An Information Infrastructure for Government Regulation Analysis and Compliance Assistance

Gloria T. Lau[1], Shawn Kerrigan[1], Haoyi Wang[1], Kincho H. Law[1], Gio Wiederhold[2]
Department of Civil & Environmental Engineering[1], Computer Science Department[2]
Stanford University, Stanford, CA 94305
glau@stanford.edu, kerrigan@stanfordalumni.org, haoyiw@stanford.edu,
law@stanford.edu, gio@cs.stanford.edu

1 Project Overview

The complexity and diversity of government regulations make understanding the regulations a non-trivial task. One of the issues is the existence of multiple sources of regulations and interpretive guides; the latter are often independent of governing bodies. In this work, we describe a research prototype system that combines text mining and knowledge management techniques to help better manage, understand and analyze regulatory documents. This regulatory information infrastructure includes three integral parts: a document repository, a tool for similarity analysis and a compliance assistance system. This paper first presents the development of a legal corpus with multiple sources of regulatory documents consolidated into a unified format. A shallow parser is developed to consolidate different regulations into a unified XML format, which is well suited for handling semi-structured data such as legal documents. Important features, such as concepts, measurements, definitions and so on, are extracted and incorporated into the corpus by using handcrafted rules and text mining tools. A regulation compliance assistance system is introduced next, where First Order Predicate Calculus (FOPC) logic sentences are implemented to help users to perform compliance check in a question and answer session. Finally, a similarity analysis for regulations is developed, where Information Retrieval (IR) and structural matching techniques are used to identify related provisions among regulations.

2 Repository Development

In order to develop a prototypic system, this work focuses on accessibility and environmental regulations. Our corpus includes regulations from Federal and state governments, as well as selected supplementary and supportive documents. Presently, regulatory documents are available in Hypertext Markup Language (HTML), Portable Document Format (PDF) or hardcopy. To ease the development of document analysis tools, we have chosen the eXtensible Markup Language (XML) as a unified format to represent regulations in our corpus because of XML's capability to handle semi-structured data. A shallow parser is first developed to consolidate documents into XML format, as well as to extract feature information as discussed below. The hierarchical structure of regulations is preserved by properly structuring provisions as XML elements. For instance, Section 4.7.4 is a subsection in Section 4.7, and thus is structured to be a child node of the XML element of Section 4.7. In addition to properly preserving the hierarchy of regulations in XML, our system also extracts referential structures, such as an explicit reference from Section 4.7.4 to Section 4.5, through a context-free parsing system. With the hierarchical structure captured in XML, different rendering tools can be used to display and view regulations in its natural organization.

To incorporate domain knowledge as well as conceptual information in regulations, we extract specific *features* from the repository and refine the XML structure. The process of feature extraction identifies the important features from the corpus that signal similarity or relatedness. Examples of features include concepts such as noun phrases, measurements specific to accessibility and environmental standards, and terminology definitions available in most regulations. Handcrafted rules and text mining tools are used to match features automatically in provisions, and the corpus of documents is refined with the extracted features tagged as additional XML elements in provisions where the features appear. Provisions can now be rendered in a web browser with useful features highlighted; for instance, users can browse through

referenced sections through hyperlinks, search the repository with suggested concepts that are identified in the current provision, as well as look up definitions of specific terms.

3 Regulation Assistance System

An online repository of government regulations allows users to retrieve interested documents with ease; however, there still remains the question of compliance with the provisions and their implicit references to others. To facilitate manipulation and interpretation of regulations, we employ a logic-based compliance check system. Logic and control processing metadata are introduced to our XML-based regulation framework to support a compliance-checking session. The purpose of the metadata is to guide users through regulations and to identify potential conflicts with the implemented rules translated into logic sentence based on provision contents. These logic metadata are associated with provisions as additional XML tags.

We employ Otter, a publicly available FOPC theorem prover developed at the Argonne National Laboratory, for logic check. A compliance-checking session begins with a web interface with questions based on the implemented logic metadata. Users may select a response from a menu of possible answers, including "Yes", "No" and "I don't know" options, where the "I don't know" option forks the compliance process along all possible answers. The system then checks user answers against the embedded logic sentences implemented based on provision requirements. Control logic metadata are used to check referenced provisions for specific compliance requirements outside of the current provision. When the system completes a check against the regulation provisions or detects a conflict between the user's answers and the regulation, it displays a summary of the question-and-answer history as well as the compliance results. The logs of the compliance session allow users to maintain a detailed compliance record which is useful for record keeping or when the regulations are to be revisited in the future.

4 Similarity Analysis

Apart from compliance checking and assistance, another capability of our prototype system is similarity analysis across different sources of regulations. We employ a combination of IR techniques and document structure analysis to extract related provisions based on a similarity measure, which is defined as a similarity score between 0 and 1. Since typical regulations are massive in size, we take a provision as the unit of comparison. The goal is to identify the most related provisions across different regulation trees using not only a traditional term match but instead a combination of feature matches, and not only content comparison but also structural analysis. This is obtained by first comparing regulations based on conceptual information as well as domain knowledge through a combination of feature matching. In addition, legal documents possess specific structures, such as the natural tree hierarchy and the referential structure. These structures also represent useful information in locating related provisions, and are therefore incorporated into our analysis for a more accurate comparison.

We first compute a base score between two provisions by matching extracted features. This design provides the flexibility to add on features and different weighting schemes if domain experts desire to do so. The scoring scheme for each of the features essentially reflects how much resemblance can be inferred between the two sections based on that particular feature. The base score is subsequently refined by utilizing the tree structure of regulations. The parent, siblings and children (the immediate neighbors) of the interested sections are compared to include similarities between the interested sections that are not previously accounted for based on a direct comparison. In other words, similarities between the immediate neighbors imply similarity between the interested pair. The referential structure of regulations is handled in a similar manner, based on the assumption that similar sections often reference each other. In essence, the heavily self-referenced structure of regulations is used to further refine the similarity score. Therefore, similarities from both near-tree neighbors and references are identified, and related provisions are retrieved based on the resulting scores. For example, a comparison between American and British standards using our similarity analysis system revealed interesting terminology differences such as "door hardware" versus "door furniture" that cannot be identified with a simple term match.