Domain-Specific Ontology Mapping by Corpus-Based Semantic Similarity

Chin Pang Cheng[1], Gloria T. Lau[1], Jiayi Pan[1], Kincho H. Law[1] and Albert Jones[2]

[1]Engineering Informatics Group
Department of Civil & Environmental Engineering
Stanford University
Stanford, CA 94305, U.S.A

[2]Enterprise Systems Group
National Institute of Standards and Technology
Gaithersburg, MD 20899-0001

Abstract

Mapping heterogeneous ontologies is usually performed manually by domain experts, or accomplished by computer programs via comparing the structures of the ontologies and the linguistic semantics of their concepts. In this work, we take a different approach to compare and map the concepts of heterogeneous domain-specific ontologies by using a document corpus in a domain similar to the domain of the ontologies as a bridge. Cosine similarity and Jaccard coefficient, two vector-based similarity measures commonly used in the field of information retrieval are adopted to compare semantic similarity between ontologies. Additionally, the market basket model is modified as a relatedness analysis measure for ontology mapping. We use regulations as the bridging document corpus and the consideration of the corpus hierarchical information in concept similarity comparison. Preliminary results are obtained using ontologies from the architectural, engineering and construction (AEC) industry. The proposed market basket model appears to outperform the other two similarity measures, with its prediction error reduced using corpus structural information.

Keywords

Heterogeneous Taxonomies, Ontology Mapping, Relatedness Analysis.

1. Introduction

The purpose of interoperation is to increase the value of information when information from multiple, heterogeneous sources is accessed, related and combined. Recent studies by the National Institute of Standards and Technology (NIST) have reported that inefficient interoperability led to significant costs to the construction as well as the automotive industries (Gallaher et al 2004, NIST 1999). One common approach to enhance communication among heterogeneous information sources is to develop interoperability or ontology standards. It has been forecasted by some that "By 2010, ontologies ….will be the basis for 80 percent of application integration projects" (Jacobs and Linden 2002). Ontologies serve as a means for information sharing and capture the semantics of a domain of interest. However, building a single, unifying model of concepts and definitions is neither efficient nor practical. Different groups or organizations operate in different contexts with different definitions. A more practical assumption is that software services that need to communicate will likely be based on distinct ontologies (Ray 2002). In practice, multiple terminology classifications or data model structures exist. For instance, in the architectural, engineering and construction (AEC) industry, there are a number of ontologies to describe the semantics of building models, such as the Industry Foundation Classes (IAI 1997), the CIMsteel Integration Standards (CIS/2) (Watson 1995), and the OmniClass Construction Classification System (CSI 2006). For model rebuilding and data exchange purposes, comparison and mapping between heterogeneous ontologies in the same industry are often inevitable.

The tasks of ontology comparison and mapping are commonly performed manually by domain experts, who are familiar with one or more industry-specific taxonomies. The manual task could be time-consuming, unscalable and inefficient. Surveys on the various approaches for ontology mapping (merging, alignment) have been reported (de Bruijn et al 2004, Euzenat et al 2004). Automated comparison and mapping based on the ontology structures and the linguistic similarity between concepts are growing in popularity in recent years. Some common approaches include term matching that relates terms with the same words, synonyms, or terms with the same root. Dictionaries are used (Ehrig and Sure 2004, van Hage, Katrenko and Schreiber 2005) to help define and compare ontology concepts. However, the reliability is not guaranteed because the set of synonyms and the definition paragraphs could be different from different sources. In addition, the use of stemmers such as Porter (1980) and Lovins (1968) to reduce derived words to their root, stem or base form (e.g. from *piling* to *pile*) is not always appropriate. For instance, suffixes like -*itis*, -*oma* and -*idae* may be specific to a particular domain and therefore cannot be considered by traditional stemmers (Grabar and Zweigenbaum 2000). In addition, many concepts have different meanings when used in different domains. For example, the concept "*finishes*" refers to the decorative texture or coating of a surface in the construction industry whereas it means to complete or to terminate in a general sense. Hence domain-specific methodologies may be more desirable. In this work, we focus on using ontologies from the construction industry and use building code regulations as the document corpus for concept similarity comparison.

With the intuition that related terms should appear in the same paragraphs or sections, concept comparison and matching by co-occurrence is proposed to map different sets of terms from heterogeneous ontologies. The co-occurrence frequency of two concepts in the corpus reveals the closeness of the two topics and acts as a means to compute the relatedness between them. The document corpus herein used should be in the same domain as the mapping ontologies in order to capture the domain-specific semantics of the concepts. Two existing relatedness analysis techniques, namely cosine similarity and Jaccard coefficient, and the suggested market basket model are proposed as similarity metrics for corpus-based ontology mapping.

2. Existing Relatedness Analysis Approaches

To find similar or related concepts in a different ontology, two pools of concepts are compared with each other to obtain the similarity score, which is a measure of relatedness of each pair of concepts. Two existing approaches, namely cosine similarity measure and Jaccard similarity coefficient, are herein introduced and then compared. Both metrics are non-Euclidean distance measures, which are based on properties of points instead of their locations in the domain space.

Consider two ontologies, $O1$ and $O2$, with m and n concepts respectively, and a corpus of N regulation sections. A frequency vector $\vec{c}_i$ is an N-by-1 vector storing the occurrence frequencies of concept i from either ontology $O1$ or $O2$ among the N documents. That is, the k-th element of $\vec{c}_i$ equals the number of times concept i is matched in section k. Therefore, the frequency matrix of ontology $O1$, denoted by C_1, is an N-by-m matrix in which the i-th column vector is $\vec{c}_i$ for $i \le m$. And the frequency matrix of ontology $O2$, denoted by C_2, is an N-by-n matrix in which the i-th column vector is $\vec{c}_i$ for $i \le n$.

2.1 Cosine similarity measure

Cosine similarity is a non-Euclidean distance measure of similarity between two vectors by finding the angle between them. This is a common approach to compare documents in text mining (Larsen and Aone 1999, Nahm, Bilenko and Mooney 2002, Salton 1989). Given two frequency vectors $\vec{c}_i$ and $\vec{c}_j$, the similarity score between concept i from ontology $O1$ and concept j from ontology $O2$ is represented using the dot product:

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

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

2.2 Jaccard similarity coefficient

Jaccard similarity coefficient (Nahm, Bilenko and Mooney 2002, Roussinov and Zhao 2003) is a statistical measure of the extent of overlapping between two vectors. It is defined as the size of the intersection divided by the size of the union of the vector dimension sets:

$$Jaccard(i, j) = \frac{\left|\vec{c}_i \cap \vec{c}_j\right|}{\left|\vec{c}_i \cup \vec{c}_j\right|}$$

Two concepts are considered similar if there is a high probability for both concepts to appear in the same sections.. To illustrate the application to the concept relatedness analysis, let N_{11} be the number of sections both concept *i* from *O*1 and concept *j* from *O*2 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 concepts *i* and *j* are not matched to. The similarity between both concepts is then computed as

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

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

3. Market Basket Model

Market-basket model is a probabilistic data-mining technique to find item-item correlation (Hastie, Tibshirani and Friedman 2001). The task is to find the items that frequent the same baskets. The *support* of each itemset *I* is defined as the number of baskets containing all items in *I*. Sets of items that appear in *s* or more baskets, where *s* is the support threshold, are the *frequent itemsets.*

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 $\{i_1, i_2, ..., i_k\}$. 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 corpus of *N* documents. Let N_{11} be the number of sections both concepts *i* and *j* are matched to, N_{10} be the number of sections concept *i* is matched to but not concept *j*, and N_{01} be the number of sections concept *j* is matched to but not concept *i*. The occurrence probability of concept *j* is computed as

$$\Pr(j) = \frac{N_{11} + N_{01}}{N}$$

and the confidence of the association rule $i \rightarrow j$ is

$$Conf(i \rightarrow j) = \Pr(j \mid i) = \frac{N_{11}}{N_{11} + N_{10}}$$

So, the forward similarity of the concepts *i* and *j*, that is the interest of the association rule $i \rightarrow j$ without absolute notation, is expressed as

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

The value ranges from -1 to 1. The value of -1 means that concept *j* is matched to all sections but concept *i* does not co-exist in any of these sections. The value of 1 is unattainable because (N_{11} + N_{01}) cannot be zero when confidence equals one. Conceptually, it represents the boundary case where the occurrence of concept *j* is not significant in the corpus, but it appears in every section that concept *i* appears.

4. Use of Corpus Hierarchy Structural Information

Although many related concepts can be captured by treating each section in a document corpus as an independent dimension in concept co-occurrence comparison, some related concepts rarely co-occur in the same sections. For example, if two concepts contain an *Is*-relationship, such as door furniture and door hardware, they may be used in the same corpus interchangeably but in different sections.

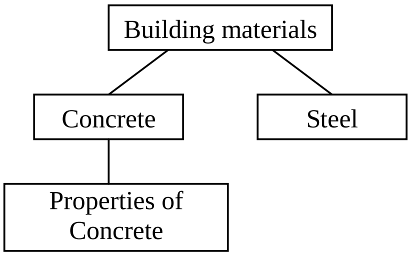


Figure 1: Example of related but rarely co-occurring concepts

Is-A-related concepts are also hard to be extracted if regarding each section as a discrete information island because the relationship between concepts such as building materials and concrete are sometimes implicit from the structures of sections. For example, the descriptions of building materials and those of concrete may not appear in the same sections; instead, the sections describing concrete are usually the subsections of the sections describing building materials. By considering sections with more levels up, the implicit relationship between building materials and concrete will become more obvious. Moreover some concepts, for instance concrete and steel, are related but may not appear in the same sections because they are in different sub-scopes of the same topic. Their computed relatedness by corpus-based similarity comparison can be increased if we can discover the fact that the sections about concrete and the sections about steel are at the same levels under the same parent section (Figure 1). As a result, the hierarchical structure of sections needs to be considered to extract the implied related concepts.

4.1 Regulations as document corpus

Regulations are used as the training document corpus because of their well-defined contents and well-organized hierarchical structures within regulations. Regulations are usually voluminous and cover a broad range of scopes. Nevertheless they are organized into many sections and sub-sections, each of which contains contents with a specific topic or scope. In addition, the fact that regulations are written with precise and concise contents helps to reduce the possibility of false negatives, i.e., the mismatched concepts.

The tree hierarchy of regulations provides additional information besides the coexistence of concept terms. Lau et al. (2005, 2006) compare sections in different regulations with the help of the hierarchy structural information of each regulation in order to locate sections in similar and to build an e-government system. The results illustrated in Lau et al. (2005, 2006) show that structural organization is resourceful information if regulations are employed to uncover semantic relationships between concepts from different ontologies.

Well-structured regulations could be simplified as a hierarchical tree, in which each section corresponds to a discrete node. Each section has a parent section, a set of sibling sections and a set of child sections (Figure 2). For a section with a particular topic, the parent section describes a broader topic, the sibling sections describe similar topics and the child sections describe more specific topics in general. The parent section, sibling sections and child sections can be taken into consideration by assuming that all the concepts matched to those sections appear also in the self section, discounted by a factor.

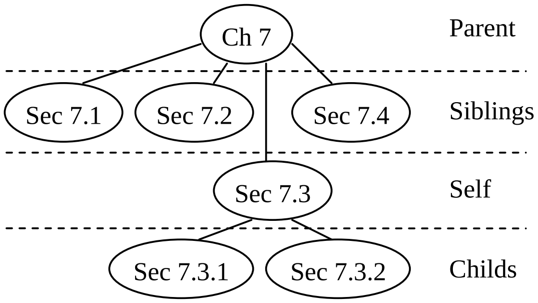


Figure 2: Hierarchically-related sections

5. Practical Demonstration

For demonstrative purpose in the construction domain, entities in the OmniClass (CSI 2006) and IfcXML (IAI 1997) classification models were selected as concepts and documents from the International Building Code (IBC) (ICBO, 2000) were used as the corpus for concept relatedness analysis (Figure 3).

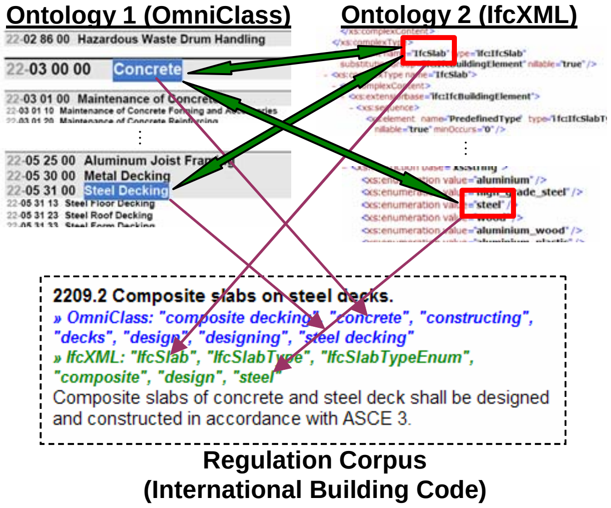

Ontology 1 (OmniClass)
Ontology 2 (IfcXML)
22-02 86 00 Hazardous Waste Drum Handling
22-03 00 00 Concrete
22-03 01 00 Maintenance of Concrete
22-03 01 10 Maintenance of Concrete Forming and Accessories
22-03 01 20 Maintenance of Concrete Reinforcing
22-05 25 00 Aluminum Joist Framing
22-05 30 00 Metal Decking
22-05 31 00 Steel Decking
22-05 31 13 Steel Floor Decking
22-05 31 23 Steel Roof Decking
22-05 31 33 Steel Form Decking
"IfcSlab"
"ifc:IfcSlab"
"true"
"IfcSlab"
"ifc:IfcBuildingElement"
PredefinedType
"true" minOccurs="0"
value="aluminium"
"steel"
value="aluminium_wood"
2209.2 Composite slabs on steel decks.
» OmniClass: "composite decking", "concrete", "constructing", "decks", "design", "designing", "steel decking"
» IfcXML: "IfcSlab", "IfcSlabType", "IfcSlabTypeEnum", "composite", "design", "steel"
Composite slabs of concrete and steel deck shall be designed and constructed in accordance with ASCE 3.
Regulation Corpus
(International Building Code)

Figure 3: Mapping between OmniClass and IfcXML using IBC as the corpus

5.1 OmniClass, IfcXML and IBC

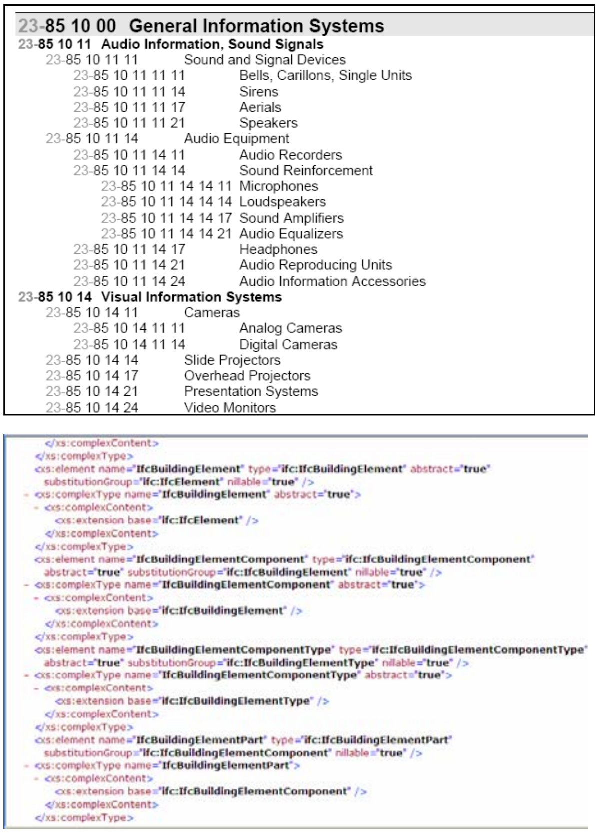
In the AEC industry nowadays, the urge for Building Information Model (BIM) leads to the establishment of various description and classification standards to facilitate data exchange. OmniClass and IfcXML by far are two of the most commonly used data models for buildings and constructions. OmniClass categorizes elements and concepts in the AEC industry and provides a rich pool of vocabularies practitioners can use in legal documents. It contains a set of object data elements that represent the parts of buildings or processes, and the relevant information about those parts. OmniClass consists of 15 tables, each of which represents a different facet of construction information. IfcXML, specialized in modeling CAD models and work process, is frequently used by practitioners to build information-rich product and process models and to act as a data format for interoperability among different software. It is a single XML schema file comprised of concept terms which are highly hierarchically structured and cross-linked.

The International Building Code (IBC) addresses the design and installation of building systems through requirements that emphasize performance. Content is founded on broad-based principles that make possible the use of materials and building designs. Structural as well as fire- and life-safety provisions covering seismic, wind, accessibility, egress, occupancy, roofs, and more are included. The version herein used is the IBC published in 2006.

5.2 Implementation

Preprocessing of the ontologies is required at the beginning stage. OmniClass is organized in a hierarchical structure and each entity is associated with a unique ID (Figure 4a), which was discarded to obtain the textual OmniClass concepts. IfcXML is organized in a XML Schema XSD format (Figure 4b) in which element names, group names and type names were extracted as IfcXML concepts. The concepts were then sorted in alphabetical order and duplicates were eliminated.

The entire preprocessed concept terms of OmniClass and IfcXML were latched to each section of the IBC XML files. The concepts are inserted into the corresponding sections which match the concepts in the stemmed form (e.g. *fire system* instead of *fire systems*; *permit* instead of *permitted*). As an example, the XML structure of Section 907.2.11.3 Emergency Voice/Alarm Communication System is changed to include the OmniClass and IfcXML concepts as shown in Figure 5. While the stemmed form of the concept terms is used in latching, the "name" attribute for each **<OMNICLASS>** element and for each **<IFCXML>** element is in the original form that OmniClass uses. The "times" attribute is the number of times the term matches the contents in that section.

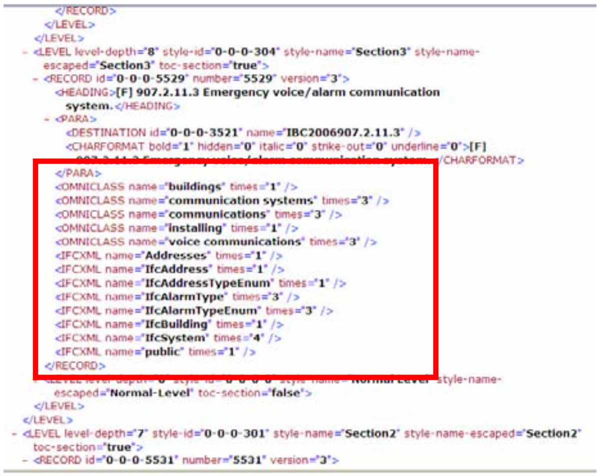


23-85 10 00	**General Information Systems**
23-85 10 11	**Audio Information, Sound Signals**
23-85 10 11 11	Sound and Signal Devices
23-85 10 11 11 11	Bells, Carillons, Single Units
23-85 10 11 11 14	Sirens
23-85 10 11 11 17	Aerials
23-85 10 11 11 21	Speakers
23-85 10 11 14	Audio Equipment
23-85 10 11 14 11	Audio Recorders
23-85 10 11 14 14	Sound Reinforcement
23-85 10 11 14 14 11	Microphones
23-85 10 11 14 14 14	Loudspeakers
23-85 10 11 14 14 17	Sound Amplifiers
23-85 10 11 14 14 21	Audio Equalizers
23-85 10 11 14 17	Headphones
23-85 10 11 14 21	Audio Reproducing Units
23-85 10 11 14 24	Audio Information Accessories
23-85 10 14	**Visual Information Systems**
23-85 10 14 11	Cameras
23-85 10 14 11 11	Analog Cameras
23-85 10 14 11 14	Digital Cameras
23-85 10 14 14	Slide Projectors
23-85 10 14 17	Overhead Projectors
23-85 10 14 21	Presentation Systems
23-85 10 14 24	Video Monitors

```
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcBuildingElement" type="ifc:IfcBuildingElement" abstract="true"
    substitutionGroup="ifc:IfcElement" nillable="true" />
- <xs:complexType name="IfcBuildingElement" abstract="true">
  - <xs:complexContent>
      <xs:extension base="ifc:IfcElement" />
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcBuildingElementComponent" type="ifc:IfcBuildingElementComponent"
    abstract="true" substitutionGroup="ifc:IfcBuildingElement" nillable="true" />
- <xs:complexType name="IfcBuildingElementComponent" abstract="true">
  - <xs:complexContent>
      <xs:extension base="ifc:IfcBuildingElement" />
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcBuildingElementComponentType" type="ifc:IfcBuildingElementComponentType"
    abstract="true" substitutionGroup="ifc:IfcBuildingElementType" nillable="true" />
- <xs:complexType name="IfcBuildingElementComponentType" abstract="true">
  - <xs:complexContent>
      <xs:extension base="ifc:IfcBuildingElementType" />
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcBuildingElementPart" type="ifc:IfcBuildingElementPart"
    substitutionGroup="ifc:IfcBuildingElementComponent" nillable="true" />
- <xs:complexType name="IfcBuildingElementPart">
  - <xs:complexContent>
      <xs:extension base="ifc:IfcBuildingElementComponent" />
    </xs:complexContent>
  </xs:complexType>
```

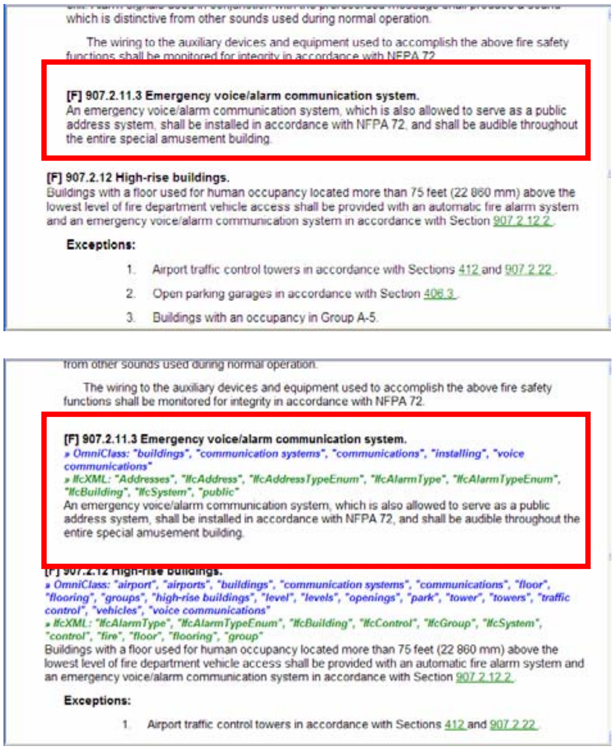
Figure 4: (Top) OmniClass table; (Bottom) IfcXML schema

With the help of XSL and CSS style sheets, the IBC can be viewed in a web browser in a reader-friendly way and the inserted **<OMNICLASS>** tags and **<IFCXML>** tags appear in blue and green respectively underneath the section heading of the corresponding matched section for reference. Figure 6 demonstrates the display of Section 907.2.11.3 Emergency Voice/Alarm Communication System before and after the OmniClass concepts were latched.



```
      </RECORD>
    </LEVEL>
  </LEVEL>
- <LEVEL level-depth="8" style-id="0-0-0-304" style-name="Section3" style-name-
    escaped="Section3" toc-section="true">
  - <RECORD id="0-0-0-5529" number="5529" version="3">
      <HEADING>[F] 907.2.11.3 Emergency voice/alarm communication
        system.</HEADING>
    - <PARA>
        <DESTINATION id="0-0-0-3521" name="IBC2006907.2.11.3" />
        <CHARFORMAT bold="1" hidden="0" italic="0" strike-out="0" underline="0">[F]
          907.2.11.3 Emergency voice/alarm communication system.</CHARFORMAT>
      </PARA>
      <OMNICLASS name="buildings" times="1" />
      <OMNICLASS name="communication systems" times="3" />
      <OMNICLASS name="communications" times="3" />
      <OMNICLASS name="installing" times="1" />
      <OMNICLASS name="voice communications" times="3" />
      <IFCXML name="Addresses" times="1" />
      <IFCXML name="IfcAddress" times="1" />
      <IFCXML name="IfcAddressTypeEnum" times="1" />
      <IFCXML name="IfcAlarmType" times="3" />
      <IFCXML name="IfcAlarmTypeEnum" times="3" />
      <IFCXML name="IfcBuilding" times="1" />
      <IFCXML name="IfcSystem" times="4" />
      <IFCXML name="public" times="1" />
    </RECORD>
  - <LEVEL level-depth="0" style-id="0-0-0-0" style-name="Normal-Level" style-name-
      escaped="Normal-Level" toc-section="false">
    </LEVEL>
  </LEVEL>
- <LEVEL level-depth="7" style-id="0-0-0-301" style-name="Section2" style-name-escaped="Section2"
    toc-section="true">
  - <RECORD id="0-0-0-5531" number="5531" version="3">
```

Figure 5: Latched OmniClass and IfcXML concepts in IBC



which is distinctive from other sounds used during normal operation.

The wiring to the auxiliary devices and equipment used to accomplish the above fire safety functions shall be monitored for integrity in accordance with NFPA 72.

[F] 907.2.11.3 Emergency voice/alarm communication system.
An emergency voice/alarm communication system, which is also allowed to serve as a public address system, shall be installed in accordance with NFPA 72, and shall be audible throughout the entire special amusement building.

[F] 907.2.12 High-rise buildings.
Buildings with a floor used for human occupancy located more than 75 feet (22 860 mm) above the lowest level of fire department vehicle access shall be provided with an automatic fire alarm system and an emergency voice/alarm communication system in accordance with Section 907.2.12.2.

Exceptions:

1. Airport traffic control towers in accordance with Sections 412 and 907.2.22.
2. Open parking garages in accordance with Section 406.3.
3. Buildings with an occupancy in Group A-5.

from other sounds used during normal operation.

The wiring to the auxiliary devices and equipment used to accomplish the above fire safety functions shall be monitored for integrity in accordance with NFPA 72.

[F] 907.2.11.3 Emergency voice/alarm communication system.
» OmniClass: "buildings", "communication systems", "communications", "installing", "voice communications"
» IfcXML: "Addresses", "IfcAddress", "IfcAddressTypeEnum", "IfcAlarmType", "IfcAlarmTypeEnum", "IfcBuilding", "IfcSystem", "public"
An emergency voice/alarm communication system, which is also allowed to serve as a public address system, shall be installed in accordance with NFPA 72, and shall be audible throughout the entire special amusement building.

[F] 907.2.12 High-rise buildings.
» OmniClass: "airport", "airports", "buildings", "communication systems", "communications", "floor", "flooring", "groups", "high-rise buildings", "level", "levels", "openings", "park", "tower", "towers", "traffic control", "vehicles", "voice communications"
» IfcXML: "IfcAlarmType", "IfcAlarmTypeEnum", "IfcBuilding", "IfcControl", "IfcGroup", "IfcSystem", "control", "fire", "floor", "flooring", "group"
Buildings with a floor used for human occupancy located more than 75 feet (22 860 mm) above the lowest level of fire department vehicle access shall be provided with an automatic fire alarm system and an emergency voice/alarm communication system in accordance with Section 907.2.12.2.

Exceptions:

1. Airport traffic control towers in accordance with Sections 412 and 907.2.22.

Figure 6: (Top) Original IBC; (Bottom) IBC with latched concepts

To limit the scope, Chapter 7 Fire-Resistance-Rated Construction and Chapter 9 Fire Protection Systems have been selected from the IBC, containing 839 sections in total. Both chapters are related to fire resistance and hence provide a combined corpus with shared terminology. 20 unique OmniClass concepts and 20 unique IfcXML concepts that are matched in these two chapters were randomly chosen for comparison. Similarity analysis was then performed between these 400 concept-concept pairs.

5.3 Result and discussion

Root mean square error is a metric to compute the difference between the predicted values and the true values so as to evaluate the accuracy of the prediction. Comparison between ontology of *m* concepts and ontology of *n* concepts involves *mn* concept-concept pairs. Therefore the RMSE is calculated as

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|true_{i,j} - predicted_{i,j}\right|}$$

Each concept-concept pair is given a true value of 1 if domain experts assure that the two concepts are similar or related, 0 otherwise. A predicted value of 1 is assigned to concept-concept pairs whose similarity score is larger than a similarity threshold *s* which could be adjusted according to similarity comparison approaches and different inclusion of hierarchy structural information. The comparison between ontology of *m* concepts and ontology of *n* concepts requires computation time of $O(mn)$.

Table 2 shows the result RMSE for each analysis metrics. Various weights of hierarchical information and threshold similarity scores have been used to illustrate their influences on the prediction errors. A threshold similarity score of 0.35 means that two concepts are considered related if the similarity score between them is greater than or equal to 0.35.

The market basket model results in the lowest RMSE in all cases of threshold similarity score and hierarchy information inclusion, illustrating that it is a better approach to find related concepts than cosine similarity measures and Jaccard similarity coefficient.

The result shows that corpus structural organization information does not necessary improve the similarity analysis. Consideration of parent, sibling and child sections in most cases worsens the relatedness analysis for cosine similarity approach because most of the elements in the frequency matrix are diluted by adding the frequencies of the concepts occurred in neighbor sections into the frequency matrix which is then normalized column by column. As a result, the similarity score of some correct concept-concept pairs is reduced to such an extent that is lower than the threshold similarity score. However, consideration of corpus structural information helps improve the prediction error for the market basket model. It can be explained by the fact that the occurrence probabilities of the concepts are usually so small that the addition to the frequency matrix does not affect the probabilities on the one hand, while on the other hand the inclusion of concepts occurred in neighbor sections can capture concept-concept pairs whose concepts are related but not occurred in the same section, leading to an increase in the association rule confidence from zero to a significant value. The resultant similarity scores of correct concept pairs are therefore raised.

Table 2: RMSE comparison of different relatedness analysis approaches

Thres-hold	$[w_p, w_s, w_c]$	Cosine Similarity	Jaccard Similarity	Market Basket
0.35	[0, 0, 0] (*no structural information*)	0.1000	0.1250	0.0850
	[0.7, 0, 0] (*parent only*)	0.1025	0.1300	0.0725
	[0, 0, 0.7] (*childs only*)	0.1025	0.1250	0.0775
	[0.7, 0, 0.7] (*parent and childs*)	0.1050	0.1275	0.0800
	[0.7, 0.3, 0.7] (*parent, siblings and childs*)	0.1075	0.1250	0.0750
0.4	[0, 0, 0]	0.1000	0.1300	0.0825
	[0.7, 0, 0]	0.1050	0.1300	0.0725
	[0, 0, 0.7]	0.1075	0.1250	0.0825
	[0.7, 0, 0.7]	0.1050	0.1300	0.0750
	[0.7, 0.3, 0.7]	0.1100	0.1300	0.0800
0.45	[0, 0, 0]	0.1025	0.1300	0.0925
	[0.7, 0, 0]	0.1050	0.1300	0.0800
	[0, 0, 0.7]	0.1025	0.1250	0.0875
	[0.7, 0, 0.7]	0.1050	0.1300	0.0800
	[0.7, 0.3, 0.7]	0.1125	0.1300	0.0950
0.5	[0, 0, 0]	0.1075	0.1300	0.0900
	[0.7, 0, 0]	0.1150	0.1300	0.0825
	[0, 0, 0.7]	0.1050	0.1300	0.0850
	[0.7, 0, 0.7]	0.1100	0.1300	0.0875
	[0.7, 0.3, 0.7]	0.1175	0.1300	0.1025

In addition, Table 2 shows that the RMSE increases with the threshold similarity score. As the threshold increases, the correctness of the predicted matches grows, which reduces the RMSE, while at the same time some correct matches yet with low similarity score will be discarded, which increases the error. The errors illustrated in Table 2 increase with the threshold score, implying that in this case the effect of discarding correct matches with low score outweighs that of improving the prediction correctness. Hence the choice of the threshold similarity score influences the quality of the ontology matching.

6. Related Work

Semantic similarity between words or concepts in a single taxonomy based on corpus statistics and lexical taxonomy has been studied (Jiang and Conrath 1997) to combine the node-based (information content) approach and the edge-based (distance) approach. Some node-based approaches (Resnik 1992, Resnik 1995) has concluded that the higher the occurrence probability of an instance of a concept in a corpus, the more information-rich the concept is. As assumed in (Resnik 1992, Resnik 1995) that a concept in a hierarchical structure subsumes the concepts lower in the hierarchy, the occurrence probability of a concept increases whereas information content decreases as one goes up the hierarchical concept structure. As for edge-based approaches, Jiang and Conrath (1997) summarized (Sussna 1993, Richardson and Smeaton 1995) to conclude that the distances between concepts in a hierarchy are not evenly distributed and network density, node depth, type of link and link strength are the determining factors of the distances, in which corpus statistics can be referred to for link strength calculation. The hybrid approach provides insight to the use of corpus statistics in semantic similarity measures. However, it requires that the concepts share the same hierarchical structure, which unfortunately does not hold true when mapping heterogeneous taxonomies.

Researchers in the field of linguistics and lexicography are also interested in the similarity and co-occurrence of concepts and words. Linguists try to classify words based on both their meanings and their co-occurrence with words while lexicographers attempt to explore word patterns and sentence patterns. Church and Hanks (1990) compares the joint probability of 2-word phrases among the whole set of all words and the individual probability of the two single words. By introducing the notion of mutual information *MI* (*x*, *y*) between two words *x* and *y*, compound relations (*computer scientist*, *United States*), semantic relations (*man woman*) and lexical relations (*coming from, set up*) can be located in a pool of words. Using similar measures, Hindle (1990) and Grefenstette (1992) derive the similarity of words from the distribution of syntactic context in a large text corpus. These researches also focus on word relatedness analysis using a document corpus. They impose no requirements on the choice of corpus and thus cannot make use of any structural characteristic of the corpus.

7. Conclusions

Three approaches have been tested to compare related concepts. Cosine similarity measure is to find the similarity of two concepts as the angle between the two frequency vectors of matched sections. It is similar to the reversed problem of finding similar documents by comparing the angle between the two vectors of n-shingles. Jaccard similarity coefficient is a statistic measure of the size of intersection relative to the size of union, that is the number of sections matched to both concepts divided by the number of sections matched to either concept. Market basket model is to discover interesting concept-to-concept association rules by using the theory of conditional probability.

The use of regulation structural information is also proposed. By making use of the information-rich well-structured hierarchical organization in regulations, more potentially related concept pairs can be extracted even though they do not co-exist in the same sections. It is achieved by considering the parent section, the set of sibling sections and the set of child sections for each section when establishing the frequency matrix before relatedness analysis.

This work focuses on ontologies from the construction domain. The International Building Code was utilized as the regulation corpus, and IfcXML and OmniClass, two commonly-used building data models, were extracted to provide the pools of concepts. With root mean square error as the performance metric, the result shows that the market basket model is the best approach for concept relatedness analysis while Jaccard similarity measure is the worst. It also shows that the regulation hierarchy information helps improve the relatedness comparison for the market basket model, but not cosine similarity and Jaccard similarity measures.

8. Acknowledgements

The authors would like to thank the International Code Council for providing the XML version of the International Building Code (2006). The authors would also like to acknowledge the supports by the National Science Foundation, Grant No. CMS-0601167, the Center for Integrated Facility Engineering 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 NIST, CIFE and NSF.

9. References

Church, K.W., and Hanks, P. “Word Association Norms, Mutual Information, and Lexicography.” *Computational Linguistics*, Volume 16, Issue 1, pages 22-29, 1990.

Construction Specifications Institute (CSI), *OmniClass Construction Classification System*, Edition 1.0, http://www.omniclass.org, 2006.

de Bruijn, J., Martin-Recuerda, F., Manov, D., and Ehrig, M., *State-of-the-art Survey on Ontology Merging and Aligning* V1.SEKT-project report D4.2.1 (WP4), IST-2003-506826, EU-IST Integrated Project (IP), EU, 2004.

Ehrig, M., and Sure, Y. “Ontology Mapping – An Integrated Approach.” *Proceedings of the First European Semantic Web Symposium*, Volume 3053, Lecture Notes in Computer Science, pages 76–91, Heraklion, Greece, May 2004.

Euzenat, J., Le Bach, T., Barasa, J., et.al. *State of the Art on Ontology Alignment*, Technical Report KWEB/2004/D2.2.3/v1.2, EU-IST Knowledge Web (KWEB), EU, 2004.

Gallaher, M., O'Connor, A., Bettbarn Jr., J., and Gilday, L. "Cost Analysis of Inadequate Interoperability in the US Capital Facilities Industry." *Technical Report GCR 04-867*, NIST, 2004.

Grabar N., and Zweigenbaum P. "Automatic Acquisition of Domain-Specific Morphological Resources from Thesauri." *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, pages 765-784, Paris, France, April, 2000.

Grefenstette, G. "Use of Syntactic Context to Produce Term Association Lists for Text Retrieval." *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval*, SIGIR'92, pages 89-97, Copenhagen, Denmark, 1992.

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

Hindle, D. "Noun Classification from Predicate-Argument Structures." *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, ACL28'90, pages 268-275, Pittsburgh, Pennsylvania, 1990.

International Alliance for Interoperability (IAI). *Guidelines for the development of industry foundation classes*, IAI, May 1997.

International Building Code (IBC) 2006, International Conference of Building Officials (ICBO), Whittier, CA, 2006.

Jacobs, J. and Linden, A. *Semantic Web Technologies Take Middleware to the Next Level*, Technical Report T-17-5338, Gartner Group, 2002 (see http://www.gartner.com/ DisplayDocument?doc_cd=109295)

Jiang, J., and Conrath, D. "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy." *International Conference Research on Computational Linguistics*, ROCLING X, Taiwan, 1997.

Larsen, B., and Aone, C. "Fast and Effective Text Mining Using Linear-Time Document Clustering." *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16-22, San Diego, California, USA, 1999.

Lau, G.T., Law, K.H., and Wiederhold, G. "Comparative Analysis of Government Regulations Using Structural and Domain Information." *IEEE Computer*, Volume 38, Issue 12, pages 70-76, Dec 2005.

Lau, G.T., Law, K.H., and Wiederhold, G. "A Relatedness Analysis of Government Regulations using Domain Knowledge and Structural Organization." *Information Retrieval*, Volume 9, Issue 6, pages 657-680, Sep 2006.

Lovins, J.B. "Development of a Stemming Algorithm." *Mechanical Translation and Computational Linguistics*, Volume 11, pages 22-31, 1968.

Nahm, U.Y., Bilenko, M., and Mooney, R.J. "Two Approaches to Handling Noisy Variation in Text Mining." *Proceedings of the ICML-2002 Workshop on Text Learning*, pages 18-27, Sydney, Australia, July 2002.

NIST. *Interoperability Cost Analysis of the US Automotive Supply Chain*, Planning Report #99-1, NIST Strategic Planning and Economic Assessment Office, 1999 (available at http://www.nist.gov/director/prog-ofc/report99-1.pdf 1999).

Porter, M.F. "An Algorithm for Suffix Stripping." *Program*, Volume 14, Issue 3, pages 130-137, 1980.

Ray, S. "Interoperability Standards in the Semantic Web." *Journal of Computing and Information Science in Engineering*, ASME, Volume 2, pages 65-69, March, 2002.

Resnik, P. "WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery." *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, pages 56-64, San Jose, CA, July 1992.

Resnik, P. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy." *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1, pages 448-453, Montreal, August 1995.

Richardson, R., and Smeaton, A. *Using WordNet in a Knowledge-Based Approach to Information Retrieval*, Technical Report CA-0395, Dublin City Univ., School of Computer Applications, Dublin, Ireland, 1995.

Roussinov, D., and Zhao, J.L. "Automatic Discovery of Similarity Relationships Through Web Mining." *Decision Support Systems*, Vol. 25, pages 149-166, 2003.

Salton, G. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, 1989.

Sussna, M. “Wordsense Disambiguation for Free-Text Indexing Using a Massive Semantic Network.” *Proceedings of the Second International Conference on Information and Knowledge Management*, CIKM-93, pages 67-74, Arlington, Virginia, 1993.

Watson, A., and Crowley, A. “CIMSteel Integration Standard”, in Scherer R.J. (Eds.), *Product and Process Modelling in the Building Industry*, A.A. Balkema, Rotterdam, pages 491-493, 1995.

van Hage, W., Katrenko, S., and Schreiber, G. “A Method to Combine Linguistic Ontology-Mapping Techniques.” *Fourth International Semantic Web Conference (ISWC)*, pages 732-744, 2005.