Improving Access to and Understanding of Regulations through Taxonomies

ABSTRACT

Industrial taxonomies have the potential to automate information retrieval, facilitate interoperability and, most importantly, improve decision making -- decisions that must comply with existing government regulations and codes of practice. However, it is difficult to find those regulations and codes most relevant to a particular decision, even though they are now in digital form, and often available online. The focus of this work is to map regulations and codes to existing industry-specific taxonomies that would improve their access and retrieval and facilitate their integration with application programs.

Keyword matching is a commonly used technique for mapping from a single taxonomy to a single regulation. In this paper, we examine techniques to address two other mapping problems: from a single taxonomy to multiple regulations and from multiple taxonomies to a single regulation. Those techniques - cosine similarity, Jaccard coefficient, and market-basket analysis - provide metrics for measuring the similarity between concepts from different taxonomies. We discuss these metrics and provide evaluations using examples from the building industry. These examples illustrate 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 extend the laws governing the country with specific guidance for corporate and public actions. Therefore, regulations are an important asset to society and they should be readily retrievable by interested individuals or businesses. For instance, manufacturing companies design and fabricate thousands of products for use by the public. These products and the processes involved in inventing and producing them are subjected to comply with numerous Federal and State regulations. The complexity, diversity, and volume of these regulations make it difficult for companies to know when they are in compliance and for the public to be confident in the safety and performance of the products. Since these regulations have the force of law, however, it is important that companies and citizens be able to locate, understand, and comply with them.

As noted in the Washington Post, "deciphering and complying with federal regulations is a legal and paperwork nightmare for many businesses (Skrzycki, 2000)." This burden has been recognized and targeted by legislation to create a digital government infrastructure that would make such regulations available in digital formats. The Office of Management and Budget's (OMB) efforts to provide such an infrastructure - formerly First-Gov and currently E-Gov – have been organized around four key portfolios: Government to Citizen, Government to Business, Government to Government, and Internal Efficiency and Effectiveness. Aligned with these portfolios, the OMB and other Federal agencies have launched various E-GOV initiatives to provide "high-quality and well-managed solutions for tax filing, federal rulemaking and e-

training among others (Office of Management and Budget (OMB), 2008)." One initiative came from the Small Business Administration (SBA) (with participation from many other federal agencies), which launched a program to build a "one-stop" portal to assist small businesses to comply with regulations (Small Business Administration, 2002).

Most government agencies and organizations now distribute regulations and codes on the web using a variety of portals.[1] Presently, the majority of these online portals provide digital information in either PDF or HTML format. As such, they are designed primarily for displaying regulatory information for experienced users who already know the relevant regulations; they cannot be used directly with other software applications. A next generation IT framework that facilitates the retrieval of regulations and allows integration of regulations with applications will empower small businesses and citizens with relevant policy and compliance information in electronic formats that support both computer and human decision-making. There is significant societal impact of such framework.

Our research aims to establish methodologies to aid in the development of this framework (Cheng, Lau, Law, Pan, & Jones, 2008; Lau, 2004, 2005). We use advanced IT modeling techniques and analysis tools to enhance the availability, understanding, and usage of those digital government regulations most relevant to both citizens and businesses. Specifically, we use ontologies of domain knowledge and advanced textual analysis techniques to transition the current state of online forms and scattered documentations to the next generation digital government portal where interactive systems and organized regulations are available. This enables us to map the domain knowledge, in the form of industry-specific taxonomies, to relevant regulatory sections. To map between taxonomies and regulations, we generate

frequency matrices for the taxonomy concepts and the regulatory sections and perform relatedness comparison using statistical analysis techniques. Structural information of the regulations is also considered to refine the relatedness scores. In this paper, we present the results of our research which we believe illustrate the potential benefits for improving access to regulations via taxonomies.

This paper is organized as follows. We summarize relevant work on regulations and taxonomies in Section 2. We introduce example regulations and taxonomies from the building industry in Section 3. In Section 4, we review the keyword-matching technique used to map one taxonomy to one regulation. Since small businesses often have to comply with more than one regulation, we extend the mapping to multiple regulations in Section 5, where we use the relatedness analysis approach that compares regulation sections based on term match, as well as a combination of feature matches, content comparison and structural analysis. In Section 6, we (1) discuss the challenges of mapping multiple taxonomies to a single regulation, (2) propose three metrics to compute the similarity between concepts from multiple taxonomies, and (3) provide an evaluation of the three metrics using precision and recall measures. In Section 7, we conclude with observations based on our research to date as well as suggestions of future work.

2. LITERATURE REVIEW

Laws are an important aspect of our society. To aid understanding of the law, (Al-Kofahi, Tyrrell, Vachher, & Jackson, 2001; Bench-Capon, 1991; Brüninghaus & Ashley, 2001; Moens, Uyttendaele, & Dumortier, 1997; Schweighofer, Rauber, & Dittenbach, 2001; Thompson, 2001) have proposed numerous techniques for abstraction and retrieval of case law.

[1] See http://www.gpoaccess.gov/cfr/, http://www.regulations.gov/, etc., for examples.

(Lau, 2004, 2005) has developed approaches for analysis of regulations and (Kerrigan, 2003; Kerrigan & Law, 2003) have suggested methods for compliance guidance for regulations. 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 for a number of industry domains. Taxonomies describe concepts and entities in a subclass hierarchy through an "is-a" relationship. Since taxonomies contain well defined entities and hierarchical relationships, computers can interpret, understand, and reason about the terms and concepts described in a taxonomy. As a result, taxonomies can facilitate information interoperation and regulation retrieval. Interoperability is important because it allows practitioners – more importantly application programs - to access, relate, and combine information from multiple, heterogeneous sources. Recent studies by the National Institute of Standards and Technology (NIST) have reported that the lack of interoperability led to significant costs to the construction as well as the automotive industries (Brunnermeier & Martin, 1999; Gallaher, O'Connor, Bettbarn, & Gilday, 2004).

Ontologies, which describe the general semantics of concepts and entity relationships that are not limited to an "is-a" hierarchy, have been proposed as a way to address interoperability problems. One recent forecast estimates that "By 2010, ontologies ….will be the basis for 80 percent of application integration projects" (Jacobs & Linden, 2002). Ontologies serve as a means for information sharing because they capture the semantics of domain-specific information in a formal and computer interpretable form. Utilizing ontologies as a means to automate much of the integration process might be able to reduce cost and time significantly. We believe that they can also be used to facilitate access to government regulations.

Building a single ontology for an entire industry domain is both inefficient and impractical. Rather, small communities that need to exchange information frequently build ontologies targeted to their own users and applications (Ray, 2002). This results in multiple terminology classifications and data model structures. For instance, the architectural, engineering and construction (AEC) community has built several ontologies that describe the semantics of buildings and their components (Begley, Palmer, & Reed, 2005; Lipman, 2006). Even though these ontologies are all targeted towards the same user group, their structures, vocabularies and coverage differ depending on the application.

Government agencies, on the other hand, often use terminology and organize regulations based on their own needs, rather than the needs of the industrial communities they serve (Fountain, 2003). Both the agencies and the communities see a clear benefit of bridging these two distinct needs. One way to build such a bridge is to enable practitioners to browse and retrieve government regulations using their own terms and vocabularies - as captured in existing industry taxonomies. This would minimize the need for users to learn new vocabularies and organizational schemes. Metadata such as taxonomies and ontologies have been leveraged to facilitate locating and retrieval of government information (Moen, 2001; Prokopiadou, Papatheodorou, & Moschopoulos, 2004). These metadata, however, capture the semantics of the government information for conceptualization rather than representing domain knowledge from industry practitioners for browsing and retrieval. To bridge the needs of policy makers and the needs of industrial communities, we need methods and tools that map taxonomies to regulations. In the remainder of this paper, we describe a collection of such methods and tools.

3. TAXONOMIES AND REGULATIONS FOR THE PILOT STUDY

In this paper, we work with taxonomies and regulatory corpus from both the building industry and the environmental protection industry (Kerrigan, 2003; Kerrigan & Law, 2003; Lau, 2004, 2005). For the building industry, we use three main taxonomies that describe the semantics of building models: the CIMsteel Integration Standards (CIS/2) for the steel building and fabrication industry (Garas & Hunter, 1998), the Industry Foundation Classes (IFC) for building CAD models of building components (International Alliance for Interoperability (IAI), 2007), and the OmniClass construction classification system (OmniClass) for the construction specification, materials and product components (Construction Specifications Institute (CSI), 2006).

Figures 1 and 2 show excerpted examples of the *OmniClass* and *IfcXML* standards. Typical of ontology standards, both OmniClass and IfcXML 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 "Street and Roadway Lighting" is associated with the ID "23-80 70 14 21". For the IfcXML taxonomy, the Industry Foundation Class objects are expressed in an XML structure that defines the hierarchical relationship between elements and entities. As a result, the first task in this pilot study is to extract the object terms from the taxonomies, so that we can use them to map to regulations. We implemented parsers for this task to preprocess the two standards to eliminate irrelevant information, such as the IDs in the OmniClass, the element names, group names and type names in the IfcXML, as well as duplicated terms from both standards.

23-80 70 00 Lighting
23-80 70 11 Luminaries for Internal Lighting
23-80 70 11 11 General Luminaries, Non Directional
23-80 70 11 14 General Luminaries, Directional
23-80 70 11 14 11 Downlights
23-80 70 11 14 14 Uplights
23-80 70 11 14 17 Direct/Indirect
23-80 70 11 14 21 Spots and Tracklight Specialties
23-80 70 11 17 Specialized Lighting by Location or Use
23-80 70 11 21 Emergency Lighting
23-80 70 11 24 Fiber Optic Lighting
23-80 70 14 Luminaries for External Lighting
23-80 70 14 11 Amenity Lighting
23-80 70 14 11 11 Lighting Bollards
23-80 70 14 11 14 Post-Top Lighting
23-80 70 14 11 17 Wall or Ceiling Mounted External Lighting
23-80 70 14 11 21 Buried Uplights
23-80 70 14 14 Exterior Floodlights
23-80 70 14 17 Exterior Spotlights
23-80 70 14 21 Street and Roadway Lighting

Figure 1: Excerpt from OmniClass Construction Classification System

```
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcReinforcingBar" type="ifc:IfcReinforcingBar"
    substitutionGroup="ifc:IfcReinforcingElement" nillable="true" />
- <xs:complexType name="IfcReinforcingBar">
  - <xs:complexContent>
    - <xs:extension base="ifc:IfcReinforcingElement">
      - <xs:sequence>
          <xs:element name="NominalDiameter" type="ifc:IfcPositiveLengthMeasure" />
          <xs:element name="CrossSectionArea" type="ifc:IfcAreaMeasure" />
          <xs:element name="BarLength" type="ifc:IfcPositiveLengthMeasure" nillable="true"
            minOccurs="0" />
          <xs:element name="BarRole" type="ifc:IfcReinforcingBarRoleEnum" />
          <xs:element name="BarSurface" type="ifc:IfcReinforcingBarSurfaceEnum"
            nillable="true" minOccurs="0" />
        </xs:sequence>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcReinforcingElement" type="ifc:IfcReinforcingElement" abstract="true"
    substitutionGroup="ifc:IfcBuildingElementComponent" nillable="true" />
- <xs:complexType name="IfcReinforcingElement" abstract="true">
  - <xs:complexContent>
    - <xs:extension base="ifc:IfcBuildingElementComponent">
      - <xs:sequence>
          <xs:element name="SteelGrade" type="ifc:IfcLabel" nillable="true" minOccurs="0" />
        </xs:sequence>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
  <xs:element name="IfcReinforcingMesh" type="ifc:IfcReinforcingMesh"
    substitutionGroup="ifc:IfcReinforcingElement" nillable="true" />
```

Figure 2: Organization of IfcXML

Compared to industry-specific taxonomies, regulations are voluminous and cover a broad range of topics. Increasingly, regulatory documents are available online and are organized in HTML or XML structure. The International Building Code (IBC) (International Code Council (ICC), 2006), which represents the code of practice in the building industry, is employed as one of the regulatory documents in this study. Figure 3 shows a provision in IBC and its representation in XML. One notable feature of regulations is that they are typically organized into sections and sub-sections, each of which contains content addressing a specific topic. The

tree hierarchy of regulations provides useful information that can be explored, for example, to locate sections that cover similar topics (Lau, 2004, 2005).

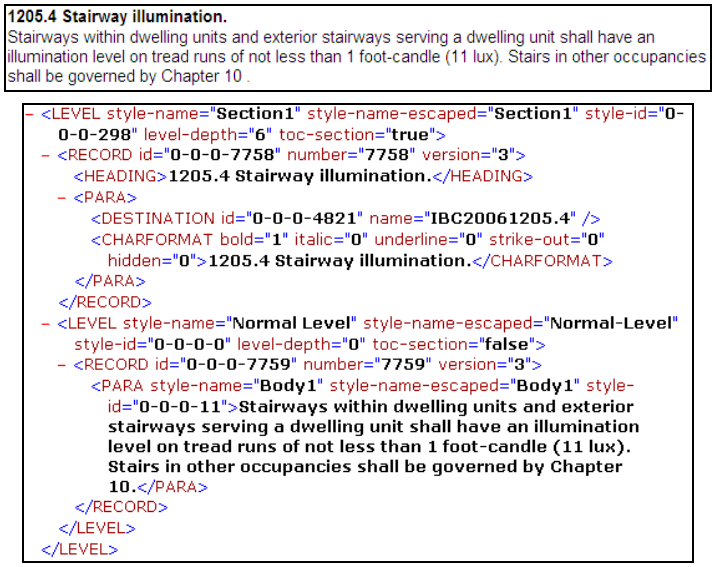

Figure 3: An IBC Provision and XML Structure

4. ONE TAXONOMY TO ONE REGULATION

Mapping one taxonomy to one regulation is a basic keyword mapping task. There are many commercial tools that perform this task. Node labels in the taxonomy tree are treated as concept keywords that get mapped to sections (or sub-sections) in the regulation where they appear. Figure 4 shows the International Building Codes latched with the OmniClass. Users can now traverse the taxonomy and browse relevant sections of the regulation. For instance, a small business owner can browse the OmniClass classification for the concept “lighting” and subsequently retrieve relevant IBC sections, such as Section 1205.2.1 titled “adjoining spaces” in Figure 4.

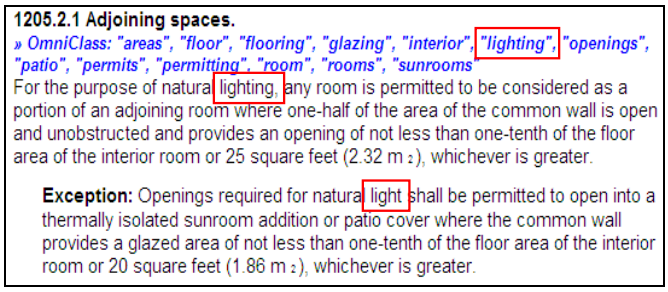

Figure 4: Regulation Latched with Taxonomy Concepts

Extending the mapping from one taxonomy to multiple regulations leads to the classic problem of information overload. For instance, suppose we want to search the Web to find state regulations in Alabama and Arizona governing chlorine levels in drinking water. If we search the drinking-water regulations from those states for the concept "chlorine", we would find over 60 total sections. The actual relevancy of any of these 60 sections to chlorine levels is not known. The problem is that Web search engines cannot take document structure into account when computing relevancy. The result is information overload. Research on intelligent retrieval and presentation of web content has begun, but the results are not yet available in commercial products (Bonnel, Lemaire, Cotarmanac'h, & Morin, 2006).

Fortunately, regulatory documents are much more organized and structured than web content. Therefore, we propose to solve the problem of information overload by clustering relevant sections from different regulations and pivoting on one regulation with which the user is most familiar. We discuss our approach in the following section.

5. ONE TAXONOMY TO MULTIPLE REGULATIONS

Simultaneous traversal of multiple regulation trees using one taxonomy is a challenging but frequently encountered problem. Consider the example above and a scenario in which an engineer from Alabama must design a water distribution system that provides water to Phoenix,

Arizona from lakes near Montgomery, Alabama. The engineer is likely to be familiar with the Alabama state code, but not the Arizona state code. Since the water distribution system will be subjected to comply with both regulations, the engineer must find the relevant sections in the Arizona code. 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. We believe this approach significantly reduces information overload.

Figure 5 shows a simple user interface for finding related provisions in the regulations from the two states. After browsing down the taxonomy tree to the concept "bacteria", users are shown a list of matched sections from the Alabama regulation. This matching is done using the technique described in Section 3. Selecting Section 335.7.5.23 of the AL code shows that there are 16 recommended sections from the Arizona regulation. There are two major challenges to developing such a system: a suitable user interface and a methodology for determining relevant regulations. Here, we discuss briefly an approach for making recommendations based on relevancies between sections from different regulations.

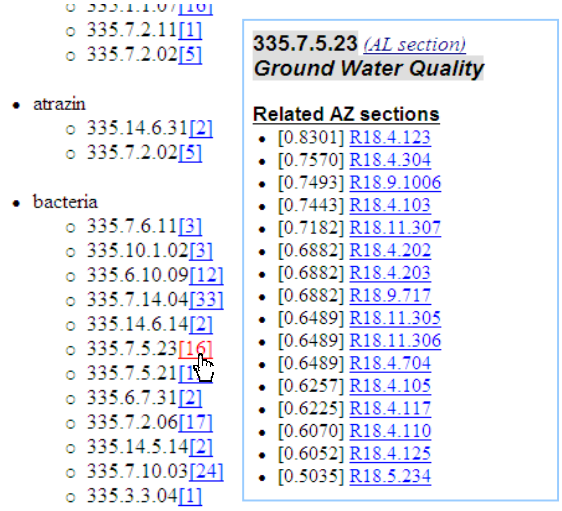

Figure 5: Concept "bacteria" mapped to Section 335.7.5.23 in AL code, which has 16 related sections in AZ code

To identify related provisions from different regulations, we use the relatedness analysis technique from Lau (2004, 2005). This technique compares sections from different regulations based on shared features using a Vector Space model (Larsen & Aone, 1999; Nahm, Bilenko, & Mooney, 2002). The goal is to identify the most strongly related provisions using not only a traditional term match but also a combination of feature matches, content comparison, and 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 provisions and the referential structure. These structures represent useful information for locating related provisions and are, therefore, used in the analysis as well. For the detailed discussion on the methodology and evaluations of results from the relatedness analysis of provisions see (Lau, 2004).

6. MULTIPLE TAXONOMIES TO ONE REGULATION

As suggested in the Introduction, multiple taxonomies have been developed for different applications within the same industry domain. Most industry practitioners are familiar with at least one of the taxonomies; but, they frequently need to deal with others for various applications (Begley et al., 2005; Lipman, 2006). Traversing regulations using multiple taxonomy trees is a challenging problem, and a potential solution is to merge multiple taxonomies into one. Techniques for merging ontologies are discussed in (Noy, 2003; Stumme & Maedche, 2001). These techniques 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 the newly merged ontology in order to browse regulations, this would defeat the original intent of using existing taxonomies to

help locate regulatory provisions. Using the argument from Section 5, we will focus on one taxonomy then derive related concepts from other taxonomies.

Figure 6 illustrates the proposed approach using the OmniClass and the IFC taxonomies, and the International Building Code (IBC) regulations discussed above. In this scenario, we assume that the user is more familiar with the OmniClass hierarchy, and thus starts browsing the IBC using this taxonomy. The OmniClass is altered from its original representation (see Figure 6) to display a widget upon mouse-over that includes an ordered list of matching IBC sections and recommended relevant IFC concepts. The user uses the term “concrete” from OmniClass to find an ordered list of matching IBC sections and relevant IFC concepts. Upon locating a list of IBC sections that are related to “concrete”, sorted in order of relevance, the user also sees a list of related IFC concepts including “beam”. Mousing-over the IFC concept “beam” brings the focal point to the IFC hierarchy, where the user is presented with the same analysis – namely the IFC elements around this concept “beam”, a ranked list of matching IBC sections, and a ranked list of relevant OmniClass concepts.

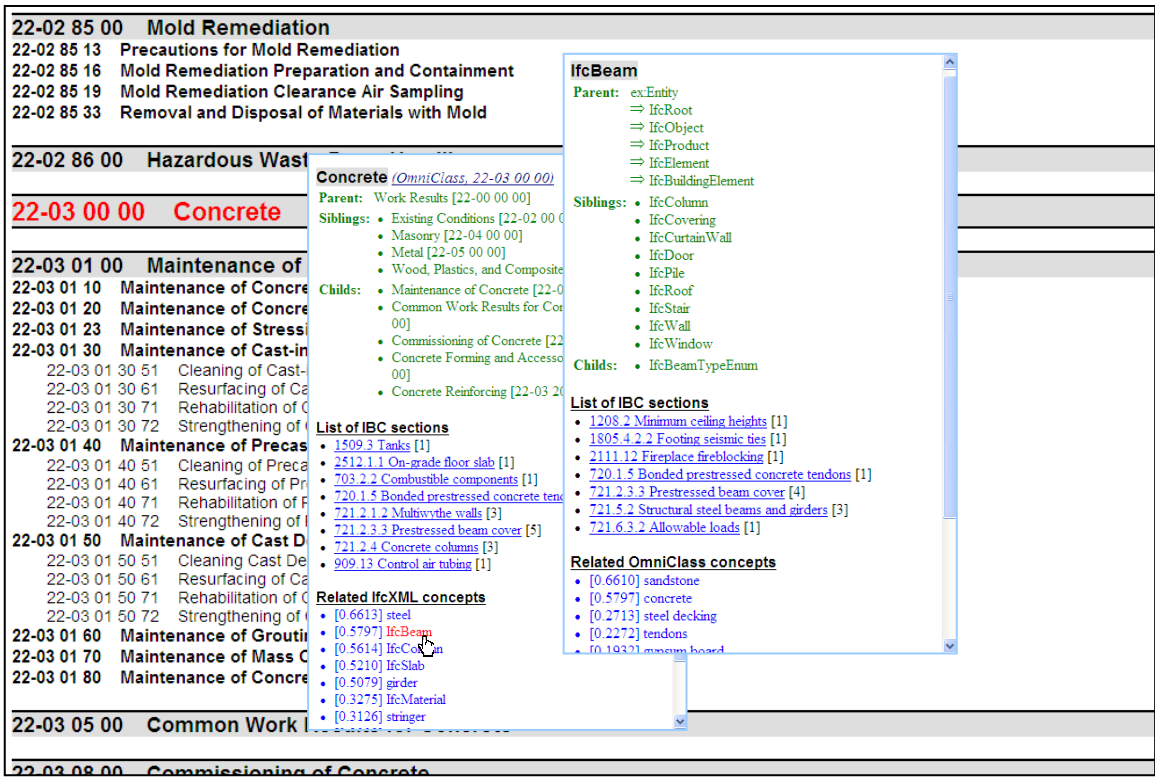

Figure 6: Traversing the IBC using OmniClass Taxonomy with Relevant Concepts from the IFC Taxonomy

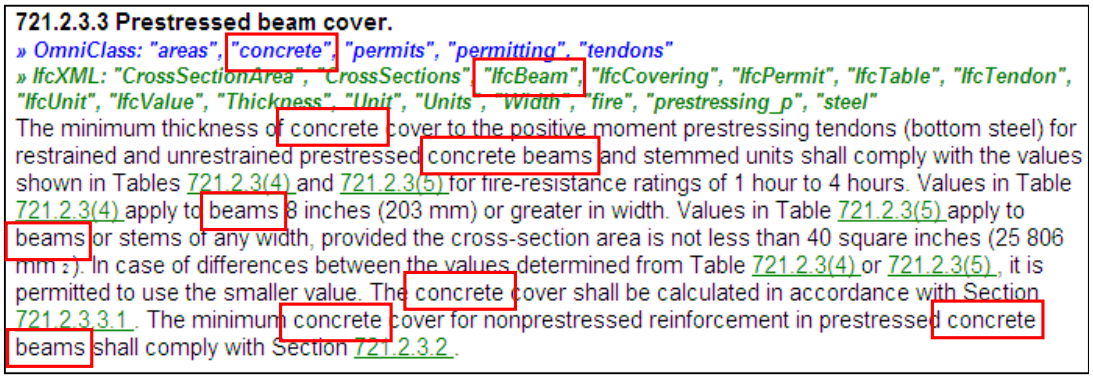

Figure 7: IBC section that OmniClass concept "concrete" and IFC concept "beam" co-occur

The task here is to identify similar or related concepts from multiple taxonomies. This is equivalent to mapping from one ontology to another. Ontology mapping has been an active research area since the semantic web movement (Mitra, 2003; Mitra & Wiederhold, 2001). That research has shown that it is difficult to develop mappings between two arbitrary ontologies. In our case, however, the ontologies are not arbitrary; they are domain specific and targeted

towards the same group of users. Therefore, we can extend the techniques presented in Section 5 by using a carefully selected document corpus to relate concepts by computing their co-occurrence frequencies. Conveniently, we have a corpus of regulatory documents that have been meticulously drafted and reviewed for accuracy. We believe that this corpus will dramatically increase the likelihood of finding accurate matches between concepts from different taxonomies.

We investigated three different methodologies for clustering relevant concepts from different taxonomies by computing a similarity score between concepts (Cheng et al., 2008): Cosine similarity, Jaccard coefficient, and Market baskets. Cosine similarity and Jaccard coefficient are vector-based similarity measures commonly used in the field of information retrieval. The market basket model is a popular technique in data mining. In Figure 6, we relate the concept "concrete" from the OmniClass taxonomy to the concept "beam" from the IFC taxonomy using these methods. Figure 7 shows one of the IBC sections in which the two concepts co-occur. As illustrated in the figure, the concepts "concrete" and "beam" appear in the IBC Section 721.2.3.3 five times and four times, respectively. The more "concrete" and "beam" co-occur in a unit of regulation, the higher their similarity score is. A unit of regulation here refers to a section. As shown in the Figure 6, their similarity score is 0.58, which ranks second among all IFC concepts that are relevant to "concrete".

6.1 Evaluations of the Three Similarity Measures

In order to evaluate the performance of the three investigated methods, we need to establish a test set where benchmark results can be determined by domain experts. The test data set consists of twenty concepts randomly selected from the OmniClass and the IFC hierarchies respectively. Three domain experts are asked to identify the related concept pairs among a total

of 400 possible pairs. The results of concept matching performed by domain experts are treated as the true matches. Pairwise similarity scores are computed using the three relatedness analysis measures described above. Concept pairs deemed related by domain experts are assigned a true value of one; all other pairs are assigned a true value of zero. As for the predicted values using the three measures, two concepts are predicted as similar or related if the computed similarity score is larger than a predetermined threshold score. Based on the three manual and independent mappings from domain experts, we compute values of precision and recall to evaluate the performance of the three measures.

Precision and recall evaluate the accuracy of predictions and the coverage of accurate pairs. Precision measures the fraction of predicted matches that are correct, i.e., the number of true positives over the number of pairs predicted as matched. Recall measures the fraction of correct matches that are predicted, i.e., the number of true positives over the number of true matches. They are computed as

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

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

The evaluation results are plotted in Figure 8. Jaccard similarity achieves the best precision result, followed by first Cosine similarity and then the market basket model. By comparing their recall values, the results are reversed; namely the market basket model is best, followed by Cosine and Jaccard similarities. There is always a tradeoff between precision and recall. This is because the more predicted matches we include by lowering the threshold similarity score, the higher the recall rate becomes - there are bound to be more true matches.

On the other hand, the precision rate is lower since we also introduce false matches inadvertently. In our evaluation, Jaccard similarity is least preferred due to its very low recall rates. Cosine similarity appears to be average among the three measures with acceptable precision and recall rates. Likely, a combination of techniques will produce optimized precision and recall values after some tuning.

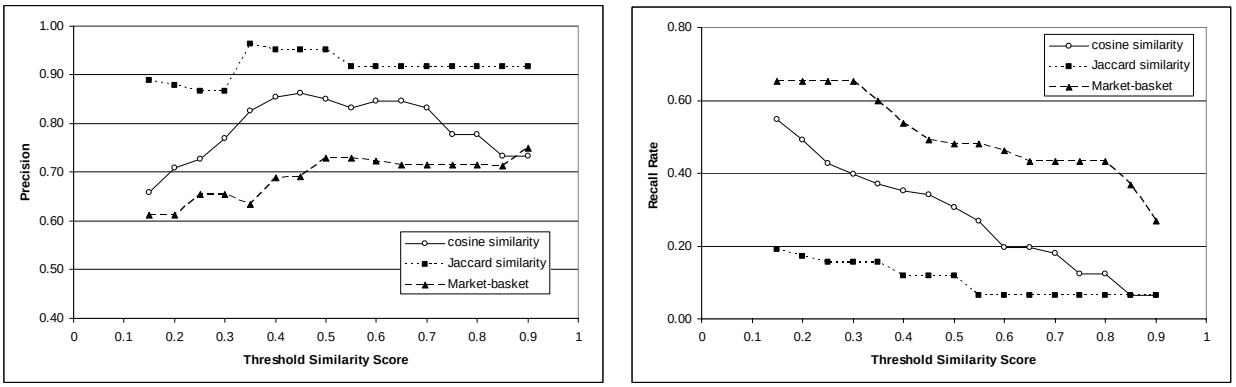

Figure 8: Precision and recall evaluations of three similarity metrics

7. CONCLUSIONS & FUTURE TASKS

One of the key components of the next generation digital government infrastructure is to better facilitate Government-to-Citizen and Government-to-Small Business communications and transactions. Coming from the government, regulations have the force of law. It is vital that citizens and small businesses have the means to access and retrieve them when needed. Some of the current digital government initiatives have already recognized the need for a one-stop shop to help small business owners access regulatory compliance information[2]. A growing amount of regulations are available online so that citizens can browse and peruse them. This assumes, however, that the information seekers are capable of untangling the massive volume and complexity of the law.

One of the complexities of regulatory documents originates from how regulations are written – they are written by government and code-issuing agencies who organize the material to suit their own needs. A concept or term may be located in multiple sections within a regulation document or in multiple regulations under different jurisdictions. From industry practitioners' standpoint, even for a single regulation document, the original regulatory hierarchy might hinder the retrieval of relevant regulations. We believe the next generation digital government infrastructure should support easy access to regulations based on citizen's and small business's mental models.

To this end, this paper proposed a system to map concepts from industry-specific taxonomies to similar concepts in related regulations. The system can potentially help industrial users locate, retrieve and relate regulations relevant to their needs, and therefore increase the accessibility and usability of regulations. Policy makers can organize and manage their regulations according to their classification systems while enabling the retrieval by the industrial communities, who often use a different organization and classification system. We used a running example from the AEC industry to illustrate the need, the usage, and the potential benefit of the mapping system.

Our future plan to improve the system will focus on two directions – (1) improving the relatedness analysis techniques by revising the metrics and extending to n-n concept-section mapping problem, and (2) designing and evaluating the user interface according to users' needs and feedbacks. We demonstrated 1-1, 1-n, n-1 mappings between taxonomies and regulations. In section comparisons, we took advantage of the hierarchical structure of regulations with well defined contents in each section to enhance the analysis. Our next step will be to incorporate the

---

[2] See www.business.gov .

hierarchical structure of taxonomies into concept comparisons. In concept comparisons, we evaluated the performance of three similarity metrics: Cosine similarity, Jaccard similarity, and the market basket model. A natural next step would be to combine and tune the three metrics in order to attain optimized precision and recall values. Further evaluation and improvement of the related analysis techniques to provide more precise results for the user's domain are being investigated.

Our current relatedness analysis approach considers the co-occurrence of concepts on the section level because we believe each regulation section contains well-defined contents for a specific scope. For sections that cover a broad scope, however, unrelated concepts may co-occur in the same section but different paragraphs. A finer granularity of analysis on the paragraph or sentence level might improve prediction accuracy; we will test this idea in the future. We also plan to implement an n-n concept-section mapping in the future, by combining the techniques of concept comparisons and section comparisons.

More user studies, in the form of focus groups composed of small business owners and industry practitioners, can help evaluate the pilot regulatory infrastructure. In the future, we would like to engage potential users to help perform formal evaluations of the similarity metrics 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 visualization tools for this purpose. An ideal user interface should facilitate access to the mapping of multiple taxonomies and the browsing of regulations by domain experts. Once the pilot framework is deemed usable by industry practitioners, we plan to investigate the adoption of the framework by regulation stakeholders, rulemakers and policy writers in government agencies. The pilot framework can become a frontend visualization tool for regulations that plugs into the current digital

government infrastructure. The burden of identifying and gathering relevant taxonomies from different regulatory domains should be shared between rulemakers and industry practitioners.

To this end, ideally, the next generation regulatory framework should accept suggestions of industry-specific taxonomies per regulation. As most ontologies are specified in a semantically descriptive syntax such as the Resource Description Framework (RDF)[3], parsers can be developed to automate the extraction of concepts from ontologies, in order to map them to regulations. We believe that such regulatory infrastructure can be successfully deployed by the government and readily adopted by small businesses, as taxonomies are being developed and standardized by industry practitioners.

8. ACKNOWLEDGMENTS

The authors would also like to acknowledge the supports by the National Science Foundation, Grant Numbers CMS-0601167 and IIS-0811975, 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 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.

[3] See www.w3.org/RDF/ for specification of RDF.

9. REFERENCES

[1] Al-Kofahi, K., Tyrrell, A., Vachher, A., & Jackson, P. (2001). *A Machine Learning Approach to Prior Case Retrieval.* Paper presented at the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri.

[2] Begley, E. F., Palmer, M. E., & Reed, K. A. (2005). *Semantic Mapping Between IAI ifcXML and FIATECH AEX Models for Centrifugal Pumps* (No. NISTIR-7223): National Institute of Standards and Technology (NIST).

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

[4] Bonnel, N., Lemaire, V., Cotarmanac'h, A., & Morin, A. (2006). *Effective Organization and Visualization of Web Search Results.* Paper presented at the 24th IASTED International Conference on Internet and Multimedia Systems and Applications, Innsbruck, Austria.

[5] Brüninghaus, S., & Ashley, K. D. (2001). *Improving the Representation of Legal Case Texts with Information Extraction Methods.* Paper presented at the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri.

[6] Brunnermeier, S. B., & Martin, S. A. (1999). *Interoperability Cost Analysis of the US Automotive Supply Chain* (No. #99-1): Program Office of Strategic Planning and Economic Analysis Group, National Institute of Standards and Technology (NIST).

[7] Cheng, C. P., Lau, G., Law, K. H., Pan, J., & Jones, A. (2008). Regulation Retrieval Using Industry Specific Taxonomies. *Artificial Intelligence and Law (accepted)*.

[8] Construction Specifications Institute (CSI). (2006). *OmniClass Construction Classification System, Edition 1.0*. Retrieved August 2, 2006, from http://www.omniclass.org

[9] Fountain, J. E. (2003). Information Institutions and Governance: Advancing a Basic Social Science Research Program for Digital Government. Research Working Paper RWP 03-004: National Center for Digital Government, John F. Kennedy School of Government, Harvard University.

[10] Gallaher, M., O'Connor, A., Bettbarn, J., Jr., & Gilday, L. (2004). *Cost Analysis of Inadequate Interoperability in the US Capital Facilities Industry* (No. GCR 04-867): National Institute of Standards and Technology (NIST).

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

[12] International Alliance for Interoperability (IAI). (2007). *Industry Foundation Classes (ifcXML), Edition 2x3*. Retrieved August 8, 2007, from http://www.iai-tech.org/

[13] International Code Council (ICC). (2006). 2006 International Building Code, Whittier, CA.

[14] Jacobs, J., & Linden, A. (2002). *Semantic Web Technologies Take Middleware to the Next Level* (No. T-17-5338): Gartner Group.

[15] 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.

[16] Kerrigan, S., & Law, K. (2003). *Logic-Based Regulation Compliance-Assistance.* Paper presented at the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003), Edinburgh, Scotland.

[17] Larsen, B., & Aone, C. (1999). *Fast and Effective Text Mining Using Linear-Time Document Clustering.* Paper presented at the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA.

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

[19] Lau, G. (2005). *Uses of Text Mining Techniques in Extracting Entity Relationships from Legal Documents.* Paper presented at the Data Mining, Information Extraction, and Evidentiary Reasoning for Law Enforcement and Counter-Terrorism Workshop, Bologna, Italy.

[20] Lipman, R. (2006). *Mapping Between the CIMsteel Integration Standards (CIS/2) and Industry Foundation Classes (IFC) Product Model for Structural Steel.* Paper presented at the Conference on Computing in Civil and Building Engineering, Montreal, Canada.

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

[22] Mitra, P., & Wiederhold, G. (2001). *An Algebra for Semantic Interoperability of Information Sources.* Paper presented at the 2nd IEEE Symposium on BioInformatics and Bioengineering, Bethesda, MD.

[23] Moen, W. E. (2001). The metadata approach to accessing government information. *Government Information Quarterly,* 18(3), 155-165.

[24] Moens, M.-F., Uyttendaele, C., & Dumortier, J. (1997). *Abstracting of Legal Cases: The SALOMON Experience.* Paper presented at the 6th International Conference on Artificial Intelligence and Law (ICAIL 1997), Melbourne, Australia.

[25] Nahm, U. Y., Bilenko, M., & Mooney, R. J. (2002). *Two Approaches to Handling Noisy Variation in Text Mining.* Paper presented at the ICML-2002 Workshop on Text Learning, Sydney, Australia.

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

[27] Office of Management and Budget (OMB). (2008). *Presidential Initiatives: Powering America's Future with Technology*. Retrieved August 27, 2008, from http://www.whitehouse.gov/OMB/egov/c-presidential.html

[28] Prokopiadou, G., Papatheodorou, C., & Moschopoulos, D. (2004). Integrating knowledge management tools for government information. *Government Information Quarterly*, 21(2), 170-198.

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

[30] Schweighofer, E., Rauber, A., & Dittenbach, M. (2001). *Automatic Text Representation, Classification and Labeling in European Law*. Paper presented at the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri.

[31] Skrzycki, C. (2000, May 23). The Regulators; Compliance Education Goes Self-Service. *The Washington Post*.

[32] Small Business Administration. (2002). *Business Compliance One Stop Workshop*. Queenstown, MD.

[33] Stumme, G., & Maedche, A. (2001). *Ontology Merging for Federated Ontologies on the Semantic Web*. Paper presented at the International Workshop on Foundations of Models for Information Integration (FMII 2001), Seattle, WA.

[34] Thompson, P. (2001). *Automatic Categorization of Case Law*. Paper presented at the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001), St. Louis, Missouri.

[35]