TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/17 pp
Volume , Number , 2008

Utilizing Statistical Semantic Similarity Techniques for Ontology Mapping – with Applications to AEC Standard Models

Jiayi Pan[§], Chin-Pang Jack Cheng[*†], Gloria T. Lau[†], Kincho H. Law[†]

§ School of Civil Engineering, Tsinghua University, Beijing 100084, China;
† Engineering Informatics Group, Stanford University, CA 94305, USA

Abstract: In the Architecture, Engineering and Construction (AEC) industry, there exist a number of ontological standards to describe the semantics of building models. Although the standards share similar scopes of interest, the task of comparing and mapping concepts among standards is challenging due to their differences in terminologies and perspectives. Ontology mapping is therefore necessary to achieve information interoperability, which allows two or more information sources to exchange data and to re-use the data for further purposes. This paper introduces three semi-automated ontology mapping approaches. These approaches adopt the relatedness analysis techniques to discover related concepts from heterogeneous ontologies. A pilot study is conducted on *IFC* and *CIS/2* ontologies to evaluate the approaches. Preliminary results from an attribute-based approach, a corpus-based approach and a name-based approach are presented to illustrate the utilization of the relatedness analysis approach for ontology mapping..

Key words: ontology mapping; similarity analysis; information interoperation; statistical analysis techniques

Introduction

To facilitate information flows between individuals in a supply chain, the interoperability issue among information sources is inevitable. Recent studies performed by the US National Institute of Standards and Technology (NIST) have reported significant costs to the construction industry due to inefficient interoperability [3]. In the context of a supply chain, the purpose of interoperation is to allow two or more information sources to exchange data and to re-use the data for further purposes. Effective interoperation therefore adds values to individual information sources and enhances efficiency and productivity in a supply chain.

To capture various phases and facets of design and construction processes, some organizations have been collaborating to build a single, unifying standard model of concepts and definitions. The earlier development of the IFC [5] and the current use of Building Information Model (BIM) have been focused on establishing unifying models to describe product, process and organization information in aspects such as design, analysis, procurement and installation (even though individual applications would likely use certain aspect and only portion of the model). As pointed out by researchers at NIST, unifying models tend to be inefficient and impractical [11]. Contrary to a unifying model, separate yet linked models differentiated by types and scopes are easier to manage and more flexible for information exchange among multi-disciplinary applications.

Short of unifying the different semantic models in the AEC industry, a mapping is needed to build the

Received:
* To whom correspondence should be addressed.
E-mail: cpcheng@stanford.edu; Tel: 1-650-8623262

linkages among the models. Examples of the ontological standards in the building industry include the Industry Foundation Classes (IFC) [5], the CIMsteel Integration Standards (CIS/2) [2], and the OmniClass Construction Classification System [1]. Each of these standards is constructed for specific purposes and from specific viewpoints. Although the standards share similar scopes of interest, it is a non-trivial task to develop a mapping due to their differences in terminologies and perspectives. Currently, ontology mapping among the various models is conducted in a labor intensive manual manner [7, 14]. This manual ontology mapping process is quite daunting as the number of heterogeneous sources to be accessed, compared and related increases. In this paper, we present a semi-automated ontology mapping approach using text mining and statistical analysis techniques. We believe the input from domain experts is necessary to achieve a usable mapping, but a semi-automated filtering system could help minimize the manual effort.

1 Feature Selection

Ontologies are composed of concepts. To map concepts between heterogeneous ontologies, we consider the shared features between concepts to compute the relevancy between concepts. The features are extracted using advanced text mining techniques and then compared using statistical analysis methods. In this paper, the three features we consider are occurrence frequency in document corpus, attributes in data models and keywords in concept names. In Section 2, we will introduce the statistical analysis measures to evaluate the similarity between features.

1.1 Corpus-Based Features

A corpus of text documents is a good indicator of similarities between concepts. Well-structured documents are generally divided into sections and sub-sections, each of which contains contents about a specific topic. Concepts and phrases that appear in the same sections are often related, as demonstrated in Figure 1. Therefore, the relatedness of concepts can be discovered based on the co-occurrence frequency of the concepts in a document. In this approach, concepts are predicted as related if they frequently appear in the same sections in a document corpus.

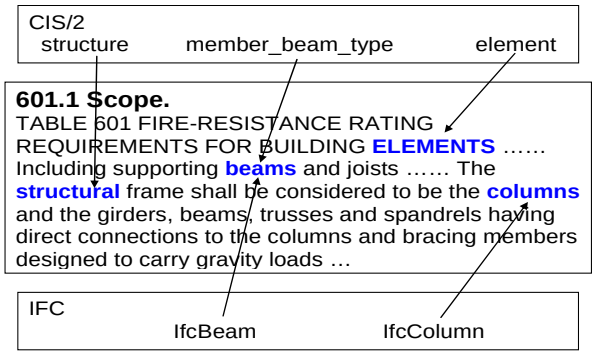

Fig. 1 Similarity analysis between concepts from heterogeneous ontologies using document corpus

A document corpus is used to relate concepts by computing their co-occurrence frequencies. This corpus must be carefully selected as it represents the relevancy among concepts from different ontologies. For this task, our corpus contains regulatory documents from the AEC domain due to their well-defined contents and well-organized structures. Regulations are rigorously drafted and reviewed, thus minimizing random co-occurrences of phrases in the same provision. As opposed to a general-purpose document corpus, we use a regulatory corpus which likely includes terms related to domain concepts.

1.2 Attribute-Based Features

The terminologies and structures used in data models with heterogeneous ontological standards may be very different, even though when they refer to the same entity. However, the sets of attribute values that are used to describe the same entity often do not differ significantly among data models. Figures 2 and 3 are examples of the *IFC2X3* and *CIS/2* standards that describe the identical structural column component with element name “C1”. As illustrated in Figures 2 and 3, the two representations are quite different. For example, IFC uses “Local Placement” to describe the column geometry and “Shape Representation” to describe the definition of the shape whereas CIS/2 uses “Element curve complex” and “Section profile I type” respectively. Owner history information is included in the definition of a column in IFC but not in CIS/2, while material information is included in CIS/2 but not in IFC. The difference in coverage can be explained by the fact that IFC is mainly used by CAD vendors and focuses on information about design description of building components; CIS/2 is emphasizes on informa-

```
#35008=IFCCOLUMN('25lBgcnZRvHB9_Bu7GGEeh',#1400005,'
C1','W14X193','Column',#35009,#35013,'C1');
   #1400005=IFCOWNERHISTORY(#1400003,#1400004,$,.NOCH
ANGE.,$,$,$,1201301536);
   #35009=IFCLOCALPLACEMENT($,#35010);
      #35010=IFCAXIS2PLACEMENT3D(#79,#35011,#35012);
         #79=IFCCARTESIANPOINT((0.0000,0.0000,1296.0000));
         #35011=IFCDIRECTION((1.,0.,0.));
         #35012=IFCDIRECTION((0.,0.,1.));
   #35013=IFCPRODUCTDEFINITIONSHAPE($,$,(#35014));
   #35014=IFCSHAPEREPRESENTATION(#1400011,'Body','Mapp
edRepresentation',(#35015));
      #1400011=IFCGEOMETRICREPRESENTATIONCONTEXT(
$,'Model',3,1.0E-5,#1400040,$);
      #35015=IFCMAPPEDITEM(#35020,#1400059);
```

Fig. 2 Excerpt from IFC that describes column "C1"

```
#8836=(ELEMENT('C1',',#8793,1)   ELEMENT_CURVE($)
   ELEMENT_CURVE_COMPLEX((#8830,#9866),(#828,#829),
   (#830,#831))
   ELEMENT_WITH_MATERIAL(#8795)
);
   #8793=ANALYSIS_MODEL_3D(",',.SPACE_FRAME.,$,3);
   #8830=SECTION_PROFILE_I_TYPE(6,'W14X193',$,$,10,.F.,#8
      831,#8832,#8833,#8834,#8835,$,$,$);
   #9866=SECTION_PROFILE_I_TYPE(7,'W14X193',$,$,10,.F.,#9
      869,#9870,#9871,#9872,#9873,$,$,$);
   #828=POINT_ON_CURVE('1',#827,PARAMETER_VALUE(0.00
      00000));
      #827=LINE('C1_LINE',#79,#826);
   #829=POINT_ON_CURVE('1',#827,PARAMETER_VALUE(144.
      0000000));
   #830=DIRECTION('C1_Z_AXIS_I',(1.0000,0.0000,0.0000));
   #831=DIRECTION('C1_Z_AXIS_J',(0.0000,0.0000,0.0000));
   #8795=MATERIAL_ISOTROPIC(1,'STEEL',$,#8798);
      #8798=MATERIAL_REPRESENTATION('STEEL',(#8801,#88
         02,#8803,#8804),#8805);
         #8801=MATERIAL_ELASTICITY('STEEL',0.3000000,290
            00.0000000,$,$);
```

Fig. 3 Excerpt from CIS/2 that describes column "C1"

tion about steel building and fabrication. Some fundamental attributes such as element name and geometry, however, are identical between the two data models. Therefore, an attribute value is a good feature that can potentially uncover similarities between concepts from heterogeneous ontologies.

To compare two data models based on attributes, the data models are parsed and transformed into a tree structure as shown in Figures 2 and 3. Every tree branch in the first data model is compared with every branch in the second data model tree. The relevancy between two branches is measured based on the similarity between their sets of attribute values, which is computed by the statistical similarity analysis measures described in Section 2. The set of attributes of a branch is defined as a set of the attributes of the branch data element as well as the attributes of all the descendant elements in the tree structure.

1.3 Name-Based Features

Some concept mappings cannot be discovered by the corpus-based approach or the attribute-based approach if the concepts do not appear in the document corpus or the data models we select. Some mappings such as (boundary_condition_logical, IfcBoundaryNodeCondition), nevertheless, are quite obvious from the name of the concept. Although the two concept names are not textually identical, they share a few terms such as "boundary" and "condition". The descriptive keywords in the concept name provide an alternative means to map concepts from different ontologies. To relate descriptive phrases, we tokenize keywords in concept names and compare the stemmed keywords. Stemming is done using Porter Stemmer [10].

2 Statistical Relatedness Analysis Measures

Consider an ontological standard of *m* concept terms and a set of *n* features. A feature vector $\vec{c}_i$ is an n-by-1 vector storing the occurrence frequencies of concept *i* among the *n* features. That is, for corpus-based features, the *k*-th element of $\vec{c}_i$ is defined as the number of times concept *i* is matched in section *k*; for attribute-based features, the *k*-th element of $\vec{c}_i$ represents the number of times attribute *k* is included in the branch with concept term *i*.

2.1 Cosine Similarity Measure

Cosine similarity is a non-Euclidean distance measure between two vectors. It is a common approach to compare documents in the field of text mining [9, 13]. Given two feature vectors $\vec{c}_i$ and $\vec{c}_j$, the similarity score between concepts *i* and *j* is represented using the dot product:

$$Sim(i, j) = \frac{\vec{c}_i \cdot \vec{c}_j}{\left|\vec{c}_i\right| \times \left|\vec{c}_j\right|} \quad (1)$$

The resulting score is in the range of [0, 1] with 1 as the highest relatedness between concepts *i* and *j*.

2.2 Jaccard Similarity Coefficient

Jaccard similarity coefficient [9, 12] is a statistical measure of the extent of overlap between two vectors. It is

defined as the size of the intersection divided by the size of the union of the vector dimension sets. It is a popular similarity analysis measure of term-term similarity due to its simplicity and retrieval effectiveness [6]. For corpus-based features, two concepts are considered similar if there is a high probability for both concepts to appear in the same sections; for attribute-based features, two concepts are considered similar if there is a high extent of overlap between the two sets of attribute values. To illustrate the application to our corpus-based approach, let N_{11} be the number of sections both concept *i* and *j* are matched to, N_{10} be the number of sections concept *i* is matched to but not concept *j*, N_{01} be the number of sections concept *j* is matched to but not concept *i*, and N_{00} be the number of sections that both concept *i* and *j* are not matched to. The similarity between both concepts is then computed as

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

Since the size of intersection cannot be larger than the size of union, the resulting similarity score is between 0 and 1.

2.3 Market Basket Model

Market-basket model is a probabilistic data-mining technique to find item-item correlation [4]. The task is to find the items that frequent the same baskets. Market-basket analysis is primarily used to uncover association rules between item and itemsets. The *confidence* of an association rule $\{i_1, i_2, ..., i_k\} \rightarrow j$ is defined as the conditional probability of *j* given itemset $\{i_i, i_2, ..., i_k\}$. The *interest* of an association rule is defined as the absolute value of the difference between the confidence of the rule and the probability of item *j*. To compute the similarities among concepts, our goal is to find concepts *i* and *j* where either association rule $i \rightarrow j$ or $j \rightarrow i$ is high-interest.

Consider a set of *n* features. Let N_{11} be the number of features both concept *i* and *j* possess, N_{10} be the number of features concept *i* possesses but not concept *j*, and N_{01} be the number of features concept *j* possesses but not concept *i*. The forward similarity of the concepts *i* and *j*, which is the interest of the association rule $i \rightarrow j$ without absolute notation, is expressed as

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

The value ranges from -1 to 1. The value of -1 means that concept *j* possess all the *n* features while concept *i* does not possess any of the features. The value of 1 is unattainable because (N_{11} + N_{01}) cannot be zero while confidence equals one. Conceptually, it represents the boundary case where the number of features that concept *j* possesses is not significant, but concept *j* possesses every feature that concept *i* possesses.

3 Preliminary Results

For illustrative purpose in the AEC domain, entities in the CIS/2 [2] and IFC [5] ontological standards are selected as concepts. A mapping is produced between the CIS/2 and IFC using the similarity analysis methods described above. For the corpus-based features, chapters from the International Building Code (IBC) are used as the document corpus to uncover the concept relevancy between CIS/2 and IFC. The IBC addresses the design and installation of building systems and is therefore leveraged for mappings in the AEC domain.

The mapping results from our system are evaluated against the manual mappings in [8]. The manual mappings include 103 CIS/2 concepts and 85 IFC concepts, which are regarded as the true matches. In computing the relevancy between a concept from the CIS/2 and one from the IFC, our system produces a pairwise similarity score between two concepts. With different similarity score thresholds, values of precision and recall are graphed for the three similarity analysis approaches and the three statistical measures.

3.1 Precision and Recall

Precision and recall are commonly used as the metrics to evaluate the accuracy of predictions and the coverage of accurate pairs of an information retrieval system or approach. Precision measures the fraction of predicted matches that are correct, i.e., the number of true positives over the number of pairs predicted as matched. Recall measures the fraction of correct matches that are predicted, i.e., the number of true positives over the number of pairs that are actually matched. They are computed as

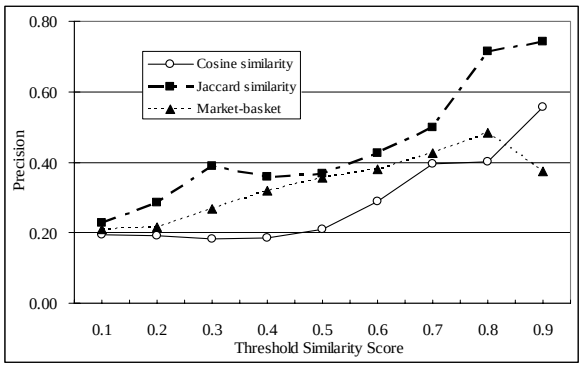

Fig. 4 Evaluation results of the three measures using precision

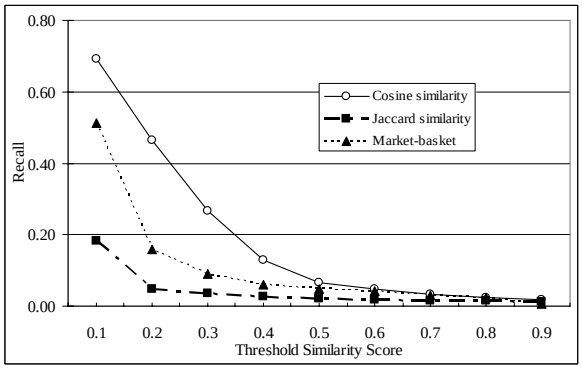

Fig. 5 Evaluation results of the three measures using recall rate

$$Precision = \frac{|True\ Matches \cap Predicted\ Matches|}{|Predicted\ Matches|} \quad (4)$$

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

3.2 Comparison of the Three Measures

Figures 4 and 5 show the results of the three statistical relatedness analysis measures using precision and recall rate. Jaccard similarity shows the highest precision yet the lowest recall rate among the three measures. The low recall rate is due to the fact that the number of predicted matches using Jaccard similarity is much smaller than the other two measures. However, the matches predicted by Jaccard similarity are more likely correct. Contrary to Jaccard similarity, cosine similarity shows the lowest precision yet the highest recall rate due to its largest amount of predicted matches for all thresholds. Market basket model appears to be the average among the three statistical measures.

3.3 Comparison of the Three Features

Figure 6 shows the precision of the three features. The results are computed using Jaccard similarity measure as it produces the highest precision for all features. As a baseline comparison, random permutation would result in a precision of about 0.2. This is due to the fact that about 20% of the possible pairwise matches are true matches according to the manual mapping in [8]. In Figure 6, we show that all three approaches result in a general precision of about 0.45. As expected, the highest scoring matches produce the highest precision, which shows that the similarity score is a good measure of the degree of relevancy between concepts. In this study, attribute-based approach outperforms the other two approaches because the CIS/2 and the IFC models are heavily populated with attributes to describe concept properties. Despite the fact that the IBC is well-structured and is in the general domain of the IFC and CIS/2, the concept appearance in the IBC is low. Due to the low coverage, the corpus-based approach results in the lowest precision.

There is always a tradeoff between precision and recall, as illustrated in Figures 4 and 5. F-measure is therefore leveraged to combine both metrics. It is a weighed harmonic mean of precision and recall. In other words, it is the weighed reciprocal of the arithmetic mean of the reciprocals of precision and recall. It is computed as

$$F-measure = \frac{2 \cdot (Predicted \times Recall)}{Predicted + Recall} \quad (6)$$

Figure 7 shows the results for the three features using the F-measure. Again, the attribute-based approach outperforms the other two approaches for most thresholds. The F-measure results in Figure 7 illustrate that the attribute-based features attain optimal performance without scarifying recall to obtain its relatively high precision in Figure 6.

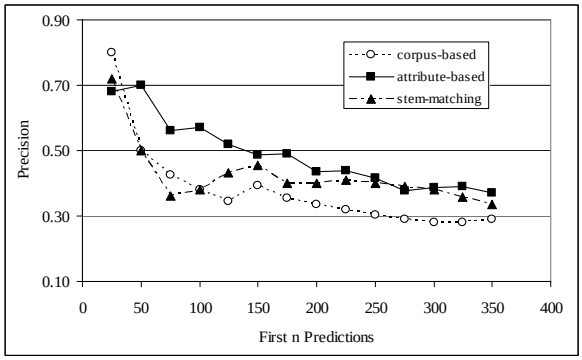

Fig. 6 Evaluation results of the three features using precision

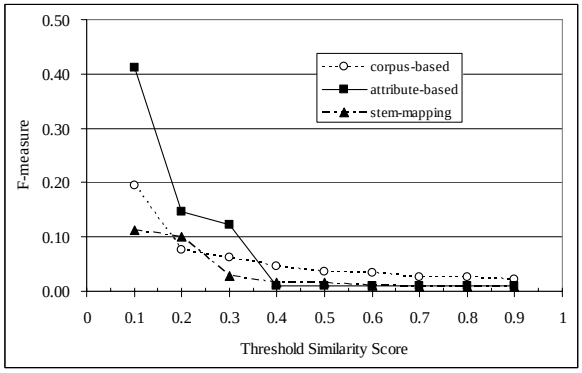

Fig. 7 Evaluation results of the three approaches using F-measure

4 Conclusions and Future Tasks

This paper presents three semi-automated approaches that utilize statistical similarity analysis techniques to relate concepts from heterogeneous ontologies. Different features are extracted as the analysis dimensions for the three approaches. Three statistical analysis measures, namely cosine similarity, Jaccard similarity and market basket model, are evaluated against each approach.

The approaches were tested and evaluated through an illustrative example of mapping CIS/2 and IFC, two commonly used ontologies in the AEC industry. Among the three statistical measures, Jaccard similarity shows the highest precision whereas cosine similarity shows the highest recall. In this study with the selected ontology concepts, features and corpus, among the three similarity features, the attribute-based approach outperforms the other two in terms of precision and the F-measure, which is a combination of precision and recall. The name-based approach helps to identify concepts that directly share the same root terms, while the other two approaches can indirectly discover related concepts through ontological structure and supporting documents. To this end, we plan to develop an improved system by combining the three approaches in the future.

5 Acknowledgements

The authors would like to acknowledge the supports by the US National Science Foundation, Grant No. CMS-0601167, the Center for Integrated Facility Engineering (CIFE) at Stanford University and the Enterprise Systems Group at the National Institute of Standards and Technology (NIST). Any opinions and findings are those of the authors, and do not necessarily reflect the views of NSF, CIFE or NIST. No approval or endorsement of any commercial product by NIST, NSF or Stanford University is intended or implied.

References

[1] Construction Specifications Institute. OmniClass Construction Classification System, Edition 1.0. 2006.

[2] Crowley A, Watson A. CIMsteel Integration Standards Release 2, SCI-P-268, the Steel Construction Institute, Berkshire, England, 2000.

[3] Gallaher M P, O'Connor A C, Dettbarn J L, et al. Cost Analysis of Inadequate Inoperability in the Capital Facilities Industry. *National Institute of Standards and Technology (NIST) Technical Report No. GCR-04-867*, 2004.

[4] Hastie T, Tibshirani R, Friedman J H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, NY, 2001.

[5] International Alliance for Interoperability (IAI). Guidelines for the Development of Industry Foundation Classes (IFC), IAI, 1997.

[6] Kim M-C, Choi K-S. A Comparison of Collocation-based Similarity Measures in Query Expansion. *Information Processing and Management: an International Journal*, 1999, **35**(1), 19-30.

[7] Lipman R. Mapping between the CIMsteel Integration Standards (CIS/2) and Industry Foundation Classes (IFC) Product Models for Structural Steel. In: Proceedings of ICCCBE. Montreal, Canada, 2006.

[8] Lipman R. Mapping between the CIS/2 and IFC Product Data Models for Structural Steel. Technical Report, NISTIR 7453, NIST, 2007.

[9] Nahm U Y, Bilenko M, Mooney R J. Two Approaches to Handling Noisy Variation in Text Mining. In: Proceedings of the ICML-2002 Workshop on Text Learning. Sydney, Australia, 2002.

[10] Porter M F. An Algorithm for Suffix Stripping. *Program*, 1980, **14**(3), 130-137.

[11] Ray S R. Interoperability Standards in the Semantic Web. *Journal of Computing and Information Science in Engineering*, 2002, **2**(1), 65-69.

[12] Roussinov D, Zhao J L. Automatic Discovery of Similarity Relationships Through Web Mining. *Decision Support Systems*, 2003, **25**, 149-166.

[13] Salton G. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Publishing Co., Inc. Boston, MA, USA, 1989.

[14] Teague T L, Palmer M E, Jackson R H F. XML for Capital Facilities. *Leadership and Management in Engineering*, 2003, **3**(2), 82-85.