Regulation Retrieval Using Industry Specific Taxonomies

Abstract Increasingly, taxonomies are being developed and used by industry practitioners to facilitate information interoperability and retrieval. Within a single industrial domain, there exist many taxonomies that are intended for different applications. Industry specific taxonomies often represent the vocabularies that are commonly used by the practitioners. Their jobs are multi-faceted, which include checking for code and regulatory compliance. As such, it will be much desirable if industry practitioners are able to easily locate and browse regulations of interest. In practice, multiple sources of government regulations exist and they are often organized and classified by the needs of the issuing agencies that enforce them rather than the needs of the communities that use them. One way to bridge these two distinct needs is to develop methods and tools that enable practitioners to browse and retrieve government regulations using their own terms and vocabularies, for example, via existing industry taxonomies. The mapping from a single taxonomy to a single regulation is a trivial keyword matching task. We examine a relatedness analysis approach for mapping a single taxonomy to multiple regulations. We then present an approach for mapping multiple taxonomies to a single regulation by measuring the relatedness of concepts. Cosine similarity, Jaccard coefficient and market basket analysis are used to measure the semantic relatedness between concepts from two different taxonomies. Preliminary evaluations of the three relatedness analysis measures are performed using examples from the civil engineering and building industry. These examples illustrate the potential benefits of regulatory usage from the mapping between various taxonomies and regulations.

Keywords Taxonomy interoperability, Regulation retrieval, Relatedness analysis, Domain specific ontology mapping

1 Introduction

Regulations are an important asset to the society. They extend the laws governing the country and provide standards, guidelines and requirements to corporate and the general public. Ideally regulations should be readily accessible and retrievable by interested individuals. Although well organized and hierarchically structured into sections and subsections, the sheer volume and complexity of regulations make any attempt to retrieve and to understand the relevant information a daunting task. To aid understanding of regulations, much prior work has been devoted to the analysis of regulations (Lau 2004, Lau et al. 2005), compliance guidance for regulations (Kerrigan 2003, Kerrigan and Law 2003), and abstraction and retrieval of case law (Al-Kofahi et al. 2001, Bench-Capon 1991, Brüninghaus and Ashley 2001, Moens et al. 1997, Schweighofer et al. 2001, Thompson 2001). However, efforts on developing methodologies and tools that facilitate the browsing and retrieval of regulations by industry practitioners according to their familiar terminology and vocabularies are relatively lacking.

Increasingly, taxonomies are being developed and used by industry practitioners to facilitate information interoperability and retrieval. Industry or application specific taxonomies represent the vocabularies that are commonly used by the practitioners. Interoperability is important because it allows practitioners to access, relate and combine information from multiple, heterogeneous sources and therefore increases the value of information. The lack of interoperability and integration poses significant economic costs to the engineering industries (Brunnermeier and Martin 2002, Gallaher et al. 2004). By capturing and representing the semantics of domain specific information in a formal and computer interpretable form, taxonomies have the potential to enable interoperability and to facilitate information retrieval. In practice, even within a

single industrial domain, there exist many taxonomies that are intended for different applications. As pointed out by Ray (2002), building a single unifying ontology for an entire domain is neither practical nor efficient. Instead, communities that need to exchange information frequently tend to develop their own ontologies. Therefore, even within an industry, multiple taxonomies and ontology standards exist. For instance, the architectural, engineering and construction (AEC) community has developed several terminology classifications and data model standards to describe the semantics of building models. Even though these standards are all targeted towards the same user group, the structures, vocabularies and coverage differ depending on the application.

Government regulations, on the other hand, are often organized according to the classification system of the agency that enforces them, rather than the mental models of the communities that use them (Fountain 2002). Multiple sources of regulations from different government agencies exist. There is a clear need and benefit to enable industry practitioners to browse and retrieve regulations utilizing their own classification models and vocabularies. 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, for example, via existing industry taxonomies. Industry practitioners are usually more familiar with the terminology and classification system represented in industry taxonomies than the agency's organization system for regulations. To browse through regulations and to locate compliance requirements, adhering to an existing taxonomy that the users are familiar with minimizes learning of new classification and vocabularies. Their mental models may be better represented using existing taxonomies rather than agency's classification for regulations.

In this paper, we present a systematic approach to map regulations to domain specific taxonomies, with the objective of facilitating the retrieval of relevant regulations. Section 2 briefly discusses the typical features of industry taxonomies and regulations considered in this study. Sections 3 and 4 review the techniques for mapping a single taxonomy to one or multiple regulations (Cheng et al. 2007, Cheng et al. 2008). Linking one taxonomy to one regulation is a trivial keyword extraction and latching task. Extending one taxonomy to multiple regulations requires clustering of relevant sections from different regulations. For this task, we reuse the relatedness analysis core previously developed to compute relevancy between regulation sections (Lau 2004). Section 5 discusses the needs and the challenges of mapping a single regulation to multiple taxonomies. The approach is to cluster relevant concepts from different taxonomies using a regulatory corpus to discover the relevancies between concepts. Three methodologies are investigated to cluster relevant concepts from different taxonomies in order to compute relevancy between those concepts. Cosine similarity and Jaccard coefficient, two vector-based similarity measures commonly used in the field of information retrieval are adopted to compute semantic relatedness between concepts from different taxonomies. The market basket model, a popular technique in data mining, is modified as another relatedness analysis measure for mapping of concepts. We explore the hierarchical structures of the regulation and the taxonomies to compute the relatedness scores. The methodologies to evaluate the similarity results and the comparison of other ontology mapping approaches are also discussed. Section 6 summarizes the results and contributions of this work. The natural next step, mapping multiple regulations to multiple taxonomies, is proposed as a future task.

2 Domain Specific Taxonomy and Regulation Corpus

Terms such as taxonomy, ontology, and classification have been used to describe conceptualization schemes for managing knowledge bases, supporting interoperability in Semantic Web, and facilitating integration across systems and applications. As noted by Gruber (1995), ontology is an "explicit specification of a conceptualization," which is "the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold among them." Taxonomy, on the other hand, is a hierarchical classification structure in which the descendants in the tree structure inherit or share some common properties held by their ancestors. For some, a taxonomy is but one example of an ontology. For others, the two terms are often synonymous and interchangeable. For most industrial ontology and taxonomy standards or classification systems, they are similar in that both describe concepts and entities, which are organized in a subclass hierarchy through the "is-a" relationship. As a result, domain specific concepts and the hierarchical relationships between those concepts in an industrial ontology standard or classification system can often be extracted to construct a domain specific taxonomy. Thus, within the context of this paper, an ontology standard or a classification system is treated as a taxonomy with an explicit hierarchical classification structure of concepts and entities.

We work with taxonomies and regulatory corpus from both the building industry and the environmental protection industry (Kerrigan 2003, Kerrigan and Law 2003, Lau 2004, Lau et al. 2005). To illustrate their organization and structure, we present briefly here 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 models, such as the CIMsteel Integration Standards (CIS/2) for the steel building and fabrication industry (Crowley and Watson 2000), the Industry Foundation Classes (IFC) initiated by the CAD vendors for design description of building components (International Alliance for Interoperability 1997), and the OmniClass construction classification system (OmniClass) for the construction specification, materials and product components (Construction Specifications Institute 2006).

Figures 1 and 2 show excerpted examples of the *OmniClass* and *IfcXML* standard respectively. 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 11 11". For the IfcXML, the Industry Foundation Class objects are expressed in an XML structure that defines the hierarchical relationship between elements and entities. Preprocessing is necessary to extract the terminologies and concepts from the ontology standards for subsequent analyses and usages.

Regulations are voluminous and cover a broad range of scopes and topics. Increasingly, regulatory documents are available online and are organized in XML structure. The International Building Code (IBC) (International Conference of Building Officials 2006), which represents the code of practice in the building industry, is employed as one of the regulatory document corpora. Figure 3 shows an IBC section and its representation in XML structure. One notable feature of regulations is that they are typically organized into sections and subsections, each

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

Fig. 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>
```

Fig. 2 Excerpt from the schema of Industry Foundation Classes, IfcXML

[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>
```

Fig. 3 An IBC section and its representation in XML structure

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 e-government services (Lau 2004, Lau et al. 2005).

3 One Taxonomy to One Regulation

Mapping one taxonomy to one regulation is a simple linguistic keyword latching task. To latch a keyword to a document, the keyword is automatically stemmed and matched with every single word or phrase of the stemmed text in the document using boolean matching. The keyword is then tagged to the document if the keyword is matched to the text. There are many commercial tools available to latch keywords from documents into a taxonomy. These tools usually import a taxonomy file and produce a hyperlinked user interface to help users locate the documents of interest. As noted earlier, an industry taxonomy is hierarchically organized as a classification tree which is generally less than 10 levels deep. Node labels in the taxonomy tree are treated as concept keywords, and they are mapped to sections in the regulation where they appear.

For mapping purposes, the concept terms from the two ontology standards are preprocessed to eliminate supplementary information such as the IDs in the OmniClass, the type names and attributes in the IfcXML, and the duplicated terms. Concept terms are then tokenized and stemmed before latching to the regulation. As regulations tend to be voluminous, we use a section or subsection (i.e. any numbered and titled section) as a unit of interest. Figure 4 shows the International Building Codes (IBC) displayed here in XML format and latched with the OmniClass concepts. With the concept terms of a taxonomy linked to the regulation document, users can easily traverse the taxonomy and browse relevant sections of the regulation.

4 One Taxonomy to Multiple Regulations

The mapping from one taxonomy to multiple regulations leads to a classic problem of information overload. It results in a Google-like user interface for each taxonomy node, where sections from different regulations are displayed. For example, a user might want to browse through state regulations governing

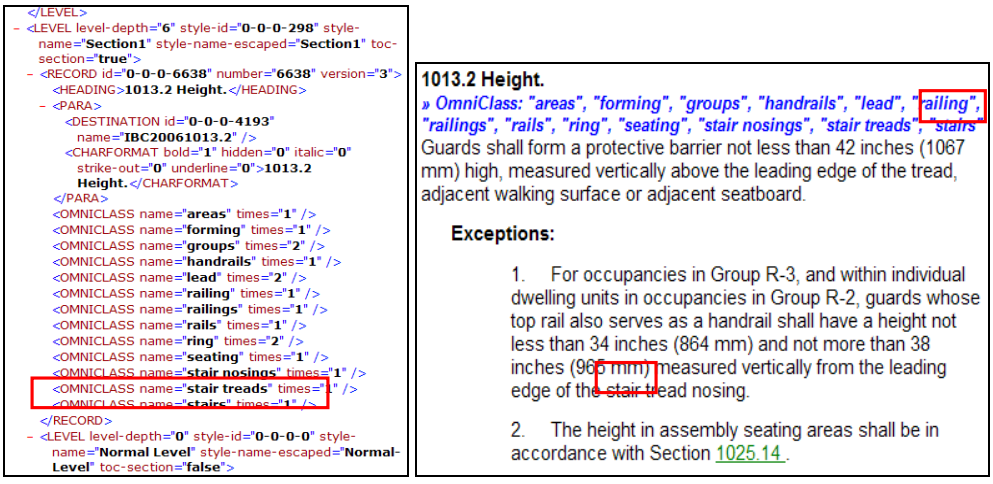

Fig. 4 Regulation in XML format latched with taxonomy concepts (left) and users' view of the modified regulation (right)

chlorine levels in drinking water. If the user search the drinking water regulations in Alabama and Arizona using the keyword “chlorine”, over 30 sections in each regulation would be located. The relevancy of these 60 regulation sections to chlorine levels is not identified. The user would quickly become frustrated with information overload. For web content, the lack of document structure poses a major challenge to search engines when computing relevancy. Therefore, intelligent retrieval and presentation of web results become a key issue for search on the Internet (Bonnel et al. 2006). Fortunately, regulatory documents are much more organized than web content, and we propose to solve the problem of information overload by clustering relevant sections from different regulations and pivoting on one regulation that the user is most familiar with. For instance, an engineer from Montgomery might be familiar with Alabama state code, but not Arizona state code. Nonetheless, if the engineer needs to design a water distribution system that provides water to Phoenix from lakes near Montgomery, searching and understanding of both state regulations would be required for compliance checking. In this case, finding the relevant Arizona regulations on chlorine levels might be a difficult job. As the engineer is more familiar with Alabama code, we believe that it is beneficial to map the taxonomy to Alabama code first, and then branch out to suggest related sections from the Arizona code. In general, focusing on one regulation as the basis for locating relevant sections and allowing users to switch the focus to other regulations significantly reduce information overload.

The discussion above poses two major challenges towards developing such a system: a suitable user interface and a methodology for determining relevant regulations. Figure 5 shows a user interface that demonstrates a scenario of locating related sections from the two state regulations. After traversing down the taxonomy tree to the concept “chlorine,” users are shown a list of matched sections from the Alabama regulation. As discussed in the previous section, matching sections to taxonomy concept is a simple keyword latching task. Selecting Section 335.7.6.15 of the Alabama 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 or related sections from other regulations as needed. As such, the user determines the amount of information he or she is exposed with in a structured retrieval model.

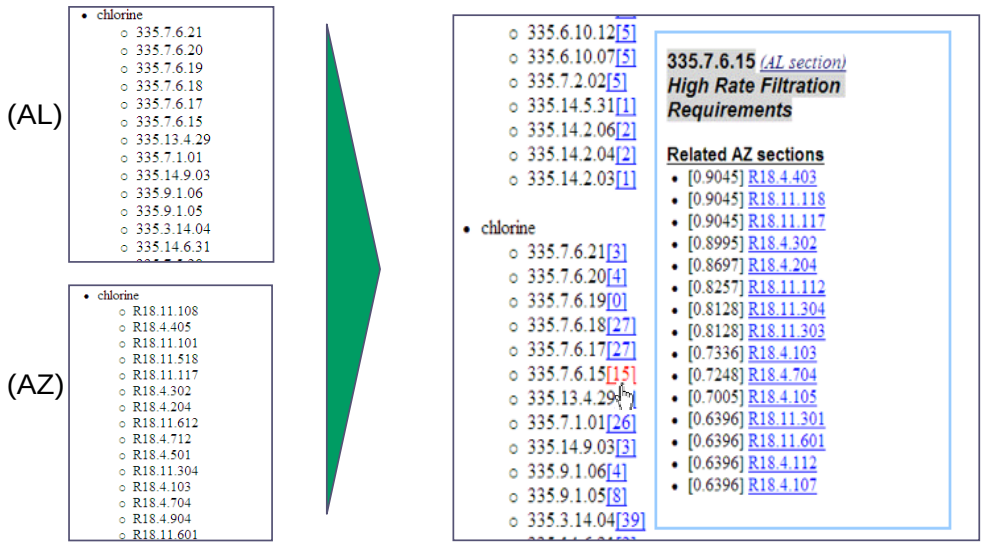

Fig. 5 Arizona regulation sections pivoting on Alabama regulations

To discover related sections from different regulations and determining their relevancies, we reuse the relatedness analysis core previously developed (Lau 2004, Lau et al. 2005), which compares sections from different regulations based on shared features using a cosine similarity measure (Larsen and Aone 1999, Salton 1989). The goal is to identify the most strongly related sections using not only a traditional term match but also a combination of feature matches, and not only content comparison but also structural analysis. Regulations are first compared based on conceptual information as well as domain knowledge through a combination of feature matching. Regulations also possess specific structures, such as a tree hierarchy of sections and the referential structure. These structures provide useful information for locating related sections and are, therefore, leveraged in the similarity analysis as well. The methodology and the relatedness analysis approach for finding similar sections from different regulations have been discussed in detail elsewhere (Lau 2004).

5 Multiple Taxonomies to One Regulation

As mentioned, multiple taxonomies exist within a single industry domain. Most industry practitioners are familiar with the vocabularies of at least one taxonomy; but, they frequently need to deal with other, possibly unfamiliar, taxonomies (Begley et al 2005, Lipman 2006). Mapping a regulation with a single taxonomy limits the usability of the system. We thus attempt to map multiple taxonomies to a regulation.

Traversing a regulation and finding relevant and related sections using multiple taxonomy trees is a challenging task. One solution is to merge the taxonomies into a new and unified taxonomy. There have been much research efforts on ontology merging (de Bruijn et al. 2004, Noy 2003, Stumme and Maedche 2001). The merged ontology which unifies and replaces the original ontologies can be used for data mediation and interoperability but not as front-end representation format. Since users would need to learn the structure and terminology of the newly merged ontology in order to browse the regulations, this would defeat the intent of using existing and thus familiar taxonomies to help search for relevant regulations. Another solution is to first map one taxonomy to the regulations, and then relate other taxonomies to the mapped taxonomy. Using the same argument from Section 4, focusing on one taxonomy that a user is familiar with is a good starting point to traverse regulations. Once the user reaches a taxonomy concept of interest, related concepts and entities from other taxonomies can be suggested and the user can shift the focus from one taxonomy to another.

Figure 6 illustrates the proposed approach using the OminClass and IFC taxonomies, and the International Building Code (IBC) regulation (International Conference of Building Officials 2006). The OmniClass is altered from its original representation to display a widget upon mouse-over that includes an ordered list of matching IBC sections and a ranked list of relevant IFC concepts. In this scenario, the user is more familiar with the OmniClass hierarchical structure and starts browsing IBC using this taxonomy. To locate the regulation sections related to the installation of steel decking, the user traverses the term “steel decking” from the OmniClass hierarchy to find the matching IBC regulation sections as well as the relevant IFC concepts. When the user moves the cursor over the IFC concept “slab”, using the same analysis, an ordered list of IBC sections that are related to slab and a ranked list of relevant OmniClass concepts

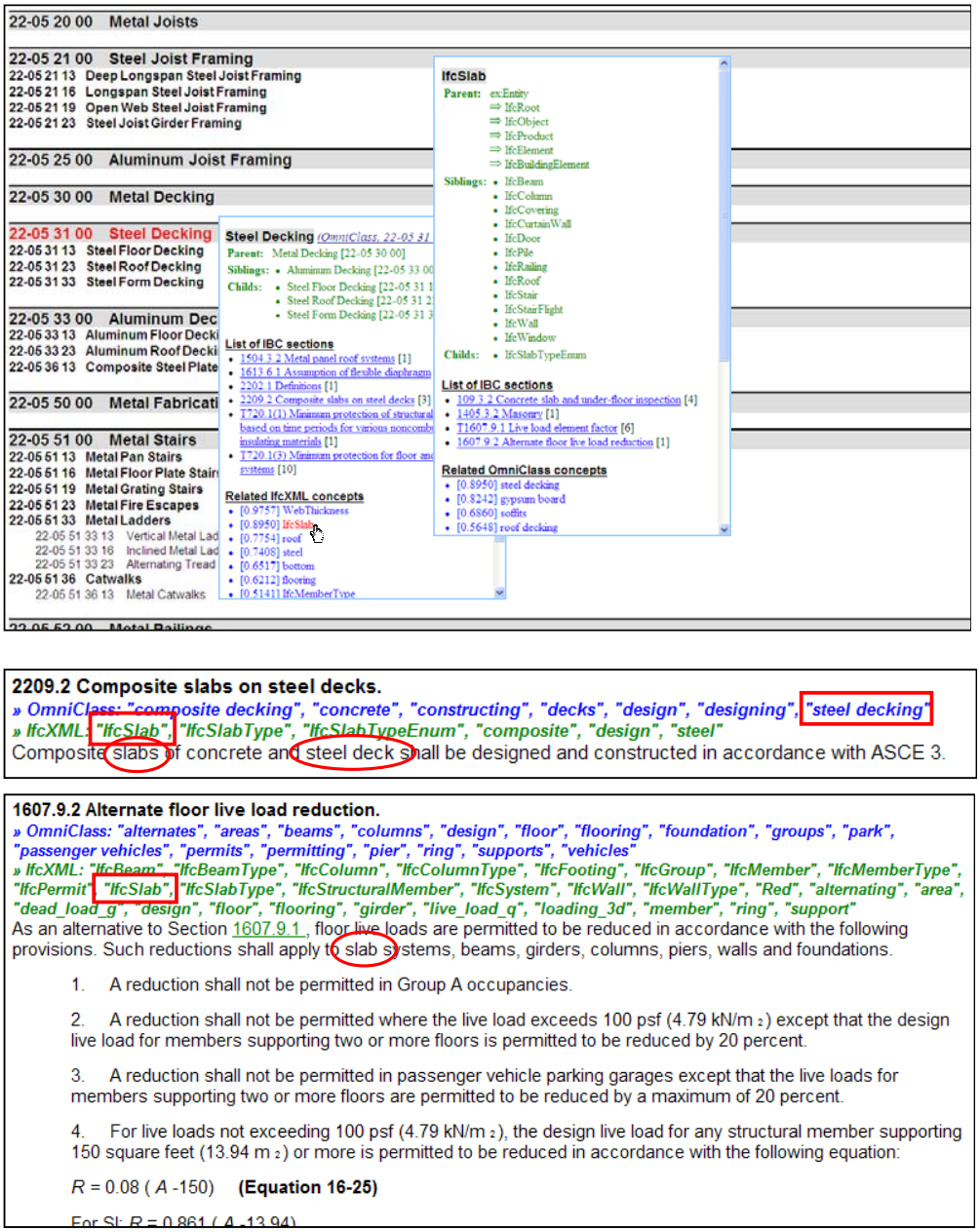

Fig. 6 Traversing the IBC using OmniClass taxonomy with relevant concepts from IFC taxonomy

are also suggested. To further illustrate the usefulness of mapping multiple taxonomies with a regulation, Figure 6 also displays two IBC sections that are related to decking design and construction. Section 2209.2, which is retrieved using the OmniClass term “steel decking”, provides the references for the composite slabs of concrete and steel decks, whereas Section 1607.9.2, which is retrieved using the IFC concept “slab”, describes load reduction for floor slab in general. Section 2209.2 is suggested because the term “steel decking” occurs in that section. However, Section 1607.9.2 cannot be retrieved using the term “steel decking” since the section uses a more generic term “slab system” instead of “decking.” If our system can relate “steel decking” to “slab,” we can present users with both sections regardless of which term the users traverse or search with.

As opposed to finding related sections from multiple regulations, the task here is to identify similar or related concepts from multiple taxonomies. Ontology mapping is an active research area and there have been many attempts to find similar concepts between ontology standards in various industry domains (Begley et al. 2005, Bicer et al. 2005, Lipman 2006). The tasks of ontology comparison and mapping are often performed manually by domain experts, who are familiar with the mapped ontologies. Such effort to identify similar concepts from multiple ontologies is often labor-intensive, non-scalable and inefficient. Research on automated or semi-automated approaches is growing in popularity, particularly for semantic web applications (Cheng et al. 2008, Li 2004, van Hage et al. 2005). It is difficult to develop mappings between two arbitrary ontologies in general. 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 earlier for mapping one taxonomy with multiple regulations, the relevancy between concepts from different taxonomies is computed using a vector comparison approach. A document corpus is used to relate concepts by considering their co-occurrence frequencies. This training corpus must be carefully selected as it represents the relevancy among concepts from different taxonomies. In this work, regulatory documents are used as the training corpus. Unlike web content, regulations are meticulously drafted and reviewed for accuracy and do not have random co-occurrences of phrases in the same section. This dramatically increases the likelihood of finding real matches. In our example, we use the same regulation document for the training corpus as well as for the targeted retrieval document.

5.1 Statistical Relatedness Analysis Measures

Consider a pool of *m* concepts and a corpus of *n* numbered and titled regulation sections. A frequency vector $\vec{c}_i$ is an *n*-by-1 vector storing the occurrence frequencies of concept *i* among the *n* sections. That is, the *k*-th element of $\vec{c}_i$ equals the number of times concept *i* is matched in section *k*. For each taxonomy, a frequency matrix *C* that aggregates the frequency vectors of all the concepts from the taxonomy can be generated. Figure 7 shows the frequency vectors for the OmniClass concept “steel decking” and for the IFC concept “slab.” In the

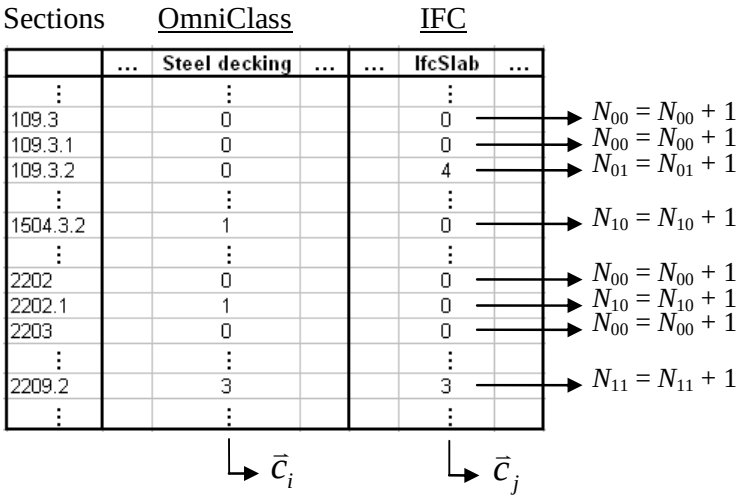

Fig. 7 Frequency vectors for OmniClass concept “steel decking” and IFC concept “IfcSlab”

frequency vector for the concept "slab," for example, the value in the row "109.3.2" is four, meaning that the concept "slab" occurs in IBC Section 109.3.2 four times. In subsequent sections, we will discuss three statistical measures to compute the similarity score among concepts. In the example shown in Figure 6, to relate the OmniClass concept "steel decking" to the IFC concept "slab", we compute their similarity score based on the defined measures. As shown in the figure, their cosine similarity score is 0.895, which ranks second among all IFC concepts that are relevant to "steel decking".

5.1.1 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 (Larsen and Aone 1999, Salton 1989). 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:

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

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

5.1.2 Jaccard Similarity Coefficient

Jaccard similarity coefficient (Roussinov and Zhao 2003, Salton 1989) 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:

$$Jaccard(i, j) = \frac{\left|\vec{c}_i \cap \vec{c}_j\right|}{\left|\vec{c}_i \cup \vec{c}_j\right|} \tag{2}$$

Jaccard similarity coefficient is a popular similarity analysis measure of term-term similarity due to its simplicity and retrieval effectiveness (Kim and Choi 1999). 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. These values can be computed by simply accumulating the number of times the corresponding matched or unmatched concepts occur. For instance, Figure 7 illustrates the calculations of N_{11}, N_{10}, N_{01} and N_{00} for the concepts "steel decking" and "slab". The similarity between both concepts is then computed as

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

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

5.1.3 Market Basket Model

The market basket model is a probabilistic data-mining technique to find item-item correlation (Hastie et al. 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 document corpus of *n* sections. Using the same notation as earlier, 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

$$\Pr(j) = \frac{N_{11} + N_{01}}{n} \tag{4}$$

which represents the individual probability of matching concept *j* to a section over the entire corpus. The confidence of the association rule $i \rightarrow j$ is

$$Conf(i \rightarrow j) = \frac{N_{11}}{N_{11} + N_{10}} \tag{5}$$

which represents the conditional probability that concept *j* is matched to a section given concept *i* is matched to that section. The forward similarity of the concept *i* and *j*, which is the interest of the association rule $i \rightarrow j$ without absolute notation, is expressed as

$$Sim(i, j) = \frac{N_{11}}{N_{11} + N_{10}} - \frac{N_{11} + N_{01}}{n} \tag{6}$$

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.

The market basket model can potentially discover the relationship that is strong in one direction but weak in another. Unlike cosine similarity and Jaccard similarity coefficient, similarity scores calculated by market basket model depend on the direction of consideration. In other words, the relatedness of concept *i* to concept *j* may be different from relatedness of concept *j* to concept *i*. The backward similarity of the concepts *i* and *j*, which is the interest of the association rule $j \rightarrow i$, is expressed as

$$Sim(j,i) = \frac{N_{11}}{N_{11} + N_{01}} - \frac{N_{11} + N_{10}}{n} \quad (7)$$

The first term on the right hand side is the conditional probability that concept *i* is matched to a section given concept *j* is matched to the section. The second term is the individual probability of matching concept *i* to a section in the whole corpus.

To illustrate this similarity asymmetry for market basket model, Table 1 shows the forward and backward similarity scores of the example concept mappings between the OmniClass and IFC taxonomies. For instance, the OmniClass concept "roof decking" and the IFC concept "slab" shows a high forward similarity score but a low backward similarity score. This implies that a section matching the concept "roof decking" will likely match the concept "slab," but not vice versa. This similarity asymmetry could be explained by their subclass is-a relationship. Roof decking is one of the many kinds of slabs in the building industry. Regulation sections describing "roof decking" are likely related to the concept "slab," but sections describing "slab" may be related to other kinds of slabs other than roof decking. The asymmetry of similarity scores may provide additional information on the types of relationship between concepts. In the later sections, when using the market basket model, the final similarity score is taken as the maximum of the forward and the backward similarity scores.

Table 2 summarizes the similarities and differences among the three measures, namely the cosine similarity, Jaccard similarity coefficient and the market basket model. All are statistical analysis measures that leverage the co-occurrence frequencies of taxonomy concepts in the regulatory corpus.

Table 1 Asymmetrical similarity scores for market basket model

OmniClass concept *i*	IFC concept *j*	$Sim(i, j)$	$Sim(j, i)$
curtain walls	IfcCurtainWall	0.992849	0.992849
sound and signal devices	IfcSwitchingDeviceType	0.998808	0.998808
roof decking	IfcSlab	0.802344	0.370313
speakers	IfcAlarmType	0.883194	0.018024
gypsum board	IfcWallType	0.568832	0.029939
concrete	IfcSlab	0.119548	0.427615

Table 2 Similarities and differences among the three statistical analysis measures

	Cosine Similarity	**Jaccard Similarity**	**Market basket Model**
Non-Euclidean	Yes	Yes	Yes
Vector-based	Yes	Yes	No
Underlying methodology	Vector space model	Set theory (Intersections and unions)	Probability theory and association rule
Symmetrical forward and backward scores	Yes	Yes	No
Range of scores	[0, 1]	[0, 1]	[-1, 1)
Usage	Often used as the baseline metric	Computationally effective	To discover potential item-item correlation

5.2 Leveraging Regulation Hierarchy Structural Information

Many related concepts can be discovered by treating each numbered and titled section in a regulation as an independent document, i.e. unit of interest. Using this approach, a concept-section frequency matrix is created to compute concept co-occurrence in documents, which correspond to individual regulation sections. This approach is generally sufficient to capture most related concepts through relatedness analysis measures. However, some related concepts rarely co-occur in the same section. For instance, if two concepts contain an "is-a" relationship, such as door furniture and door hardware, they may be used in the same regulation interchangeably but in different sections.

The is-a-related concepts are also difficult to find if each regulation section is treated as if it were an independent document in relatedness analysis. The is-a-related concepts are sometimes implicit from the hierarchical structures of a regulation. For example, as shown in Figure 8, "historic buildings" is a sub-concept of "existing structures." The two is-a-related concepts are hard to discover if regulation sections are treated independently in co-occurrence analysis because the descriptions of "historic buildings" and those of "existing structures" may not appear in the same regulation sections. Instead, the sections describing "historic buildings" are usually the subsections of the sections describing "existing structures." If we consider the subsections and the parent sections in the calculation of similarity score between "historic buildings" and "existing structures," the implicit relationship between the two concepts might become more obvious. Other related concepts such as "moved structures" and "historic buildings" may not appear in the same section since they are located in different branches under the same topic. As opposed to appearing in the same section, the sections describing "moved structures" and the sections describing "historic buildings" are organized as the subsections of the same section node. The computed relatedness between "moved structures" and "historic buildings" may increase if we consider the sibling sections in the computation of similarity score. In order to extract the implicit relationship of related concepts, it may be worthwhile to consider the hierarchical structure of the regulation sections.

Regulations contain well-organized hierarchical structures with sections and subsections of specific scope or topic. There are organizational and referential structures explicitly defined in the regulation. The organizational structure of a well-organized regulation can be represented as a hierarchical tree, where each section corresponds to a discrete node. As illustrated in Figure 9, 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 scope, the parent section covers a broader scope, the sibling sections cover parallel scopes, and the child sections cover more specific scopes. Section 4 briefly discusses the usage of the regulation hierarchy structural information to find related sections from different regulation trees. The results show that the hierarchical structure in regulations helps increase prediction accuracy of related sections (Lau 2004, Lau et al. 2005). Here, regulation hierarchical structure will be leveraged to uncover semantic relationships between related concepts from different taxonomies in the same manner. In addition to the co-occurrence of concepts in individual regulation sections, our computation includes the consideration of the co-occurrence concepts in the parent, sibling and child sections.

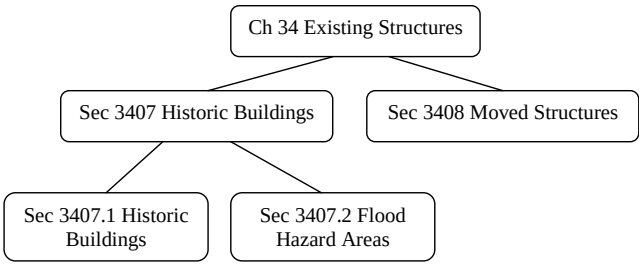

Fig. 8 Example of related but rarely co-occurring concepts

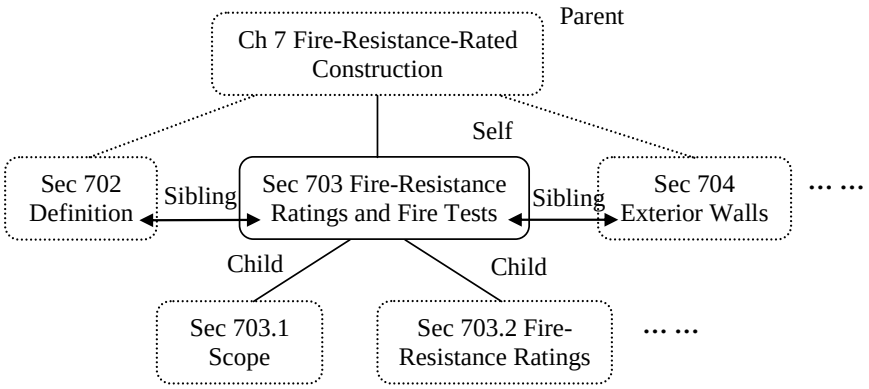

Fig. 9 Tree hierarchy of sections in regulations

The concept-section frequency matrix *C* is modified to take the parent section, sibling sections and child sections into consideration. To include the parent section, for instance, 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 numbers of occurrence in the sibling sections and the child sections are then added 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 regulation section *k*. The *k*-th element of frequency vector $\vec{c}_i$, which is the number of times concept *i* is matched to section *k*, is updated as

$$\vec{c}_i(k) := \vec{c}_i(k) + w_p \vec{c}_i(Par(k)) + w_s \sum_{u \in Sib(k)} \vec{c}_i(u) + w_c \sum_{v \in Child(k)} \vec{c}_i(v) \qquad (8)$$

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

5.3 Leveraging Taxonomy Hierarchy Structural Information

Taxonomy is a formal representation of domain information using a group of concepts and a set of relationships that are defined among those concepts. The parent, sibling and child relationships, referred as *psc*-relationships hereafter, are the fundamental relationships of a concept node in a hierarchically structured taxonomy tree. For a taxonomy concept, the parent concept represents the superclass, the sibling concepts capture the parallel entities, and the child concepts represent the subclasses. Considering the inheritance property in the taxonomy tree, the *psc* concepts for a given (self) concept from one taxonomy may be related to a concept from another taxonomy. Furthermore, their *psc* concepts may be semantically related. Therefore, consideration of the hierarchy structural information in the taxonomies may reveal or reinforce certain degree of relevancy

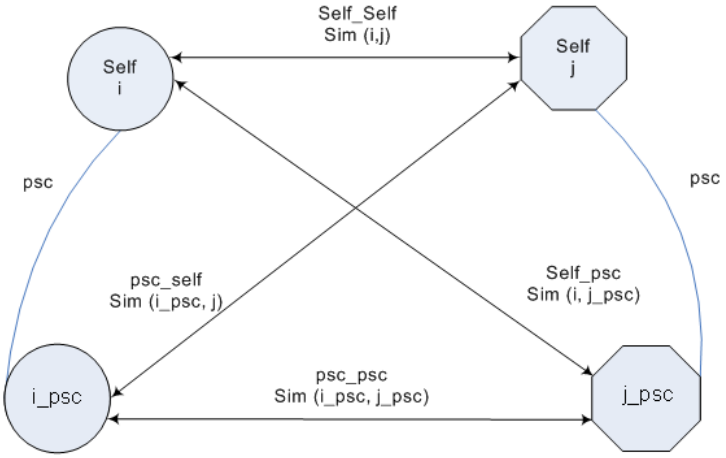

Fig. 10 Types of relationship among two concepts and their hierarchical neighbors

between two related concepts. The method to incorporate taxonomy hierarchy in relatedness analysis is discussed in the following.

Figure 10 shows the two concepts, *i* and *j*, from two taxonomies and their *psc* concepts, *i_psc* and *j_psc*, respectively. The similarity score between concepts *i* and *j* is denoted *Sim*(*i, j*). We compute the average similarity score, denoted as *Sim*(*i, j_psc*), between concept *i* and the set of *psc* concepts for concept *j*. Note that the average similarity score can be readily calculated from the basic similarity scores for each concept pair from the two taxonomies as discussed in Section 5.1. The average similarity score *Sim*(*i_psc, j*) between concept *j* and the set of *psc* concepts for concept *i*, and the average similarity score *Sim*(*i_psc, j_psc*) between the two sets of *psc* concepts can be computed in a similar fashion.

To include the *psc*-relationships in the relatedness analysis, the similarity score for each concept pair is updated by including the similarity scores from the *self* and *psc* concepts with some weighting factors:

$$\begin{cases} Sim\_new(i, j) = \alpha_0 Sim(i, j) + \alpha_1 [Sim(i, j\_psc) + Sim(i\_psc, j)]/2 \\ \qquad\qquad + \alpha_2 Sim(i\_psc, j\_psc) \\ \alpha_0 + \alpha_1 + \alpha_2 = 1 \;;\; \alpha_0 > \alpha_1 > \alpha_2 \end{cases} \tag{9}$$

where α_0, α_1 and α_2 represent the weighting factors for concept pairs (*self-self*), (*self-psc*) and (*psc-psc*), respectively. The weighting factors are chosen such that the updated similarity score is within the range between 0 and 1. Since the major focus of interest is the relatedness between concept *i* and concept *j*, the similarity score *Sim*(*i, j*) between the two self concepts should have the greatest influence to the final similarity score. Therefore, α_0 is chosen to be larger than α_1 and α_2 is the smallest among the three weighting factors.

5.4 Evaluation Results

Concepts from the OmniClass and the IFC taxonomies are extracted. For every taxonomy, each concept is assigned a consecutive number and twenty numbers are randomly generated. Twenty concepts are then randomly selected from the OmniClass and the IFC taxonomies respectively, and pairwise similarity scores are computed using the three statistical relatedness analysis measures described in Section 5.1. The results of concept mapping performed by domain experts are treated as true matches and are used to evaluate the predicted results. Root mean

square error (RMSE), precision, recall and F-measure are used as performance metrics to evaluate and compare the three measures and the use of regulation structural information as well as taxonomy structural information. The predicted results are also compared to the baseline results using a terminology-based approach and a lexicon-based approach.

Three domain experts are asked to identify the related concept pairs among the total of 400 possible pairs. A true value of one is assigned to the concept pairs that the domain expert classifies as related. All other concept pairs are assigned a true value of zero. As for the predicted results, two concepts are predicted as similar or related if the computed similarity score is larger than certain threshold score. Related concept pairs are assigned a predicted value of one whereas other pairs are assigned zero. Based on each of the three manual mappings, values of RMSE, precision, recall and F-measure are calculated for the three measures, the baseline matchers, and different regulation and taxonomy structural information inclusions. The averages of the values are then taken as the final results.

5.4.1 Root Mean Square Errors (RMSE)

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

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|true_{i,j} - predicted_{i,j}\right|} \qquad (10)$$

Figure 11 shows the results of the three measures compared using RMSE for threshold similarity scores ranging from 0.15 to 0.9. Neither regulation hierarchy structural information nor taxonomy hierarchy structural information is considered. As illustrated in Figure 11, 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 discovering related concept pairs from different taxonomies, using sections from the regulation as independent documents in the co-occurrence computation. Cosine similarity appears to be

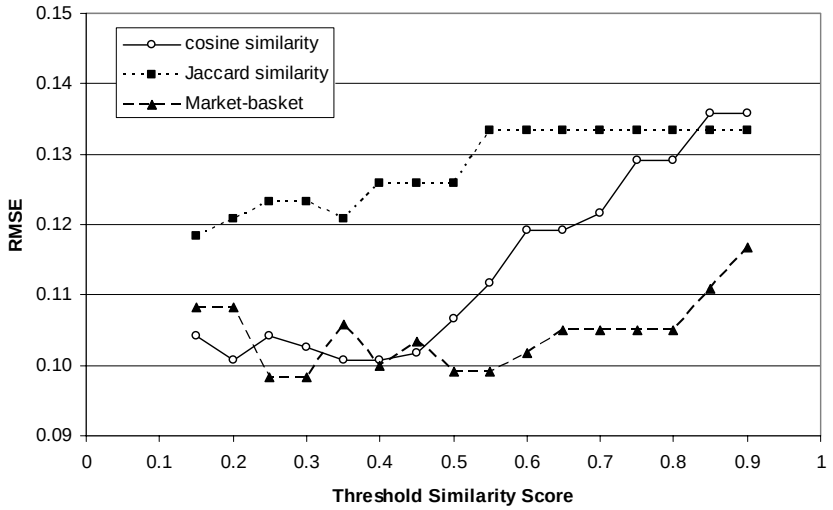

Fig. 11 Evaluation results of the three measures using RMSE

average among the three statistical analysis measures.

5.4.2 Precision, Recall and F-measure

We use precision, recall and F-measure values to compare the three similarity analysis measures and the use of regulation and taxonomy hierarchy structural information. While RMSE takes both correctness and incorrectness of prediction into consideration, precision and recall emphasize correctness only. Precision and recall evaluate the accuracy of predictions and the coverage of accurate pairs. Precision measures the fraction of predicted matches that are correct, that is, the number of true positives over the number of pairs predicted as matched. Recall measures the fraction of correct matches that are predicted, that is, the number of true positives over the number of pairs that are actually matched. They are computed as

$$Precision = \frac{\left|True\ Matches \cap Predicted\ Matches\right|}{\left|Predicted\ Matches\right|} \tag{11}$$

$$Recall = \frac{\left|True\ Matches \cap Predicted\ Matches\right|}{\left|\text{True}\ Matches\right|} \tag{12}$$

There is always a tradeoff between precision and recall. F-measure is therefore used 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

$$F - Measure = \frac{2 \cdot \left(Precision \times Recall\right)}{Precision + Recall} \tag{13}$$

Figure 12 shows the results for the three relatedness analysis measures using precision, recall and F-measure. 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 appears to be average among the three measures, consistent with the RMSE results.

As the market basket model outperforms cosine and Jaccard similarities using both RMSE and the F-measure, we will evaluate the impact of regulation hierarchy and taxonomy hierarchy using the market basket model as the similarity measure of choice. As shown in Figure 13, the effect of including regulatory structure in the analysis is inconclusive. In general, including regulation hierarchical information increases recall rate but reduces precision, as more regulatory nodes are being considered to locate related concepts. Considering neighboring nodes increases the chance to find related concepts that rarely co-occur, and thus improves the recall rate; however, including neighboring nodes also raises the likelihood to be affected by the noises of co-occurrence, and therefore decreases the precision rate. Including the parent section produces a slightly higher F-measure in most threshold scores, likely due to the fact that parent relationship is one to one which minimizes the impact on precision. Others such as sibling and child relationships, are one to many; including such relationships may reduce precision with only minor improvement in recall.

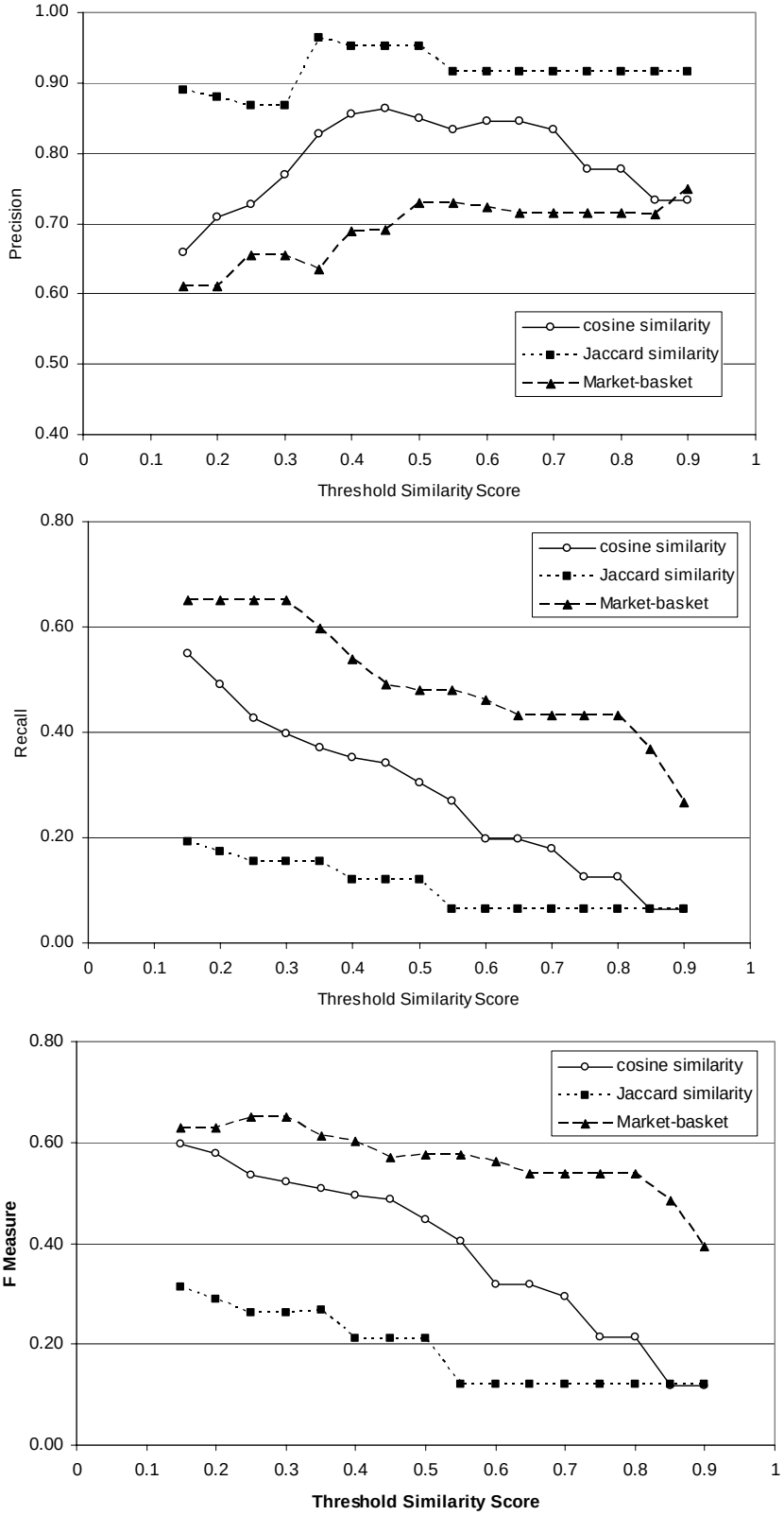

Fig. 12 Evaluation results of the three measures using F-measure

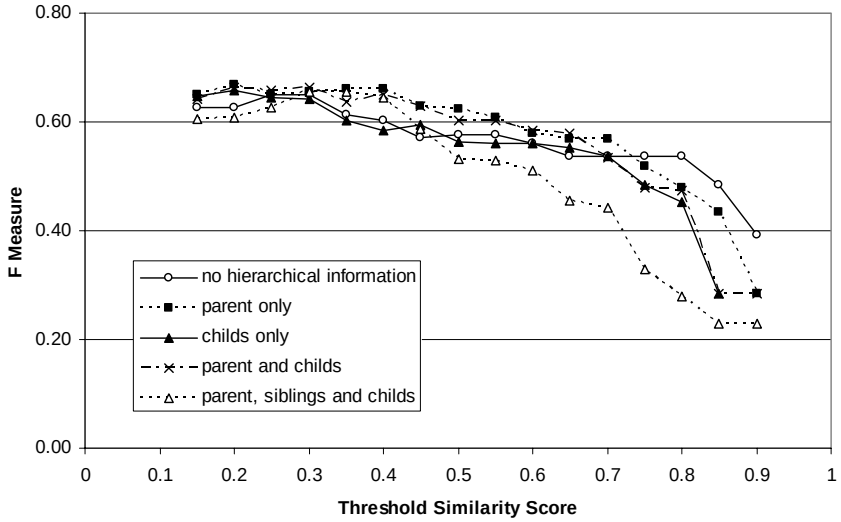

Fig. 13 Evaluation results of market basket model with regulation hierarchy inclusion using F-measure

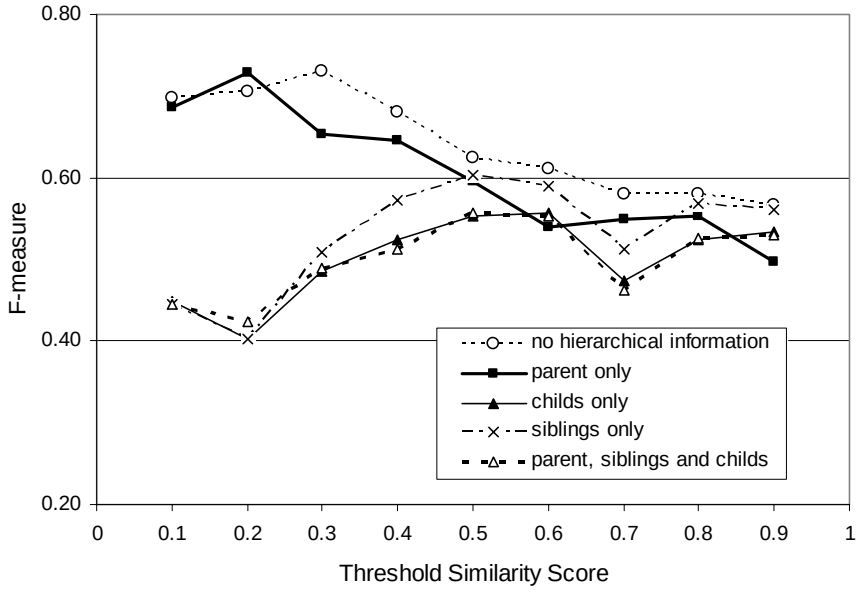

Fig. 14 Evaluation results of market basket model with taxonomy hierarchy inclusion using F-measure

To evaluate the influence of taxonomy hierarchical information in the relatedness analysis of concepts, Equation (9) is used to update the basic similarity scores with α_0, α_1 and α_2 chosen as 0.7, 0.2 and 0.1, respectively. Regulatory hierarchical information is not considered in the calculation. Figure 14 shows that in this test case, consideration of taxonomy hierarchical structure does not improve the F-measure results. In fact, although not shown here, including *psc* concepts in relatedness analysis does improve the recall rate by 20 percent in general. It should be pointed out that the OmniClass and IFC taxonomies have a relatively flat hierarchy and many concepts possess over 20 siblings and children, many of which are not strongly related to the self concepts. The increase in the number of incorrect matches when taxonomy hierarchy is considered seriously reduces the precision rate. As a result, the decrease in precision rate overweighs

the improvement in recall rate, leading to a lower F-measure value when taxonomy hierarchical information is incorporated.

5.4.3 Comparison of the Domain-based Approach with the Terminology-based Approach and Lexicon-based Approach

In addition to comparing the three measures with one another, we also evaluated our domain-based approach to a traditional lexicon-based approach and a simple terminology-based approach. Ontology mapping is an active research area due to the growing number of autonomously developed ontologies. Automated and semi-automated ontology mapping is commonly performed using rule-based (Li et al. 2000, Mitra 2003), terminology-based (Aumueller et al. 2005, Noy and Musen 2003), structure-based (Melnik et al. 2002, Milo and Zohar 1998), and lexicon-based methods (Madhavan et al. 2001, Palopoli et al. 1999). Our approach is comparable to a lexicon-based approach, where dictionary and thesaurus are used to enumerate related terms such as synonyms and homonyms. Our approach is also similar to a terminology-based approach, where the spelling of concept terms is used to identify semantic similarity linguistically. In our analysis, we use a domain specific corpus, i.e., a domain-appropriate regulation, and the preprocessed concept terms to uncover the semantic relationships among entities from heterogeneous taxonomies.

Some mappings such as (sound and signal devices, IfcSwitchingDeviceType) are quite obvious from the name of the concept terms. Although the two concept names are not textually identical, they share the term "device." The descriptive keywords in the concept name provide an alternative means to map concepts from different ontologies. To relate descriptive phrases in our baseline terminology-based matcher, we tokenize keywords in concept names and stem the tokens using Porter Stemmer (Porter 1980). As illustrated in Figure 15, each keyword token represents a row in the frequency matrix. In each frequency vector, a value of one is assigned to the rows of the keyword tokens that appear in the concept name. The semantic relatedness between concepts is determined according to the amount of keyword tokens they share. The three statistical analysis measures described in Section 5.1 are then used to compute the similarity scores.

A thesaurus is necessary to compare our approach to a lexicon-based mapping method. One of the most common thesauri is the WordNet (Miller et al. 1993), which is a well-known lexical resource for the English language. Synonyms in WordNet are interlinked by means of conceptual-semantic and

	OmniClass		IFC	
	track rail fasteners	chemical sampling and analysis of soils	IfcMechanicalFastener	IfcStructuralAnalysisModel
analys		1		1
chemic		1		
fasten	1		1	
mechan			1	
model				1
rail	1			
sampl		1		
soil		1		
structur				1
track	1			

	OmniClass		IFC	
	metal	permitting	IfcPermit	IfcSteel
allow		1	1	
alloy	1			
blade				1
brand				1
let		1	1	
license		1	1	
metal	1			
metallic	1			
permit		1	1	
steel				1
sword				1

Fig. 15 Frequency matrices using terminology-based approach (left) and lexicon-based approach (right)

lexical relations. It is one of the most widely adopted synonym sources for ontology matching techniques including CUPID (Madhavan et al. 2001), Learned Ontology Model (LOM) (Li 2004), and Version Matching Approach (VMA) (Wang et al. 2007). As shown in Figure 15, each synonym of a concept, as appeared in WordNet, represents a row in the frequency matrix.

Table 3 shows the result for comparing our domain-based ontology mapping method with a terminology-based method and a lexicon-based method using WordNet. The results show that our domain-based approach, in this case, using the regulatory document related to the taxonomies outperforms the terminology-based method as well as the lexicon-based method in terms of precision, recall and F-measure. The terminology-based matcher has a perfect precision because the concept terms that share the same keyword tokens are usually related. The terminology-based method, however, cannot identify semantically related concepts that are expressed using different terminology. Some examples of matches that are found by our domain-based matcher but not by the terminology-based method include (door hardware, IfcBuildingElementComponent), (steel decking, IfcSlab), (sound and signal devices, IfcAlarmType), etc.

Although both the lexicon-based method and our domain-based method utilize knowledge corpus as a bridge to discover semantic knowledge, the lexicon-based method results in much lower recall rate and F-measure in all cases. Since

Table 3 Precision and recall comparisons of domain-based ontology mapping approach to terminology-based and lexicon-based approaches (P: Precision, R: Recall, F: F-measure)

Similarity score threshold	Approaches	Cosine Similarity			Jaccard Similarity			Market basket Model		
		P	R	F	P	R	F	P	R	F
0.2	Lexicon-based	0.50	0.03	0.06	0.00	0.00	n/a	0.00	0.00	n/a
	Terminology-based	1.00	0.19	0.32	1.00	0.19	0.32	1.00	0.19	0.32
	Domain-based	0.79	0.53	0.63	0.91	0.17	0.29	0.70	0.71	0.70
0.3	Lexicon-based	0.50	0.03	0.06	1.00	0.03	0.06	0.50	0.03	0.06
	Terminology-based	1.00	0.19	0.32	1.00	0.14	0.25	1.00	0.19	0.32
	Domain-based	0.83	0.41	0.55	0.90	0.15	0.26	0.75	0.71	0.73
0.4	Lexicon-based	1.00	0.03	0.06	1.00	0.03	0.06	0.50	0.03	0.06
	Terminology-based	1.00	0.19	0.32	1.00	0.12	0.21	1.00	0.19	0.32
	Domain-based	0.91	0.36	0.52	1.00	0.12	0.21	0.80	0.59	0.68
0.5	Lexicon-based	1.00	0.03	0.06	1.00	0.03	0.06	1.00	0.03	0.06
	Terminology-based	1.00	0.14	0.25	1.00	0.12	0.21	1.00	0.14	0.25
	Domain-based	0.90	0.31	0.46	1.00	0.11	0.20	0.81	0.51	0.63
0.6	Lexicon-based	1.00	0.03	0.06	1.00	0.03	0.06	1.00	0.03	0.06
	Terminology-based	1.00	0.12	0.21	1.00	0.03	0.06	1.00	0.14	0.25
	Domain-based	0.92	0.20	0.33	1.00	0.07	0.13	0.81	0.49	0.61

many concepts have different meanings when used in different domains, the synonyms and definitions of identical concepts could be very different because of contexts. WordNet is a generic linguistic thesaurus rather than a domain specific document corpus. As a result, it contains little and imprecise information of the terminology used by the OmniClass and IFC taxonomies. The result shows that domain-related corpora, such as regulations and technical specifications, are useful in discovering the semantic relationships across multiple ontologies.

6 Conclusion & Future Tasks

In this paper, we consider the use of industry specific taxonomies to facilitate the retrieval of regulation sections pertinent to the subject of interests. Regulatory documents are written by government agencies who organize the materials to suit their own needs and intents, which may not fulfill the needs of the communities that use them. From industry practitioners' standpoint, the original hierarchy might not be the easiest retrieval model for regulations. In this work, we propose a systemic approach to map industry specific taxonomies to regulations in order to increase usability of regulations by industry practitioners.

This paper began by briefly describing the linking of a taxonomy to a regulation by latching the concept terms to the regulation sections. For the retrieval of related sections from multiple regulations using a single taxonomy, a relatedness analysis approach is suggested to compute relevancy between those sections. To compare sections from different regulations, cosine similarity is used to measure the relatedness and, regulation hierarchical information is leveraged to enhance the analysis. For the mapping of multiple domain specific taxonomies to a regulation, we propose to find related concept terms between the taxonomies. Specificially, regulatory document, a domain specific corpus, is employed to perform the relatedness analysis on the concept terms. For the relatedness comparison of concept pairs, three similarity measures are tested, and regulation hierarchical structures as well as taxonomy hierarchical structures are considered in the computation of similarity scores. Among the three measures, we have shown, using the taxonomies and regulation corpus employed in this study, that the market basket model performs the best in terms of RMSE and F-measure, which is a combination of precision and recall. When comparing with the terminology-based approach and the lexicon-based approach, our domain-based approach, utilizing domain related document as the training corpus, generally results in higher precision, recall and F-measure.

In summary, the 1-1, 1-n, and n-1 mapping between taxonomies and regulations have been demonstrated. We plan to implement an n-n concept-section mapping in the future, by combining the techniques of concept comparisons and section comparisons. Furthermore, we plan to engage potential users to help perform formal evaluations of the similarity measures and the usability of the system. To improve usability, a better user interface is much needed, and we plan to investigate the need to implement or adopt such visualization tool. An ideal user interface should facilitate access to the mapping of multiple taxonomies and the browsing of regulations by industry practitioners, rule makers and domain experts.

7 Acknowledgements

The authors would like to acknowledge the supports by the US National Science Foundation, Grant No. CMS-0601167 and IIS-0811460, the Center for Integrated Facility Engineering (CIFE) at Stanford University and the Enterprise Systems Group at the National Institute of Standards and Technology (NIST). The authors would like to thank the International Code Council (ICC) for providing the XML version of the International Building Code 2006. Any opinions and findings are those of the authors, and do not necessarily reflect the views of NSF, CIFE, NIST or ICC. No approval or endorsement of any commercial product by NIST, NSF, ICC or Stanford University is intended or implied.

References

Al-Kofahi K, Tyrrell A, Vachher A et al (2001) 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

Aumueller D, Do H H, Massmann S et al (2005) Schema and ontology matching with COMA++. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp 906-908

Begley E F, Palmer M E, Reed K A (2005) Semantic Mapping Between IAI ifcXML and FIATECH AEX Models for Centrifugal Pumps. Technical Report NISTIR 7223, NIST

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

Bicer V, Laleci G, Dogac A et al (2005) Artemis Message Exchange Framework: Semantic Interoperability of Exchanged Messages in the Healthcare Domain. ACM Sigmod Record 34(3)

Bonnel N, Lemaire V, Cotarmanac'h A et al (2006) 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

Brüninghaus S, Ashley K D (2001) 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

Brunnermeier S B, Martin S A (2002) Interoperability Costs in the US Automotive Supply Chain. Supply Chain Management: An International Journal 7(2): 71-82

Cheng C P, Lau G T, Law K H (2007) Mapping Regulations to Industry-Specific Taxonomies. In Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL), Stanford, CA, USA

Cheng C P, Lau G T, Pan J et al (2008) Domain-Specific Ontology Mapping by Corpus-Based Semantic Similarity. In Proceedings of 2008 NSF CMMI Engineering Research and Innovation Conference, Knoxville, Tennessee, USA

Cheng C P, Pan J, Lau G T et al (2008) Relating Taxonomies with Regulations. In Proceedings of the 9th Annual International Conference on Digital Government Research (dg.o2008), Montreal, Canada

Construction Specifications Institute (2006) OmniClass Construction Classification System, Edition 1.0.

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

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

Fountain J E (2002) 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

Gallaher M P, O'Connor A C, Dettbarn J L et al (2004) Cost Analysis of Inadequate Inoperability in the Capital Facilities Industry. Technical Report, GCR 04-867, National Institute of Standards and Technology (NIST)

Gruber T R (1995) Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies 43(5/6): 907-928

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

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

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

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

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

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

Larsen B, Aone C (1999) 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

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

Lau G, Law K, Wiederhold G (2005) 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

Li J (2004) 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

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

Lipman R (2006) Mapping Between the CIMsteel Integration Standards (CIS/2) and Industry Foundation Classes (IFC) Product Model for Structural Steel. In Proceedings of the 11th International Conference on Computing in Civil and Building Engineering, (ICCCBE XI), Montreal, Canada, pp 3087-3096

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

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

Miller G A, Beckwith R, Fellbaun C et al (1993) Five Papers on WordNet. Technical Report, Cognitive Science Laboratory, Princeton, NJ

Milo T, Zohar S (1998) 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

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

Moens M-F, Uyttendaele C, Dumortier J (1997) 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

Noy N F (2003) Tools for Mapping and Merging Ontologies. In: Staab S, Stude R (eds) Handbook on Ontologies. Springer-Verlag, pp 365-384

Noy N F, Musen M A (2003) The PROMPT suite: interactive tools for ontology merging and mapping. International Journal of Human-Computer Studies 59(6): 983-1024

Palopoli L, Sacca D, Terracina G et al (1999) 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

Porter M F (1980) An Algorithm for Suffix Stripping. Program 14(3): 130-137

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

Roussinov D, Zhao J L (2003) Automatic Discovery of Similarity Relationships Through Web Mining. Decision Support Systems 25: 149-166

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

Schweighofer E, Rauber A, Dittenbach M (2001) 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

Stumme G, Maedche A (2001) 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

Thompson P (2001) 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

van Hage W, Katrenko S, Schreiber G (2005) A Method to Combine Linguistic Ontology-Mapping Techniques. In Proceedings of the Fourth International Semantic Web Conference (ISWC)

Wang H, Akinci B, Garrett J H (2007) Formalism for Detecting Version Differences in Data Models. Journal of Computing in Civil Engineering 21(5): 321-330