files of the CityGML schema. For the IFC schema, definitions of IFC entities were extracted from the IFC documentation HTML files (buildingSMART International 2007) using a HTML parser that was also developed based on JDOM. Tokenization of entity names and definitions is needed before similarity comparison can be conducted. In this study, the Porter Stemmer in Apache Lucene (Apache, 2012) was used for the tokenization process. After tokenization, a table of all the distinct tokens in entity names and definitions was generated, which was used for generating feature vectors for similarity comparison.

4. RESULTS AND DISCUSSION

As IFC has 1008 entities and CityGML has 607 entities, 611,856 (1008 x 607) IFC-CityGML entity pairs were generated for similarity analysis using the linguistic-based method. The similarity scores for each entity pair were then calculated using the methods presented in Section 3. The results were evaluated and compared with reference to the results from the manual instance-based method, which were taken as the true matches.

4.1. Validation of framework using results from the instance-based method

To validate the proposed linguistic-based framework, results from the instance-based manual mapping of entities were generated as true matches. In the manual instance-based method, domain experts refers to the class hierarchical relationship defined in the IFC and CityGML schemas and the object representation in IFC and CityGML models. IFC and CityGML models such as the models shown in FIG. 5 were studied and compared to obtain the true matches for result evaluation. IFC representation and CityGML representation of the same component were extracted from models and compared with each other. Take window components as an example. The IFC and CityGML entities used to represent a window are tabulated in TABLE 1. The entities are divided into three levels: (1) the object level, (2) the middle level, and (3) the value level. The object level contains the information of object name and ID. The middle level connects the values of objects such as length and height to the placement system and geometric representation. At the value level, all the entities store the key values of the window component, such as coordinates, length, and boundary. All the object representations can be divided into these three levels. The entities at the same level are considered to be the true matches, such as the "*IfcPolyline"* entity in IFC and the "*LinearRing"* entity in CityGML. By inspecting the entities representing the same object, the related entities (i.e. true matches) were located (e.g. "IfcWindow" in IFC and "window" in CityGML). However, the instance-based method is limited to the instances collected and may not cover all the possible true matches. On the other hand, the linguistic-based method can generate a list of candidates for a particular entity in a semi-automatic manner. Therefore, the search space for the instance-based method is limited.



```
#1822=IFCWINDOW('3uLk5Utsz5exvK686zwvSK',#33,'Window- Dc
 #33=IFCOWNERHISTORY(#32,#2,$,.NOCHANGE.,$,$,$,0);
 #1821=IFCLOCALPLACEMENT(#1523,#1820);
 #1816=IFCPRODUCTDEFINITIONSHAPE($,$,(#1812,#1815));
  #1812=IFCSHAPEREPRESENTATION(#27,'Body','MappedReprese
   #27=IFCGEOMETRICREPRESENTATIONCONTEXT($,'Model',3,1.E
   #1811=IFCMAPPEDITEM(#1806,#1810);
    #1806=IFCREPRESENTATIONMAP(#1805,#1757);
     #1757=IFCSHAPEREPRESENTATION(#27,'Body','SurfaceMoc
      #1756=IFCFACEBASEDSURFACEMODEL((#1578,#1679,#1755)
       #1578=IFCCONNECTEDFACESET((#1532,#1539,#1546,#155
        #1532=IFCFACE((#1531));
         #1531=IFCFACEOUTERBOUND(#1530,.T.);
          #1530=IFCPOLYLOOP((#1526,#1527,#1528,#1529));
           #1526=IFCCARTESIANPOINT((120.,22.,120.));
```

```
▼<bldg:opening>
 ▼<bldg:Window gml:id="083nqB6vjFcxbEuMY0_Kh2">
   ▼<bldg:lod4MultiSurface>
     ▼<gml:MultiSurface>
       ▼<gml:surfaceMember>
         ▼<gml:Polygon>
           ▼<gml:exterior>
             ▼<gml:LinearRing>
               ▼<gml:posList srsDimension="3">
                   3388.8125855223216 -8437.764783832645
                   3388.8125855223216 -8437.764783832645
                 </gml:posList>
               </gml:LinearRing>
             </gml:exterior>
           </gml:Polygon>
         </gml:surfaceMember>
```

Fig. 5 Sample IFC and CityGML models of the same building for obtaining true matches using the instance-based method

Table 1 The entities for representing a window in IFC and in CityGML

Level	IFC Entities	CityGML Entities
Object	IfcWindow	opening window
Middle	IfcOwnerHistory IfcLocalPlacement IfcProductDefinitionShape IfcShapeRepresentation IfcGeometricRepresentationSubcontext IfcMappedItem IfcRepresentationMap IfcAxis2Placement3d IfcShapeRepresentation IfcGeometricSet	MultiSurface surfaceMember Polygon
Value	IfcPolyline IfcCartesianPoint	LinearRing exterior

In the linguistic-based method, Cosine Similarity, Jaccard Similarity Coefficient and Market Basket Model were used to measure the relatedness between the IFC entities and the CityGML entities. The tf-idf and name features were also considered for evaluation. The results of the linguistic-based method showed its potential to locate the possible true matches. By comparing the entity definitions, similar entities were discovered. For example, the "*windowtype*" entity in CityGML and the "*IfcWindow*" entity in IFC have a Cosine Similarity score of 0.103, which is the highest score among all the other entities for the "*IfcWindow"* entity. The relatedness between the two entities is thus discovered. The linguistic-based methods could also locate those similar entities that are not related in entity names. For instance, the "curve" entity in CityGML and the "*IfcPolyline*" entity in IFC are highly related; however, they use different terminology for the entity names. The two entities resulted in a Cosine Similarity score of 0.583, which is also the highest among all the entities for the "*IfcPolyline"* entity. By looking at the results with high similarity scores, the highly related entity pairs could be discovered. More results are shown in TABLE 2. For instance, in the results from instance-based mapping, the "*IfcRepresentation*" entity stores the geometry of building objects and will be mapped to "*MultiSuface*", "*MultiCurve*" or "*MultiSolid*" in CityGML. The linguistic-based method also unveils this mapping relationship. As shown in TABLE 2, the "*MultiSuface*", "*MultiCurve*" or "*MultiSolid*" are within the first ten mapping candidates for the "*IfcRepresentation*" entity. Other samples in mapping discovered by the linguistic-based mapping also include the "*IfcBuildingElement*" to "*BuildingInstallationType*" and "*IfcFurnishingElement*" to "*BuildingFurnitureType*", which all have high rankings in the mapping candidates. The linguistic-based mapping considers the definitions of entities, so it can discover mapping pairs with different names, such as "*IfcPolyline"* and "*Curve*". In this sense, the linguistic-based method could provide a reasonable pool of potential candidate entity pairs to facilitate the mapping process. To evaluate the effectiveness of the linguistic-based mapping, the results are tested against true matches from instance-based mapping. By setting different thresholds for the true match generation, the precision and recall results could be generated and compared, as shown in FIG. 6 and FIG. 7. The evaluation was made in three steps. Firstly, for a given ranking threshold (e.g. 10), results from the linguistic-based method (e.g., "*IfcPolyLine*" with "*LinearRing*" with a rank of 9) were randomly picked. Secondly, domain experts were then asked to evaluate the result using true matches from the instance-based method. Finally, the precision and recall were calculated based on the numbers of true matches, predicted matches, and correctly predicted matches. For example, in TABLE 3, if ranking threshold is 10, the number of predicted matches is 5. As shown in FIG. 6, Cosine Similarity resulted in a higher precision than Jaccard Similarity Coefficient and Market Basket Model. The low precision of Jaccard Similarity Coefficient and Market Basket Model is due to the difference in vocabulary between IFC and CityGML. Both Jaccard Similarity Coefficient and Market Basket Model are based on the duplication of words in the text. If the two schemas utilize very different vocabulary to describe the same entity, Jaccard Similarity Coefficient and Market Basket Model would return a low score. The definitions of CityGML entities utilize a vocabulary of 1734 words, while those of IFC entities utilize 1603 words. The two schemas only share a list of 726 words, which are 42% of the words in CityGML and 45% of the words in IFC. Given such a low duplication in vocabulary, Jaccard Similarity Coefficient and Market Basket Model could not get a high score. On the other hand, the Cosine Similarity score not only considers the duplication of words in the two definitions, but also considers the duplication times of the words. This is also proved by FIG. 8, which shows that the average cut-off scores of Cosine Similarity are always higher than those of Jaccard Similarity and Market Basket Model at the same number of ranks.

Table 2 Example mapping results of the linguistic-based method

IFC Entity	Remark	CityGML Entity	Remark	Cosine Similarity	Rank
IfcWindow	The entity to store information about windows in walls in IFC.	**WindowType**	The entity to store information about windows in walls in CityGML.	0.103	1
IfcPolyline	The entity to store information about connected line segments with orientation	**Curve**	The entity to store information about curves	0.58	1
IfcRepresentation	The entity to define the geometry of objects	**MultiSurface**	The entities in CityGML to define the geometry of objects	0.26	5
IfcRepresentation		**MultiCurve**		0.26	6
IfcRepresentation		**MultiSolid**		0.26	7
IfcBuildingElement	Stores major part of a building, examples are foundation, floor, roof, wall	**BuildingInstallationType**	The entity to store information about building installations, such as chimneys, stairs, antennas	0.25	2
IfcFurnishingElement	The entity to store information about furniture in buildings	**BuildingFurnitureType**	The entity to store information about furniture in buildings	0.30	3
IfcBuilding	General information about building, such as address, height	**AbstractBuilding Type**	Stores general information about building, such as construction time, height	0.27	12
IfcAddress	Stores address information	**AddressType**	Stores address information	0.28	4
IfcBoundedSurface	The entity to store information of surfaces bounding an object	**CompositeSurface**	The entity to store information of many connected surfaces	0.58	1
IfcStair	The entity to represent stairs in IFC	**AbstractSurfaceType**	The abstract entity to store information about surfaces	0.15	1

4.2. Comparison among Cosine Similarity, Jaccard Similarity, and Market Basket Measures

The ranking according to Cosine Similarity for a window representation is shown in TABLE 3. It is discovered that the entities on the same level, such as the "*IfcPolyline*" entity in IFC and the "*LinearRing*" entity in CityGML, tend to have a higher rank. Some of the relatedness could not be discovered by simply looking at their entity names. For example, the "*IfcWindow*" entity in IFC and the "*opening*" entity in CityGML were regarded as true matches by domain experts. The entity pair is suggested in the linguistic-based method with a rank of 2, as shown in TABLE 3. However, the relationship could not be discovered by simply looking at their entity names due to the use of different terminology.

Table 3 Ranking result between the "IfcWindow" entity and other related entities in IFC and the "window" entity and other related entities in CityGML using Cosine Similarity

IFC Name to CityGML Text	opening	window	MultiSurface	surfaceMember	Polygon	exterior	LinearRing
IfcWindow	2	1	49	122	344	91	129
IfcOwnerHistory	66	294	79	36	305	509	214
IfcLocalPlacement	130	122	16	533	31	394	141
IfcProductDefinitionShape	478	605	140	537	281	420	265
IfcShapeRepresentation	317	147	119	305	348	484	28
IfcGeometricRepresentation-Subcontext	468	154	31	90	158	188	169
IfcMappedItem	484	605	1	545	81	414	455
IfcRepresentationMap	404	415	1	413	31	507	185
IfcAxis2Placement3D	413	410	204	513	143	558	105
IfcGeometricSet	505	605	37	180	51	242	97
IfcPolyline	460	605	149	79	16	158	9
IfcCartesianPoint	542	386	236	385	47	509	137

As shown in FIG. 7, the results using Cosine Similarity and Jaccard Similarity yield a slightly higher recall than the result using Market Basket Model. FIG. 7 also shows the correctness of the proposed linguistic-based approach. At the ranking threshold of 200, the recall value is 0.53, indicating we could find at least half of the true matches in the top 200 predicted matches. Note that IFC has 1008 entities. In other words, 50% of true matches could be identified by only inspecting 20% of the entities. The results of precision evaluation in FIG. 6 also indicate that if we inspect the top ten results of the Cosine Similarity measurement, all of them were correctly predicted matches. This shows the potential of the linguistic-based method. In the following, the results using *tf-idf* and entity name consideration will be evaluated.

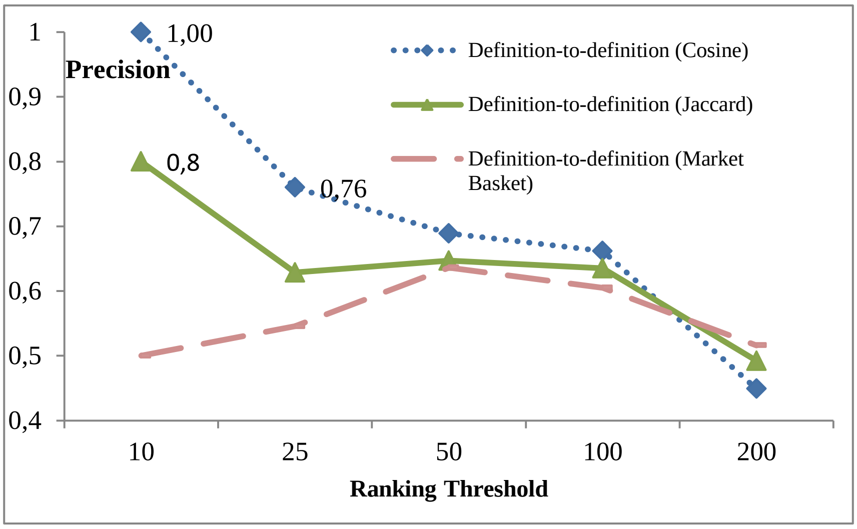


Fig. 6 Precision of the results using different similarity measures

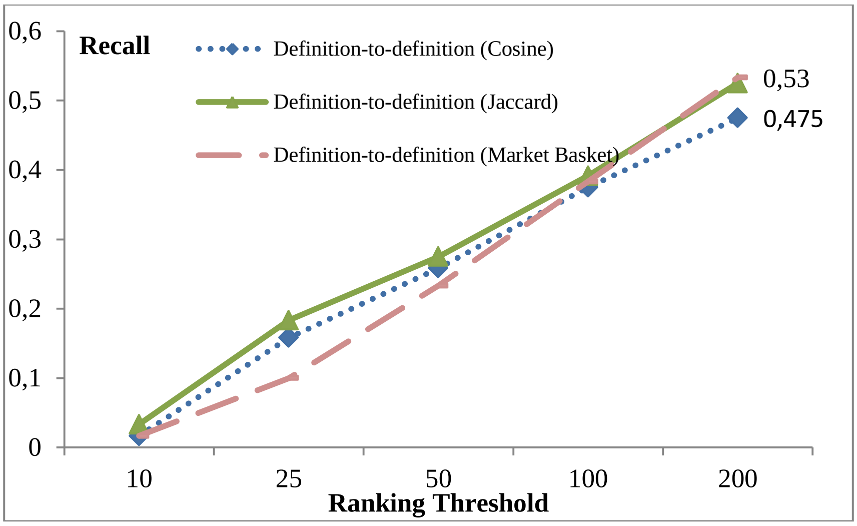


Fig. 7 Recall of the results using different similarity measures

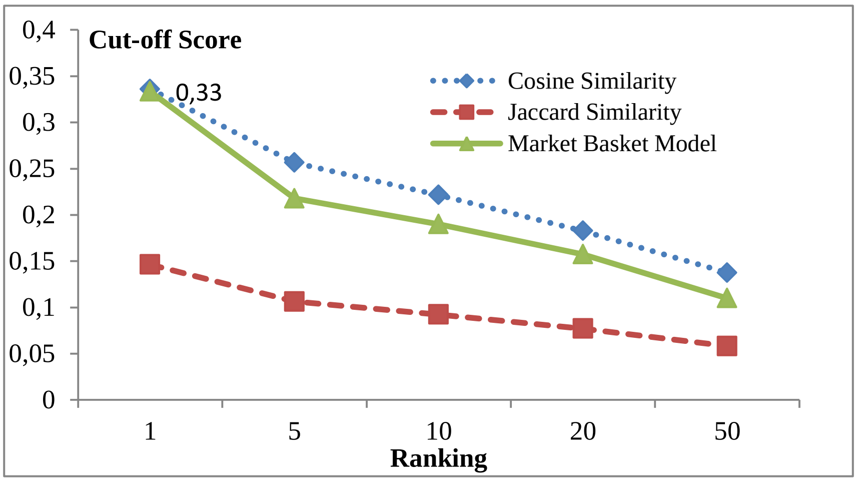


Fig. 8 Cut-off scores at different ranks of the results using different similarity measures

4.3. Evaluation of the entity name feature consideration

The entity names were also considered in the mapping because sometimes entity names do not appear in the definition to avoid repetition. The names were compared to other entity names as well as the definitions. As shown in FIG. 9 and FIG. 10, the name-to-name comparison could not generate a good score because the two schemas (IFC and CityGML) were developed for different domains and use different terminology to name objects (e.g. "*IfcPolyline*" in IFC versus "*LinearRing*" in CityGML). The Cosine Similarity scores of name-to-definition comparison were compared to those of definition-to-definition comparison. FIG.9 shows that IFC_name-to-CityGML_definition comparison has the highest recall while name-to-name comparison and CityGML_name-to-IFC_definition comparison have the lowest. FIG. 10 shows that name-to-definition comparisons and definition-to-definition comparison generally have a higher precision rate while name-to-name comparison has the lowest. The low performance of name-to-name comparison in terms of recall and precision indicates that we cannot simply use entity names in the schema mapping process, especially when the two schemas are from different domains and likely use different terminology. The results also indicate that definition-to-definition comparison performs better than name-to-name comparison, and consideration of entity names in definition-to-definition may improve the mapping accuracy. It is consistent with the conclusion made by (Lipman, 2009).

The consideration of entity names in definition-to-definition comparison was tested and evaluated. As discussed in Section 3.2, a combined score is calculated by putting different weights to name-to-name comparison, name-to-definition comparisons and definition-to-definition comparison. Since the definition-to-definition comparison between IFC and CityGML shows better precision and recall results while the name-to-name comparison shows the worst, a higher weight is put to definition-to-definition comparison and a lower weight is put to name-to-name comparison. In this case, we assigned a weight of 0.5, 0.3 and 0.1 to definition-to-definition comparison, name-to-definition comparison and name-to-name comparison, respectively. As shown in FIG. 9 and FIG. 10, the results using the combined scores show a higher precision than the name-to-definition comparison but a lower recall than the definition-to-definition comparison.

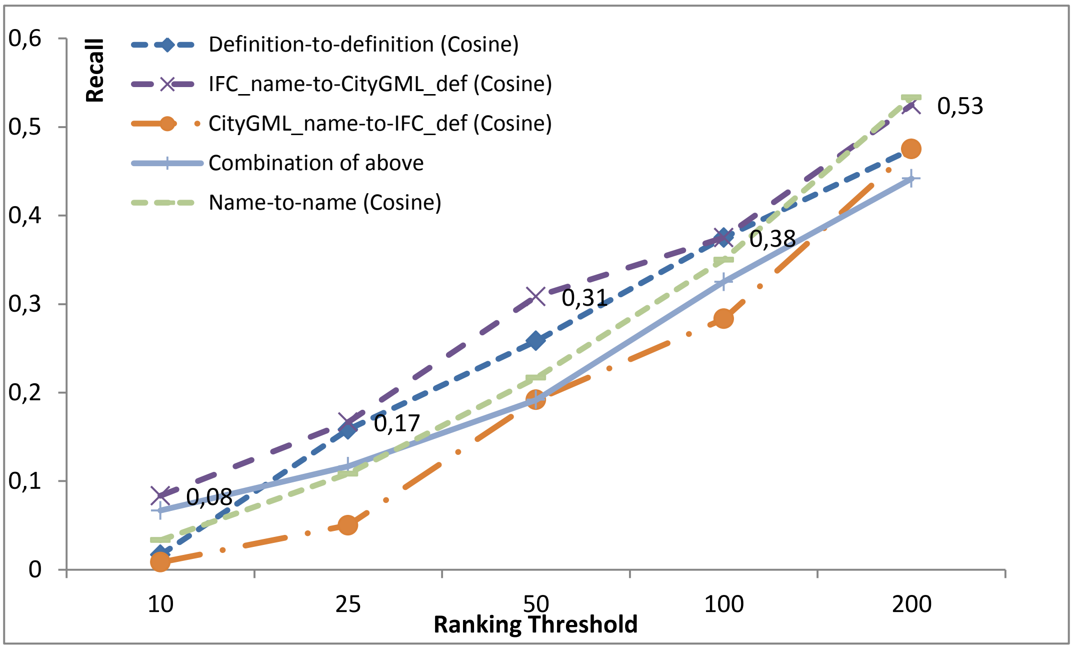


Fig. 9 The recall results considering entity names in comparison

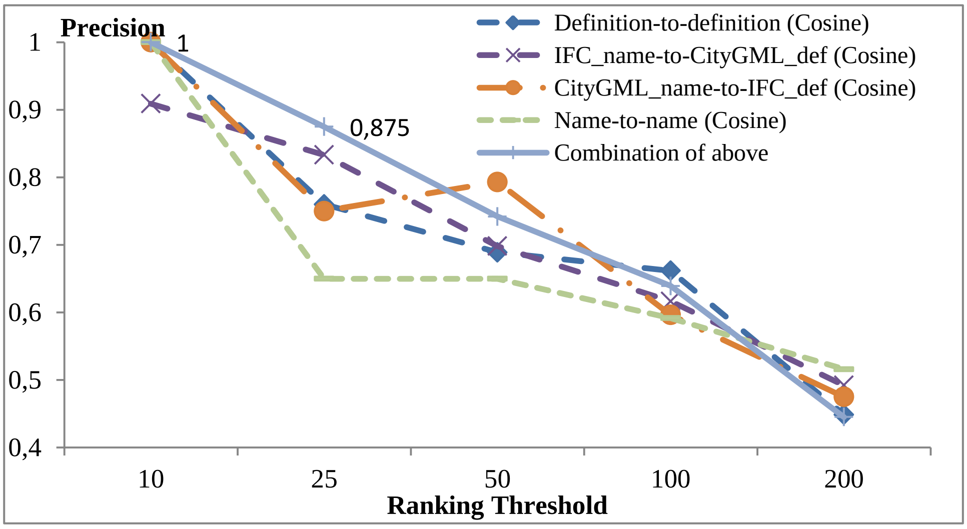


Fig. 10 The precision results considering entity names in comparison

4.4. Evaluation of the Term Frequency - Inverse Document Frequency (TF-IDF) Consideration

Consideration of *tf* and *idf* can improve the similarity analysis for document sets with a lot of duplications and commonly-appeared words. To evaluate the effect of *tf-idf* consideration in our approach, the results of definition-to-definition comparison and IFC_name-to-CityGML_definition comparison using Cosine Similarity are compared with those with the consideration of *tf-idf*. As shown in FIG. 13 and FIG. 14, the *tf-idf* consideration does not improve the results in terms of recall and precision. For the definition-to-definition comparison, the *tf-idf* consideration has substantially worsened the recall and precision results. This can be explained by the definition of *tf-idf*. The *tf-idf* technique heavily relies on the distribution of terminology utilized in the definitions. Of the 2611 words that appear in all the definitions, 81.7% of them only appear less than 10 times in all the documents. If all the definitions are considered to be one document, 78.3% of all words only appear less than 10 times. The high frequency of rare words reduces the effect of *tf-idf* measure, thus making the recall and precision results with the *tf-idf* consideration low. Nonetheless, for schemas in the same domain with similar sets of terminology, the *tf-idf* consideration is expected to improve the results.

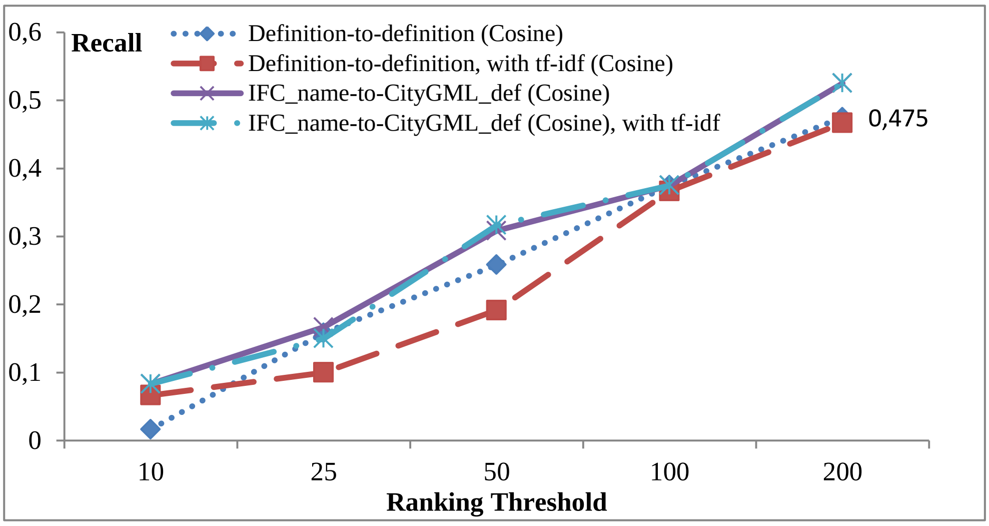


Fig. 13 The recall of similarity measurements with and without tf-idf consideration

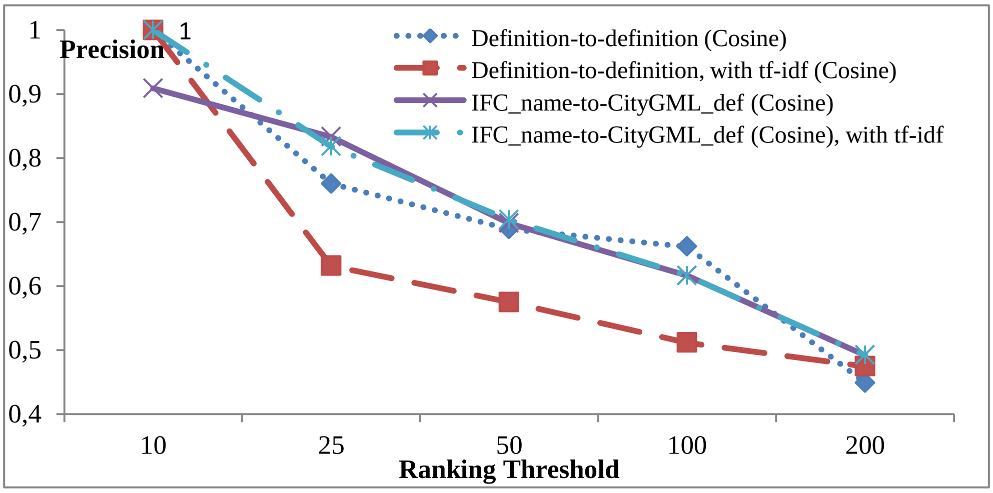


Fig. 14 The precision of similarity measurements with and without tf-idf consideration

4.5. Discussions on the Benefits and Limitations of the Linguistic-based Approach

Mapping heterogeneous schemas by inspecting individual entity names and definitions is labour intensive and may take weeks. However, semi-automatic approach using linguistic and text mining techniques could help limit the search space and complete within several minutes. Section 3 describes the methodology of the linguistic-based method and Section 4 illustrates the results of mapping candidates. The results show the following benefits of the linguistic-based method.

1. The linguistic-based method can generate reliable mapping candidates. Since the entities referring to similar concepts will likely have similar entity definitions, the linguistic-based method using the entity definitions and names could generate accurate mapping candidates although the entity names may use different terminology to represent the same concepts. As shown in FIG. 7, the recall at the ranking threshold of 100 can reach 0.375, which means that by inspecting the first 100 candidates from the result, we could find 37.5% of the true matches. Considering the large number of entities in IFC (1008) and CityGML (607), the results from the linguistic-based method narrows the search space and reduce the human effort for mapping discovery.

2. The linguistic-based method does not require the human domain knowledge. As the linguistic-based method automatically compares the semantic similarity among entities in heterogeneous schemas, people who perform the linguistic-based method are not required to possess the expertise in the domains of the schemas. Domain experts will be needed only after the first screen of entities using the linguistic-based method. It allows the linguistic-based method to be applied in domains other than the building and construction area which IFC and CityGML focus on.

3. The linguistic-based method could also suggest candidates for 1-to-M mapping. The traditional mapping methods on the entity level only consider the entity names and may not be able to find mapping between one entity and many entities. For instance, the entity "*AbstractOpening*" in CityGML could be mapped to "*IfcWindow*" or "*IfcDoor*", but the mapping could not be discovered by traditional name-to-name comparison. The linguistic-based mapping method, which utilizes the definitions of entities, could broaden the scope of comparison and generate more 1-to-M mapping results. Another example from the linguistic-based mapping results is the mapping between the CityGML entity "*AbstractBoundarySurfaceType*" and the IFC BRep entities such as "*IfcBoundedSurface*", "*IfcFaceOuterBound*", "*IfcFaceBound*", "*IfcBoundingBox*", and "*IfcClosedShell*". The 1-to-M mapping could be discovered in the top 25 mapping results.

Although the semi-automatic linguistic-based method could facilitate the mapping discovery process, the method has the following limitations that need further investigation.

1. The accuracy of the linguistic-based method depends on the preciseness of the entity names and definitions. This work assumes that the people who developed the schema and provided the entity definitions were domain experts and able to describe the entities using correct and appropriate words. We may include a wider range of entity definitions from other sources such as domain specific dictionaries or resources from the general domain such as the WordNet (Miller, 1995). However, we still believe that the same word may have different meanings in different domains, and domain specific resources and particularly resources specific to the mapping schemas would be the most appropriate.

2. The comparisons conducted in this study consider only the words in the entity names and definitions in the IFC and CityGML schemas. Since IFC and CityGML are from different domains, they may use different terminology for the same concepts in both entity names and definitions. For example, the term "*Room*" used in CityGML and the term "*Space"* used in IFC have similar meanings and are used separately to represent rooms that contain a volume and bounded by some surfaces. However, the difference in terminology between IFC and CityGML causes a mismatch of the two terms in the linguistic-based method, leading to a rank of 879 between the CityGML entity "*RoomType*" and the IFC entity "*IfcSpace"*. The problem may be solved if thesauruses of words in the entity names and/or definitions are considered in the linguistic-based comparisons. However, this may dramatically increase the running time and may generate many false mappings.

5. CONCLUSION AND FUTURE WORK

This paper presents a semi-automatic linguistic-based approach to map the IFC schema and the CityGML schema, which are related in scope but from different domains. Conventional linguistic-based approaches only consider the similarity of entity names and often are conducted manually. In this approach, entity names and definitions are used to evaluate the relatedness between IFC and CityGML entities. Linguistic and text mining techniques such as stemming and tf-idf are leveraged in the approach in a semi-automatic manner. Cosine Similarity, Jaccard Similarity Coefficient and Market Basket Model are used to calculate the similarity score of each entity pair. The linguistic-based method was evaluated using the results from the instance-based mapping. The results show that the top 200 mapping candidates from the linguistic-based method could achieve a 53% discovery of true matches. In addition, the precision results indicate that an accuracy of 100% and 87.5% on average can be achieved for the top 10 results and the top 25 results, respectively. This show that the linguistic-based method proposed in this paper can help narrow down the search space and provide a semi-automatic way when mapping heterogeneous data schemas, even though they are from different domains.
The proposed framework could result in tools that could facilitate mapping discovery of entities in the mapping process of CityGML and IFC. By introducing definition comparison in the mapping process, our framework could provide suggestions that are not possible in a traditional name-to-name comparison, such as mapping “AbstractOpening” in CityGML and “IfcWindow” or “IfcDoor” in IFC. In this paper, we did not provide details of geometry transformation, such as coordinate system transformation or CSG/Swept Solid to BRep transformation since it is not the focus of this study. However, readers could refer to (Cheng et al., 2013) for more details for geometry transformation.

The benefits and limitations of the linguistic-based method proposed in this paper have been discussed in Section 4.6. Since the proposed linguistic-based method can be used to map data schemas other than IFC and CityGML, the method will be tested on more schema mapping cases in the future for validation and evaluation purposes. Special attention will be given to the schemas in the same domain, such as CityGML and COLLADA, which are both representative schemas in the GIS domain. The future work will also include the improvement of the similarity analysis calculation and the consideration of thesaurus, as suggested in Section 4.6.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the support by the Hong Kong Research Grants Council, Grant No. 622812. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the Hong Kong Research Grants Council.

7. REFERENCES

Apache (2012). "Apache Lucene Core." Retrieved December 31, 2014, from http://lucene.apache.org/core/.

Bansal, V. K. (2007). Potential of gis to find solutions to space related problems in construction industry. *In:* Ardil, C. (ed.) *Proceedings of world academy of science, engineering and technology, vol 26, parts 1 and 2, december 2007.*

Benner, J., Geiger, A.and Leinemann, K. Flexible generation of semantic 3d building models. Gröger/Kolbe (Eds.), *Proceedings of the 1st International Workshop on Next Generation 3D City Models*, 2005 Bonn. EuroSDR Publication.

buildingSMART International (2007). "IFC2x Edition 3 Technical Corrigendum 1." Retrieved 4.20, 2013, from http://www.buildingsmart-tech.org/ifc/IFC2x3/TC1/html/index.htm.

Cheng, C. P., Lau, G. T., Pan, J., Law, K. H.and Jones, A. Domain-specific ontology mapping by corpus-based semantic similarity. *Proceedings of 2008 NSF CMMI Engineering Research and Innovation Conference*, 2008 Knoxville, Tennessee.

Cheng, J. C. P., Deng, Y. and Du, Q. (2013). Mapping Between BIM Models and 3D GIS City Models of Different Levels of Detail, *13th International Conference on Construction Applications of Virtual Reality*. London, United Kingdom.

Cheng, M. Y. and Chen, J. C. (2002). Integrating barcode and GIS for monitoring construction progress. *Automation in Construction,* 11**,** 23-33.

Eastman, C., Teicholz, P., Sacks, R.and Liston, K. (2008). *BIM Handbook: A Guide to Building Information Modeling for*

Owners, Managers, Designers, Engineers and Contractors, Wiley Publishing.

Garrett, J., Akinci, B.and Wang, H. (2004). Towards domain-oriented semi-automated model matching for supporting data exchange. *International Conference on Computing in Civil and Building Engineering (ICCCBE).* Weimar , Bauhaus-Universität.

Gröger, G., Kolbe, T. H., Nagel, C.and Häfele, K.-H. (2012). *OpenGIS City Geography Markup Language Encoding Standard 2.0,* Open Geospatial Consortium Inc. .

Hassanain, M., Froese, T.andVanier, D. (2001). Development of a maintenance management model based on iai standards. *Artificial Intelligence in Engineering,* 15**,** 177-193.

Hunter, J. and R. Lear (2012). "JDOM 1.1.2." Retrieved December 31, 2014, from http://www.jdom.org/dist/binary/archive/jdom-1.1.2.zip.

Isikdag, U.and Zlatanova, S. (2009). Towards defining a framework for automatic generation of buildings in citygml using building information models. *In:* Lee, J. & Zlatanova, S. (eds.) *3D Geo-information Sciences.* Berlin Heidelberg: Springer

Lawrence, M., Pottinger, R. and Staub-French, S. (2010) Coordination of data in heterogenous domains. *The 2010 IEEE 26th International Conference*, Long Beach, CA, USA. 167-170.

Lawrence, M., Pottinger, R., Staub-French, S. and Nepal, M. P. (2014). Creating flexible mappings between Building Information Models and cost information. *Automation in Construction*, 45, 107-118.

Lipman, R. (2009). Details of the mapping between the cis/2 and ifc product data models for structural steel. *Journal of Information Technology in Construction,* Vol. 14**,** 1-13.

Ma, Z., Wei, Z., Song, W.and Lou, Z. (2011). Application and extension of the ifc standard in construction cost estimating for tendering in china. *Automation in Construction,* 20**,** 196-204.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM,* 38**,** 39-41.

Nagel, C., Stadler, A.and Kolbe, T. H. Conceptual requirements for the automatic reconstruction of building information models from uninterpreted 3d models. *Academic Track of Geoweb 2009 Conference*, Vancouver, 2009.

Pan, J., Cheng, J. C. P., Lau, G. T.and Law, K. H. (2008). Utilizing statistical semantic similarity techniques for ontology mapping--with applications to aec standard models. *Tsinghua Science & Technology,* 13**,** 217-222.

Pauwels, P., Van Deursen, D., Verstraeten, R., De Roo, J., De Meyer, R., Van De Walle, R.and Van Campenhout, J. (2011). A semantic rule checking environment for building performance checking. *Automation in Construction,* 20**,** 506-518.

Rahm, E.and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *the VLDB Journal,* 10**,** 334-350.

Robinson, G. R.and Kapo, K. E. (2004). A gis analysis of suitability for construction aggregate recycling sites using regional transportation network and population density features. *Resources Conservation and Recycling,* 42**,** 351-365.

Strzalka, A., Bogdahn, J., Coors, V.and Eicker, U. (2011). 3d city modeling for urban scale heating energy demand forecasting. *HVAC&R Research***,** 526-539.

Su, X., Andoh, A. R., Cai, H., Pan, J., Kandil, A.and Said, H. M. (2012). GIS-based dynamic construction site material layout evaluation for building renovation projects. *Automation in Construction,* 27**,** 40-49.

Wang, H., Akinci, B.and Garrett Jr, J. H. (2007). Formalism for detecting version differences in data models. *Journal of Computing in Civil Engineering,* 21**,** 321-328.

Wang, H., Akinci, B., Garrett Jr, J. H.and Reed, K. A. (2008). Formalism for applying domain constraints in domain-oriented schema matching. *Journal of Computing in Civil Engineering,* 22**,** 170-180.

Willett, P. (2006). The porter stemming algorithm: Then and now. *Program: Electronic Library and Information Systems,* 40**,** 219-223.

Wu, I. C.and Hsieh, S. H. (2007). Transformation from ifc data model to gml data model: Methodology and tool development. *Journal of the Chinese Institute of Engineers,* 30**,** 1085-1090.