# Relating Taxonomies with Regulations

Chin Pang Cheng, Jiayi Pan
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
(cpcheng, pjy@stanford.edu)

Gloria T. Lau, Kincho H. Law
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
(glau, law@stanford.edu)

Albert Jones
Enterprise Systems Group
NIST
Gaithersburg, MD 20899-0001
(albert.jones@nist.gov)

## ABSTRACT

Increasingly, taxonomies are being developed for a wide variety of industrial domains and specific applications within those domains. These taxonomies attempt to represent formally the vocabularies commonly used by domain practitioners. These formal representations have the potential to automate information retrieval and improve decision-making. Those decisions must comply with existing government regulations and codes of practice, which are not always known to the practitioners. Although both are available in digital form online, practitioners cannot retrieve easily relevant regulations and codes that apply to particular decisions.

To address this problem, we propose an approach to relate regulations with existing industry-specific taxonomies. The mapping from a single taxonomy to a single regulation is a trivial keyword matching task. In this paper, we examine techniques to map a single taxonomy to multiple regulations, as well as to map multiple taxonomies to a single regulation. Those techniques include Cosine similarity, Jaccard coefficient and market-basket analysis. These techniques provide a metric that measures the similarity between concepts from different taxonomies. We describe these techniques and metrics, and evaluate them using examples from the building industry. These examples show the potential regulatory benefits from the mapping between various taxonomies and regulations.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models*, J.1 [**Administrative Data Processing**]: *law.*

## Keywords

Heterogeneous Ontologies, Taxonomy Interoperability, Relatedness Analysis, Regulation Retrieval.

## 1. INTRODUCTION

Government regulations are an important asset of the society. They extend the laws governing the country with specific guidance for corporate and public actions. Ideally, regulations should be readily retrievable by interested individuals. Much prior research focused on the abstraction and retrieval of case law [1, 3, 5, 30, 38, 40], analysis of regulations [19, 20], and compliance guidance for regulations [15, 16]. Relatively little research, however, has been devoted to methodologies and tools that allow practitioners to *intelligently browse and retrieve* relevant regulations utilizing familiar terms and vocabularies.

Increasingly, taxonomies are being developed to capture and represent those terms and vocabularies formally for a wide variety of industrial domains. These taxonomies can facilitate information integration and regulation retrieval. Interoperability is important because it allows practitioners to access, relate, and combine information from multiple, heterogeneous sources. Recent studies by the National Institute of Standards and Technology showed that integration inefficiencies led to significant costs to the construction as well as the automotive industries [7, 14].

Ontologies have been proposed as a way to remove these inefficiencies. One recent forecast estimates that "By 2010, ontologies ….will be the basis for 80 percent of application integration projects" [32]. Ontologies capture the semantics of domain-specific information in a formal and computer interpretable form. They have the potential to automate much of the integration process, thereby reducing cost and time significantly.

Building a single ontology for an entire domain, however, has proven to be neither efficient nor practical. Rather, small communities that need to exchange information frequently tend to build their own distinct ontologies [36]. For instance, the architectural, engineering, and construction (AEC) community has built several ontologies that describe the semantics of buildings. Even these ontologies are all targeted towards the same user group, their structures, vocabularies and coverage differ depending on the specific application.

Government regulations, on the other hand, are organized and classified by the needs of the agency that enforces them, not by the needs of the communities that must use them [6]. Consequently, there is a clear need and benefit of bridging these two distinct needs. One way to build such a bridge is to develop methods and tools that enable practitioners to browse and retrieve government regulations using their own terms and vocabularies.

In this paper, we present a systematic method to map regulations to industry-specific taxonomies. First, we use a trivial keyword extraction task to map a single taxonomy to a single regulation. To map one taxonomy to multiple regulations, we must cluster relevant sections from different regulations. To do this, we reuse the relatedness analysis core from [19] to compute relevancy between those sections. We describe three methodologies to compute this relevancy: Cosine similarity, Jaccard coefficients, and market-basket models. Cosine similarity and Jaccard

coefficient are vector-based measures commonly used in the field of information retrieval. We have adopted them to compare semantic similarity between ontologies. The market basket model is a popular technique in data mining; we have modified it to be a relatedness analysis measure for ontology mapping. We discuss our preliminary evaluations of the three metrics. We conclude with our proposal to address methods for mapping multiple taxonomies to multiple regulations.

## 2. Illustrative Ontology Standards and Regulatory Corpus

We work with taxonomies and regulatory corpus from both the building industry and the environmental protection industry [15, 16, 19, 20]. To illustrate their organization and structure, we present briefly the ontology standards and classification systems that are commonly used in the building industry. For the AEC industry, there are a few ontologies that describe the semantics of building. They include the CIMsteel Integration Standards (CIS/2) for the steel building and fabrication industry [8], the Industry Foundation Classes (IFC) initiated by the CAD vendors for design description of building components [12], and the OmniClass construction classification system (OmniClass) for the construction specification, materials, and product components [34].

Figures 1 and 2 show excerpted examples of the *OmniClass* and *IfcXML* standards. Typical of ontology standards, both are organized hierarchically with implicit "is-a" type relationships defined accordingly. OmniClass consists of 15 tables, each of which represents a different facet of construction information. Each term is associated with a unique ID. For example, the term "Sound and Signal Devices" is associated with the ID "23-85 10 1111."

For the IfcXML, the Industry Foundation Class objects are expressed in an XML structure that defines the hierarchical relationship between elements and entities. To extract the object terms for mapping purposes, the two standards are preprocessed to eliminate the miscellaneous information - such as the IDs in the OmniClass and the element names, group names and type names in the IfcXML - as well as the duplicated terms.

**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

**Figure 1: Excerpt from OmniClass Construction Classification System**

```
    </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 2: Organization of IfcXML**

Regulations are voluminous and cover a broad range of scopes and topics. Increasingly, regulatory documents are available online and organized in XML structure. The International Building Code (IBC) [13], which represents the code of practice in the building industry, is employed as one of the regulatory document corpuses. Figure 3 shows a provision in IBC and its representation in an XML structure. One notable feature of regulations is that they are typically organized into sections and sub-sections, each of which contains contents with a specific topic or scope. The tree hierarchy of regulations provides useful information that can be explored, for example, to locate similar sections and to build an e-government system [19, 20].

**[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.

```
<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>
</RECORD>
<LEVEL level-depth="0" style-id="0-0-0-0" style-name="Normal
Level" style-name-escaped="Normal-Level" toc-section="false">
<RECORD id="0-0-0-5530" number="5530" version="3">
<PARA style-id="0-0-0-15" style-name="Body3" style-name-
escaped="Body3">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.</PARA>
</RECORD>
</LEVEL>
</LEVEL>
```

**Figure 3: An IBC Provision and XML Structure**

## 3. ONE TAXONOMY TO ONE REGULATION

Mapping one taxonomy to one regulation is a simple keyword latching task. There are many commercial tools available to latch keywords from documents into a taxonomy. Industry taxonomies are hierarchical classification systems, which are generally less than 10 levels deep. Node labels in the taxonomy tree are treated as concept keywords, and they are mapped to the sections in the regulation where they appear. As regulations tend to be voluminous, we use a section or subsection as a unit of interest. Figure 4 shows the International Building Codes (IBC) [13] latched with the OmniClass. Users can then traverse the taxonomy and browse relevant sections of the regulation.

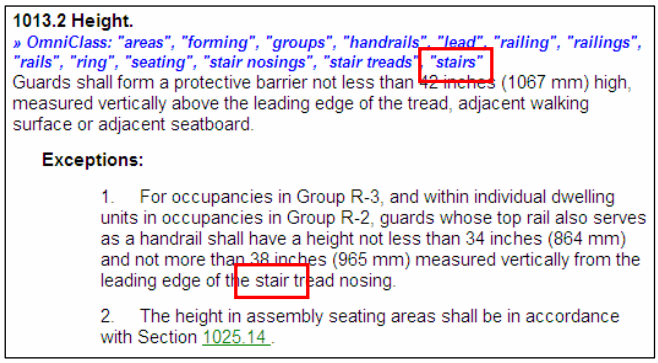



**Figure 4: Regulation Latched with Taxonomy Concepts**

Extending the mapping from one taxonomy to multiple regulations unfortunately leads to the classic problem of information overload. For instance, suppose we want to search the Web to find state regulations governing chlorine levels in drinking water. If we search the drinking-water regulations in Alabama and Arizona for the concept "chlorine," we would find over 30 sections in each. The actual relevancy of these 60 sections to chlorine levels is not known. The problem is that Web content ignores the actual structure of the documents. Consequently, search engines cannot take that structure into account when computing relevancy. The result is that users quickly become frustrated with information overload [4].

Fortunately, regulatory documents are much more structured than web content. We propose to solve the problem of information overload by clustering relevant sections from different regulations based on that structure. We discuss our approach in the following sections.

## 4. ONE TAXONOMY TO MULTIPLE REGULATIONS

Simultaneous traversal of multiple regulation trees using one taxonomy is a challenging but real problem. It is not uncommon for industry practitioners to be familiar with one particular regulation only. For example, an architect from Montgomery might be familiar with Alabama state code, but not Arizona state code. Nonetheless, if he were to design a water distribution system that provides water to Phoenix from lakes near Montgomery, he would need an understanding of both [9].

In this scenario, finding the relevant Arizona regulations on chlorine levels might pose a serious problem. We believe that it is beneficial to map the taxonomy to Alabama code first, and then branch out to recommend related sections from the Arizona code. In general, focusing on one regulation as the basis for finding relevant sections from other regulations significantly reduces information overload.

Figure 5 shows a simple user interface that shows a scenario of finding related provisions between regulations from the two states. After browsing down the taxonomy tree to the concept "chlorine," users are shown a list of matched sections from the Alabama regulation. As discussed in Section 3, matching sections to the taxonomy concept is simply keyword latching. Selecting Section 335.7.6.15 of the AL code shows that there are 15 recommended sections from the Arizona regulation. A user can stay focused on the regulation of their choice, and at the same time acquire relevant sections from other regulations as needed.

There are two major challenges to developing such a system: a suitable user interface and a methodology for determining relevant regulations. In this paper, we focus on methodologies for making recommendations based on relevancies between sections from different regulations.

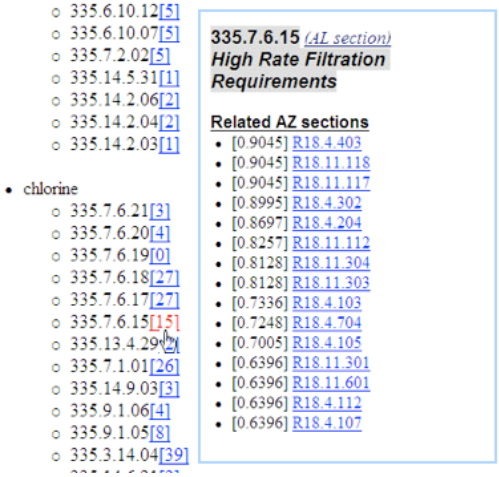



**Figure 5: Chlorine mapped to Section 335.7.6.15 in AL code, which have 15 related sections in AZ code**

To identify related provisions from different regulations, we reuse the relatedness analysis core from [19, 20]. That analysis compares sections from different regulations using a Cosine similarity measure (see Section 5.1) [18, 31]. The goal is to identify the most strongly related provisions using (1) traditional term matches and a combination of feature matches, and (2) content comparisons as well as structural analysis. Regulations are first compared based on a combination of conceptual information and domain knowledge using feature matching. Regulations also possess specific structures, such as a tree hierarchy of provisions and the referential structure. These structures contain useful information for locating related provisions and are, therefore, used in the analysis as well. For the detailed discussion on the evaluations of results from the relatedness analysis of provisions see [19].

# 5. MULTIPLE TAXONOMIES TO ONE REGULATION

As noted above, multiple taxonomies have been developed for different applications within the same industry domain. Most practitioners are familiar with at least one of them; but, they frequently need to deal with others for various applications [2, 23]. Therefore, traversing a single set of regulations using multiple taxonomy trees poses a real but non-trivial problem. There are many research efforts on ontology merging [33, 39]. These efforts produce a merged ontology that can be used for data interoperability but not as a front-end representation format. Since users would need to learn this new merged ontology in order to browse regulations, this would defeat the original intent of using the existing taxonomies.

Using the same argument from Section 4, we believe that focusing on one familiar taxonomy is the right starting point to traverse regulations. Once users reach a taxonomy node of interest, related concepts from other taxonomies can be suggested and users can switch their focal point from one taxonomy to another.

Figure 6 illustrates the proposed approach using the OmniClass [34] and the IFC [12] taxonomies, and the International Building Code (IBC) regulations [13]. This figure assumes the practitioner is familiar with OmniClass and IBC, but not IFC. The practitioner uses the term "steel decking" from OmniClass to find an ordered list of matching IBC sections and relevant IFC concepts. Upon locating a list - sorted in order of relevance - of IBC sections that are related to "steel decking," the user also sees a list of related IFC concepts including "slab." Mousing-over the IFC concept "slab" brings the focal point to the IFC hierarchy, where the user is presented with the same analysis – namely the IFC elements related to the concept "slab," a ranked list of matching IBC sections, and a ranked list of relevant OmniClass concepts.

Once these related concepts have been found, the next task is to develop mapping between them. Ontology mapping has been an active research area since the semantic web movement began [28, 29]. In general, it is very difficult to develop mappings between two arbitrary ontologies. In our case, however, the problem is slightly more manageable because our ontologies are very industry specific and are targeted towards the same group of users.

Similar to the techniques presented in Section 4, we compute the relevance among concepts from different ontologies using a vector comparison approach. A document corpus is used to relate concepts by computing their co-occurrence frequencies. This training corpus must be carefully selected since it represents the relevancy among concepts from different taxonomies. Conveniently, we have a corpus of regulatory documents that has been meticulously drafted and reviewed for accuracy. Unlike web content, regulations do not have random co-occurrences of phrases in the same provision. This dramatically increases the likelihood of finding real matches.

Consider a pool of $m$ concepts 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 $i$ 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 $k$. In subsequent sections, we will discuss three metrics to compute the similarity score among concepts. In Figure 6, the Cosine similarity scores for several concepts related to "steel decking" are shown. The score for "slab" is 0.895, which ranks second among all IFC concepts that are relevant to "steel decking."

## 5.1 Cosine Similarity

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 [18, 31]. Given two frequency vectors $\vec{c}_i$ and $\vec{c}_j$, the similarity score between concepts $i$ and $j$ is represented using the dot product:

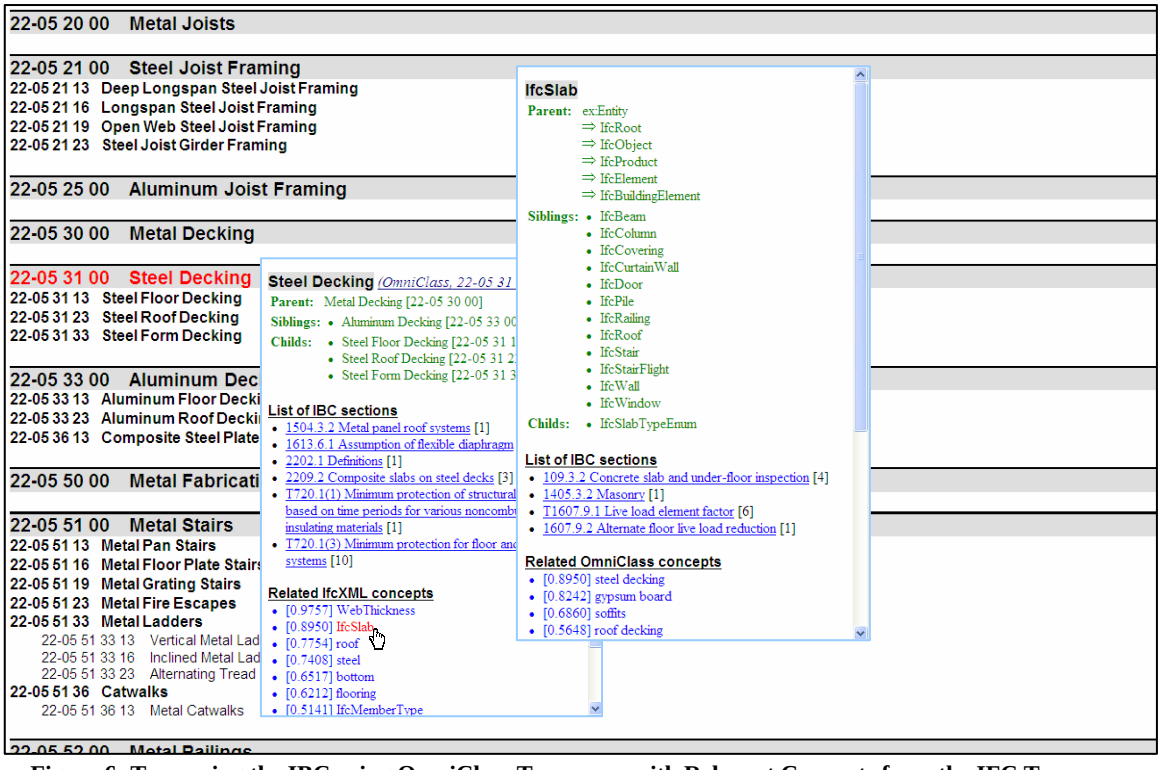



**Figure 6: Traversing the IBC using OmniClass Taxonomy with Relevant Concepts from the IFC Taxonomy**

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

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

## 5.2 Jaccard Similarity Coefficient

Jaccard similarity coefficient [31, 37] 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:

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

Jaccard similarity coefficient is a popular measure of term-term similarity due to its simplicity and retrieval effectiveness [17]. 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 our problem, 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

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

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

## 5.3 Market-Basket Model

Market-basket model is a probabilistic data-mining technique to find item-item correlation [11]. 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 concept *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 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

$$Conf(i \rightarrow j) = \frac{N_{11}}{N_{11} + N_{01}}$$

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

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

The value ranges from -1 to 1. The value of -1 means that concept *j* appears in every section while concept *i* does not co-occur in any of these sections. 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 occurrence of concept *j* is not significant in the corpus, but it appears in every section that concept *i* appears.

## 5.4 Use of Regulation Hierarchy Structural Information

Many related concepts can be uncovered by treating each section in a regulation as an independent document. A concept-document matrix is generated to compute concept co-occurrence in documents, which are really regulatory sections. This approach is generally sufficient in revealing most related concepts, but some related concepts rarely co-occur in the same sections. For example, if two concepts contain an Is-A-relationship, like door furniture and door hardware, they may be used in the same regulation interchangeably but in different sections.

*Is-A*-related concepts are also hard to find when each section is treated as if it were an independent document. The relationship between *Is-A*-related concepts, such as "building materials" and "concrete" as shown in Figure 7, 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 section. Instead, the sections describing "concrete" are usually the subsections of the sections describing "building materials." If we consider a subsection and its parent section in the computation of the similarity score between "building materials" and "concrete," the implicit relationship between building materials and concrete might become more obvious. Therefore, the hierarchical structure of sections needs to be considered to extract non-trivial related concepts.

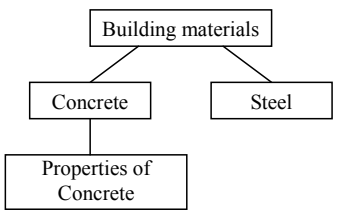



**Figure 7: Example of related but rarely co-occurring concepts**

Regulations contain well-organized hierarchical structures with sections and sub-sections of specific topic or scope. There are organizational and referential structures explicitly defined in

regulations. Section 4 briefly discussed the usage of the regulation hierarchical information to locate related sections from different regulation trees. The results show that regulatory structure sometimes helps to reduce prediction error of related provisions [19]. Here, we will use the regulation hierarchy to uncover semantic relationships between concepts from different taxonomies.

Well-structured regulations can be represented as a hierarchical tree, where each section corresponds to a discrete node. As illustrated in Figure 8, each section has a parent section, a set of sibling sections, and a set of child sections. In general, for a section with a particular topic, the parent section describes a broader topic, the sibling sections describe parallel topics, and the child sections describe more specific topics. In our computation, we will consider the co-occurrence of concepts in a broader scope, namely the parent, sibling and child sections.

**Figure 8: Tree hierarchy of sections in regulations**

The frequency matrix C is modified to take the parent section, sibling sections and child sections into consideration. To include the parent section, the weighted numbers of occurrence for all the concepts in the parent section are added to the numbers of occurrence in the self section. Similarly, the sibling sections and child sections are then included with a discounted weight. In our formulation below, we will denote *Par(k)*, *Sib(k)* and *Child(k)* as the parent section, set of sibling sections and set of child sections of Section k. The *k*-th element of frequency vector $\vec{c}_i$, i.e., the number of times concept *i* is matched to Section *k,* is updated as

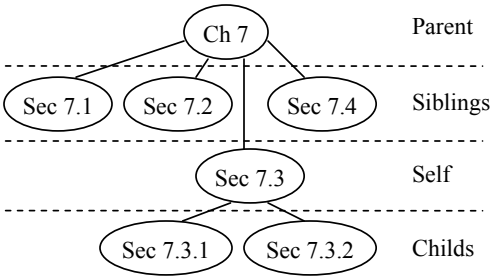
$$c_i(k) := c_i(k) + w_p c_i(Par(k)) + w_s \sum_{u \in Sib(k)} c_i(u) + w_c \sum_{v \in Child(k)} c_i(v)$$

where $w_p$, $w_s$ and $w_c$ are the weights of the parent, sibling and child sections respectively.

## 5.5 Evaluations of the Measures

In summary, we randomly selected twenty concepts from the OmniClass and the IFC hierarchies respectively and computed pairwise similarity scores using the three approaches described above. In addition, we interviewed domain experts and used their matches as the true matches. We used root mean square error (RMSE), precision, recall and F-measure as performance metrics to evaluate and compare both the three measures and the use of regulation structural information. A baseline ontology matcher is compared to the three measures using precision and recall as the evaluation metrics.

Three domain experts identified the related concept pairs among a total of 400 possible pairs. Related concept pairs are assigned a true value of one; all other pairs are assigned a true value of zero. As for the predicted values, two concepts are predicted as similar or related if the computed similarity score is larger than certain threshold scores. Given these true and predicted values, we computed values of RMSE, precision, recall and F-measure for the three measures, the baseline ontology matcher, and different regulation structural information inclusions. The averages of the results from the three true answers are then taken as the final results. Details are given in the following sections.



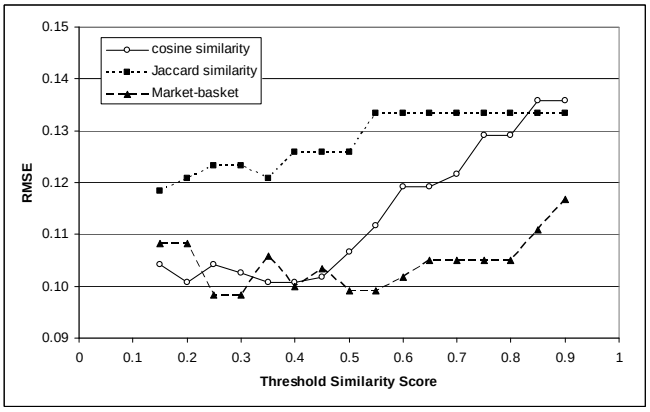

### *5.5.1 Root Mean Square Errors (RMSE) among the Three Measures*

Root mean square error (RMSE) is a metric to evaluate the accuracy of the predicted values against the true values. Comparison between ontology of *m* concept terms and ontology of *n* concept terms involves *m* by *n* concept-concept pairs. Therefore the RMSE is calculated as

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

Figure 9 shows the results of the three measures using RMSE and provisions from regulations as independent documents in the co-occurrence computation. The graph shows results for threshold similarity scores ranging from 0.15 to 0.9. We conclude, when no regulation hierarchy structural information is considered, that the market-basket model results in the lowest RMSE for most threshold similarity scores. This means that the market-basket model outperforms the other two measures in locating related concept pairs from different ontologies.

**Figure 9: Evaluation results of the three measures using RMSE**




### *5.5.2 Precision, Recall and F-measure among the Three Measures*

We used precision, recall and F-measure values to compare the three similarity analysis measures when we included regulation hierarchy structural information. While RMSE takes both correctness and incorrectness of prediction into consideration, precision and recall emphasize correctness only. Precision and recall evaluates the accuracy of predictions and the coverage of accurate pairs. It does this by measuring 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

$$Precision = \frac{|True\ Matches \cap Predicted\ Matches|}{|Predicted\ Matches|}$$

$$Recall = \frac{|True\ Matches \cap Predicted\ Matches|}{|\text{True}\ Matches|}$$

There is always a tradeoff between precision and recall. F-measure is, therefore, leveraged to combine both metrics. It is a weighed harmonic mean using 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

$$F - Measure = \frac{2 \cdot (Precision \times Recall)}{Precision + Recall}$$

Figure 10 shows the results for the three measures. The market-basket model shows the highest F-measure values in all cases, again, consistent with the RMSE results. In fact, market-basket model achieves the highest recall rate with relatively high precision in all cases. Jaccard similarity is not preferred due to its low F-measure values, resulted from its very low recall rates. Cosine similarity falls in between; this is consistent with the RMSE results.

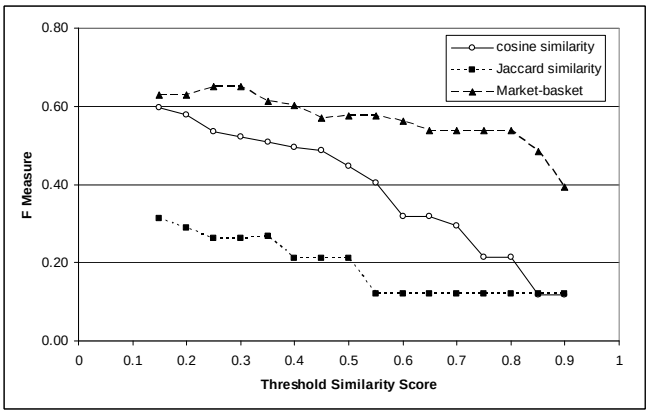



**Figure 10: Evaluation results of the three measures using F-measure**

Since the market-basket model outperforms Cosine and Jaccard similarities in both cases, we will evaluate the impact of regulation hierarchy using the market-basket model only.

As shown in Figure 11, the effect of including regulatory structure in the analysis is inconclusive. In general, it increases recall and reduces precision, as more regulatory nodes are considered to locate related concepts. The inclusion of parent section produced a slightly higher F-measure for most threshold scores. This is likely due to the fact that the parent relationship is one to one which minimizes the impact on precision. Other relationships, such as sibling and child, are not one to one; the number of such relationships, therefore, could heavily tax precision with only minor increase in recall.

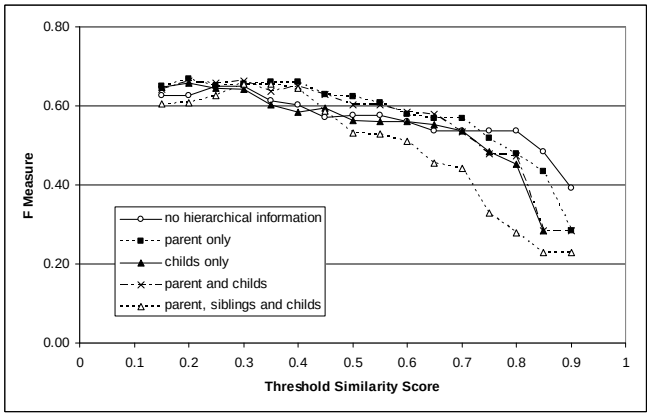



**Figure 11: Evaluation results of market-basket model using F-measure**

### *5.5.3 Comparison of the Domain-based Model with the Lexicon-based Model*

In addition to comparing the three measures with one another, we also evaluated our domain-based approach to a traditional lexicon-based approach. Ontology mapping is an active research topic, and common mapping methods discover the semantic similarity between ontology elements using rule-based [22, 28], lexicon-based [24, 35] and structure-based methods [25, 27]. Our approach is comparable to a lexicon-based approach, where dictionary and thesaurus are used to enumerate synonyms, homonyms, and so on. In our analysis, we use a domain-specific corpus of regulations to uncover such semantic relationships.

A thesaurus is necessary to compare our approach to a lexicon-based one. A common thesaurus and well-known lexical resource for the English language is WordNet [26]. It is also one of the most widely adopted synonym sources for ontology matching techniques including CUPID [24], Learned Ontology Model (LOM) [21], and Version Matching Approach (VMA) [41]. Table 1 shows the result for comparing a domain-based ontology mapping method with a lexicon-based element matcher using WordNet.

Table 1 shows that our approach outperforms the lexicon-based matcher in terms of precision and recall. Some examples of matches that are found by our domain-based matcher but not by the lexicon-based matcher are: (sound and signal devices, IfcSwitchingDeviceType), (steel decking, IfcSlab), (door hardware, IfcBuildingElementComponent), and (sound and signal devices, IfcAlarmType). The reliability of lexicon-based matchers is not guaranteed because their use of stemmers to reduce derived words to their root form is not always appropriate for the domain [10]. In addition, many concepts have different meanings when used in different domains, so that their synonyms and definitions could be different.

We should note that WordNet is a generic linguistic thesaurus and not an industry-specific taxonomy. Consequently, it contains little and imprecise information related to the terminology used by the OmniClass and IfcXML. The result shows that domain-related corpora, such as regulations and technical specifications, are useful in discovering the semantic relationships across multiple ontologies.

**Table 1: Precision and recall comparisons of domain-based ontology mapping to lexicon-based ontology mapping**

| Score threshold | Approaches | Cosine Similarity | | Jaccard Similarity | | Market-basket Model | |
|---|---|---|---|---|---|---|---|
| | | P | R | P | R | P | R |
| 0.2 | Lexicon-based Matcher | 0.50 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Domain-based Matcher | 0.79 | 0.53 | 0.91 | 0.17 | 0.70 | 0.71 |
| 0.3 | Lexicon-based Matcher | 0.50 | 0.03 | 1.00 | 0.03 | 0.50 | 0.03 |
| | Domain-based Matcher | 0.83 | 0.41 | 0.90 | 0.15 | 0.75 | 0.71 |
| 0.4 | Lexicon-based Matcher | 1.00 | 0.03 | 1.00 | 0.03 | 0.50 | 0.03 |
| | Domain-based Matcher | 0.91 | 0.36 | 1.00 | 0.12 | 0.80 | 0.59 |
| 0.5 | Lexicon-based Matcher | 1.00 | 0.03 | 1.00 | 0.03 | 1.00 | 0.03 |
| | Domain-based Matcher | 0.90 | 0.31 | 1.00 | 0.11 | 0.81 | 0.51 |
| 0.6 | Lexicon-based Matcher | 1.00 | 0.03 | 1.00 | 0.03 | 1.00 | 0.03 |
| | Domain-based Matcher | 0.92 | 0.20 | 1.00 | 0.07 | 0.81 | 0.49 |

## 6. CONCLUSIONS & FUTURE TASKS

Hierarchically structured regulatory documents are written by government agencies who organize the material to suit their own needs. In this paper, we proposed a system to map concepts from industry-specific taxonomies to similar concepts in those regulations to increase their usability by industry practitioners.

To measure similarity, we proposed and evaluated three measures: Cosine similarity, Jaccard coefficients, and market-based models. We used a running example from the AEC industry to illustrate the need, the usage, and the benefit of the mapping system and measures. Using that example, we considered 1-1, 1-n, n-1 mappings between taxonomies and regulations and showed that market-based models were superior to the other two.

We plan to implement an n-n concept-section mapping in the future, by combining the techniques of concept comparisons and section comparisons. In section comparisons, the hierarchical structure of regulations is used to enhance the analysis; we also plan to incorporate the hierarchical information of taxonomies into concept comparisons. In the future, we plan to engage potential users to help perform formal evaluations of the similarity metrics and the usability of the system.

## 7. ACKNOWLEDGMENTS

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 (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 and NIST.

## 8. DISCLAIMER

Certain commercial software products may be identified in this paper. This use does not imply approval or endorsement by NIST, nor does it imply these products are necessarily the best available for the purpose.

## 9. REFERENCES

[1] K. Al-Kofahi, A. Tyrrell, A. Vachher and P. Jackson. "A Machine Learning Approach to Prior Case Retrieval," In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, pp. 88-93, 2001.

[2] E.F. Begley, M.E. Palmer and K.A. Reed. Semantic Mapping Between IAI ifcXML and FIATECH AEX Models for Centrifugal Pumps, Technical, 2005.

[3] T.J.M. Bench-Capon. Knowledge Based Systems and Legal Applications, Academic Press Professional, Inc., San Diego, CA, 1991.

[4] N. Bonnel, V. Lemaire, A. Cotarmanac'h and A. Morin. "Effective Organization and Visualization of Web Search Results," In Proceedings of the 24th IASTED International Conference on Internet and Multimedia Systems and Applications, Innsbruck, Austria, pp. 209-216, 2006.

[5] S. Br□inghaus and K.D. Ashley. "Improving the Representation of Legal Case Texts with Information Extraction Methods," In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, pp. 42-51, 2001.

[6] J.E. Fountain. Information Institutions and Governance: Advancing a Basic Social Science Research Program for Digital Government, Technical Report, National Center for Digital Government, John F. Kennedy School of Government, Harvard University, 2002.

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

[8] F. Garas and I. Hunter. "CIMSteel (Computer Integrated Manufacturing in Constructional Steelwork) - Delivering the Promise," Structural Engineering, 76 (3), pp. 43-45, 1998.

[9] M.P. Gibbens. CalDAG 2000: California Disabled Accessibility Guidebook, Builder's Book, Canoga Park, CA, 2000.

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

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

[12] Industry Foundation Classes (IFC), International Alliance for Interoperability (IAI), 1997.

[13] International Building Code 2000, International Conference of Building Officials (ICBO), Whittier, CA, 2000.

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

[15] S. Kerrigan. A Software Infrastructure for Regulatory Information Management and Compliance Assistance, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 2003.

[16] S. Kerrigan and K. Law. "Logic-Based Regulation Compliance-Assistance," In Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003), Edinburgh, Scotland, pp. 126-135, Jun 24-28, 2003.

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

[18] B. Larsen and C. Aone. "Fast and Effective Text Mining Using Linear-Time Document Clustering," In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, pp. 16-22, 1999.

[19] G. Lau. A Comparative Analysis Framework for Semi-Structured Documents, with Applications to Government Regulations, Ph.D. Thesis, Civil and Environmental Engineering, Stanford University, Stanford, CA, 2004.

[20] G. Lau, K. Law and G. Wiederhold. "Legal Information Retrieval and Application to E-Rulemaking," In Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL 2005), Bologna, Italy, pp. 146-154, Jun 6-11, 2005.

[21] J. Li. "LOM: A Lexicon-based Ontology Mapping Tool," In Proceedings of the Information Interpretation and Integration Conference (I3CON) and the Performance Metrics for Intelligent Systems (PerMIS) Workshop, Gaithersburg, MD, Aug 25, 2004.

[22] W. Li, C. Clifton and S. Liu. "Database Integration using Neural Network: Implementation and Experiences," Knowledge and Information Systems, 2 (1), pp. 73-96, 2000.

[23] R. Lipman. "Mapping Between the CIMsteel Integration Standards (CIS/2) and Industry Foundation Classes (IFC) Product Model for Structural Steel," In Proceedings of the Conference on Computing in Civil and Building Engineering, Montreal, Canada, pp. 3087-3096, Jun 14-16, 2006, 2006.

[24] J. Madhavan, P.A. Bernstein and E. Rahm. "Generic Schema Matching with Cupid," In Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Rome, Italy, pp. 49-58, Sept 11-14, 2001.

[25] S. Melnik, H. Garcia-Molina and E. Rahm. "Similarity Flooding: A Versatile Graph Matching Algorithm," In Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose, CA, pp. 117-128, Feb 26-Mar 1, 2002.

[26] G.A. Miller, R. Beckwith, C. Fellbaun, D. Gross and K. Miller. Five Papers on WordNet, Technical Report, Cognitive Science Laboratory, Princeton, NJ, 1993.

[27] T. Milo and S. Zohar. "Using Schema Matching to Simplify Heterogeneous Data Translation," In Proceedings of the 24th International Conference On Very Large Data Bases, New York, NY, pp. 122-133, 1998.

[28] P. Mitra. An Algebraic Framework for the Interoperation of Ontologies, Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, CA, 2003.

[29] P. Mitra and G. Wiederhold. "Resolving Terminological Heterogeneity in Ontologies," In Proceedings of Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence (ECAI), Lyon, France, pp. 45-50, 2002.

[30] M.-F. Moens, C. Uyttendaele and J. Dumortier. "Abstracting of Legal Cases: The SALOMON Experience," In Proceedings of the 6th International Conference on Artificial Intelligence and Law (ICAIL 1997), Melbourne, Australia, pp. 114-122, 1997.

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

[32] NIST. Interoperability Cost Analysis of the US Automotive Supply Chain, Technical Report, #99-1, http://www.nist.gov/director/prog-ofc/report99-1.pdf, NIST Strategic Planning and Economic Assessment Office, 1999.

[33] N.F. Noy. "Tools for Mapping and Merging Ontologies," In S. Staab and R. Stude (Eds.), Handbook on Ontologies, Springer-Verlag, pp. 365-384, 2003.

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

[35] L. Palopoli, D. Sacca, G. Terracina and D. Ursino. "A Unified Graph-based Framework for Deriving Nominal Interscheme Properties, Type Conflicts and Object Cluster Similarities," In Proceedings of the 4th IFCIS International Conference On Cooperative Information Systems (CoopIS), Edinburgh, Scotland, pp. 34-45, Sep 2-4, 1999.

[36] S. Ray. "Interoperability Standards in the Semantic Web," Journal of Computing and Information Science in Engineering, 2, pp. 65-69, 2002.

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

[38] E. Schweighofer, A. Rauber and M. Dittenbach. "Automatic Text Representation, Classification and Labeling in European Law," In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, pp. 78-87, 2001.

[39] G. Stumme and A. Maedche. "Ontology Merging for Federated Ontologies on the Semantic Web," In Proceedings of the International Workshop on Foundations of Models for Information Integration (FMII 2001), Seattle, WA, pp. 16-18, 2001.

[40] P. Thompson. "Automatic Categorization of Case Law," In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri, pp. 70-77, 2001.

[41] H. Wang, B. Akinci and J. Garrett. "A Formalism for Detecting Version Differences in Data Models," Journal of Computing in Civil Engineering, 21 (5), pp. 321-330, 2007.