An E-Government Information Architecture for Regulation Analysis and Compliance Assistance

Gloria T. Lau
Civil & Env. Engineering
Stanford University
glau@stanford.edu

Shawn Kerrigan
Civil & Env. Engineering
Stanford University
kerrigan@stanfordalumni.org

Kincho H. Law
Civil & Env. Engineering
Stanford University
law@stanford.edu

Gio Wiederhold
Computer Science
Stanford University
gio@cs.stanford.edu

ABSTRACT
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. In this work, we propose an information infrastructure for regulation analysis, which includes a document repository and tools for compliance assistance and similarity analysis. A regulatory repository is developed based on an XML format, and important features, such as concepts and measurements, are extracted using handcrafted rules and a text mining tool. Our framework provides compliance assistance using a reasoning tool based on First Order Predicate Calculus logic, where users are alerted of detected conflicts or otherwise compliance with the regulation. A relatedness analysis is performed by comparing the extracted features as well as structural and referential information from regulations. Examples of an electronic-rulemaking scenario and a compliance checking procedure are shown to demonstrate current capabilities of the prototype system.

Categories and Subject Descriptors
I.7.2 [**Document and Text Processing**]: Document Preparation – *Markup Languages*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *linguistic processing*; J.1 [**Administrative Data Processing**]: Law.

General Terms
Management, Documentation, Standardization, Languages, Theory, Legal Aspects.

Keywords
Legal Informatics, E-government, E-rulemaking, Shallow Parsing, Similarity Analysis, Text Mining, Compliance Check.

1. INTRODUCTION
Government regulations should ideally be understandable and retrievable with ease by practitioners as well as the general public. In reality, regulations are voluminous, heavily cross-referenced and often ambiguous. Multiple sources of regulations, for instance, from the Federal, State and local governments, amend, complement and potentially conflict with one another. There are many reference guides, that are published independent of governing bodies, attempting to help the public to better understand and comply with the regulations. As a result, the regulations, amending provisions and interpretive manuals together create a massive volume of semi-structured documents with similar content but potential differences in format, terminology and context. It becomes a difficult task for individuals to search through multiple codes with multiple terms to locate related provisions, if there is any. Nonetheless, there is a need to identify as much relevant information as possible, since as Berman and Hafner have noted, "[a] vast amount of information … must be collected and integrated in order for the legal system to function properly [6]. " An information infrastructure that can consolidate, check for compliance, compare and contrast different regulatory documents will greatly enhance and aid the *understanding* of regulations.

To motivate the problem, we give a classic example of such complexity and conflict found across different regulations as shown in Figure 1 [16]. Both Federal and California regulations provide design requirements of a curb ramp. However, the Federal regulation [3] focuses on wheelchair traversal, which is in conflict with the California regulation (this provision is from the 1998 version) [11] focusing on the visually impaired when using a cane. The conflict is captured by the clash between the term "flush" and the measurement "1/2 inch lip beveled at 45 degrees." Clearly, a framework for regulation comparison and compliance assistance is much desired to alert users of related information.

ADA Accessibility Guidelines 4.7.2: Slope
Slopes of curb ramps shall comply with 4.8.2. The slope shall be measured as shown in Figure 11. Transitions from ramps to walks, gutters, or streets shall be **flush** and **free of abrupt changes**. Maximum slopes of adjoining gutters, road surface immediately adjacent to the curb ramp, or accessible route shall not exceed 1:20.

California Building Code 1127B.5.5: Beveled lip
The lower end of each curb ramp shall have **a ½ inch (13mm) lip beveled at 45 degrees** as a detectable way-finding edge for persons with visual impairments.

Figure 1: Two conflicting provisions

In this paper, we describe a research prototype system that combines text mining and knowledge management techniques to help better manage, understand and analyze regulatory documents. The example domains include accessibility and environmental regulations. This paper is organized as follows: related work is reviewed in Section 2, where several legal expert systems, feature extraction and Information Retrieval (IR) techniques are discussed. We then present the development of a legal corpus with different sources of regulatory documents consolidated into a unified XML format. Extraction of important features, e.g., concepts, measurements, references and so on, is described in Section 3. A regulation compliance assistance system follows in Section 4. We describe the implementation of First Order Predicate Calculus (FOPC) logic sentences to help users to perform a compliance check in a question and answer session. Section 5 discusses the work on applying information retrieval and structural matching techniques to perform a relatedness analysis between provisions. Preliminary results are shown to illustrate the identification of hidden relatedness of different provisions. Potential application of relatedness analysis for aiding the electronic-rulemaking (e-rulemaking) process is shown in Section 6. A brief summary and discussion on future works are given in Section 7.

2. RELATED WORK

Representation of laws and regulations has been an active research area for decades. There has been a great deal of work on building expert systems for the law [29, 30]. T. Bench-Capon provided a review on the applications of knowledge-based systems for legal applications, particularly the research and development efforts related to the Alvey DHSS Demonstrator project in the UK [5]. The reference includes several hundred citations that appeared before 1990 that are related to logic and rule based approaches and their application in legal systems. Much of the earlier work in IT and law has focused on building systems to optimize decisions with respect to laws, particularly tax law [23]. While legal knowledge representation and reasoning has been an active research topic [1, 2], an integrated approach covering the management of regulations, efficient access and retrieval of documents and tools for compliance checking is missing. This research investigates the issues related to the development of a formal regulatory information management system that supports similarity analysis as well as compliance assistance, based on a consolidated legal repository.

To aid legal research, one can use traditional textual comparison techniques from the field of Information Retrieval (IR). Some examples are the Boolean model or the Vector model [4], with most being bag-of-word type of analysis (i.e. word order insensitive). This type of analysis is insufficient since it ignores the structure of regulations, namely that 1) regulations are organized into deep hierarchies, 2) sections are heavily cross-referenced, and 3) terms are well defined within regulations. A decent similarity analysis tool for a legal corpus should make use of the structure of regulations mentioned above to provide a better comparison. In addition, traditional IR techniques do not take into account the domain-centered nature of legal documents. Laws are developed based on specific areas of application and jurisdiction, where a general index term extraction would fail to capture any related domain knowledge that are available. Example of domain knowledge includes ontologies and field-specific handbooks. Feature extraction provides some help to this end.

Feature extraction is an important step in repository development when the data is voluminous. It is a form of pre-processing, e.g., combining input variables to form a new variable, and most of the time features are constructed by hand based on some understanding of the particular problem being tackled [7]. Automation of this process is also possible; in particular, in the field of information retrieval, software tools exist to fulfill "the task of feature extraction … to recognize and classify significant vocabulary items" [7]. The IBM Intelligent Miner for Text [14] and the Semio Tagger [26] are both examples of fully automated key phrase extraction tools.

Apart from comparing the body text of provisions, the heavily referenced nature of regulations provides extra information about provisions. Link analysis [9] is the natural improvement to the similarity measure. Academic citation analysis [8] is closest in this regard; however the algorithm cannot be directly transported to our domain. Citation analysis assumes a pool of documents citing one another, while our problem here are separate *islands* of information where within island documents are highly referenced; across islands they are not. We are therefore in search of a different algorithm that will better serve our needs.

3. DEVELOPMENT OF AN XML REGULATION REPOSITORY

In order to develop a prototypic system, this work focuses on accessibility and environmental regulations. For accessibility regulations, our corpus currently includes two Federal documents: the Americans with Disabilities Act Accessibility Guidelines (ADAAG) [3] and the Uniform Federal Accessibility Standards (UFAS) [28]. In addition, Chapter 11 of the International Building Code [18], titled Accessibility, is included to reflect the similarity and dissimilarity between federal and private agency mandated regulations. Related sections from the British Standard BS8300 [10] and the Scottish Technical Standards [27] are also included to show the differences and dissimilarities between American and European regulations. For environmental regulations, we currently cover the US Code of Federal Regulations Title 40 (40 CFR): Protection of the Environment [13], along with drinking water provisions from the California Code of Regulations Title 22 (22 CCR) [12]. Our corpus also includes selected supplementary and supportive documents that focus on regulations covering hazardous waste and the management of used oil.

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. Figure 2 shows a schematic of our repository development process. A shallow parser is first developed to consolidate documents from HTML or PDF into XML format. We also extract feature information which will be discussed later this section. The hierarchical structure of regulations, as shown in Figure 3, is preserved by properly structuring provisions as XML elements. For instance, Section 4.7.4 is a provision in Section 4.7, and thus is structured

to be a child node of the XML element of Section 4.7. With the hierarchical structure captured in XML, different rendering tools can be used to display and view regulations in its natural tree shape. Figure 4 shows an example in which a browsing tool called SpaceTree [17], developed at the Human-Computer Interaction Lab at University of Maryland, is used to render regulations as a dynamically rescalable tree.

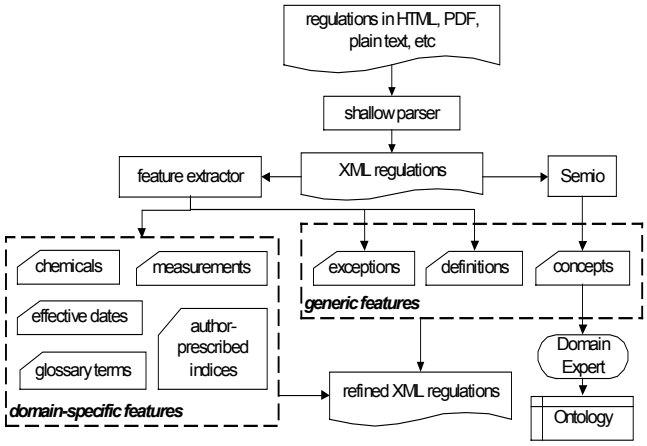


Figure 2: Repository development with feature extraction

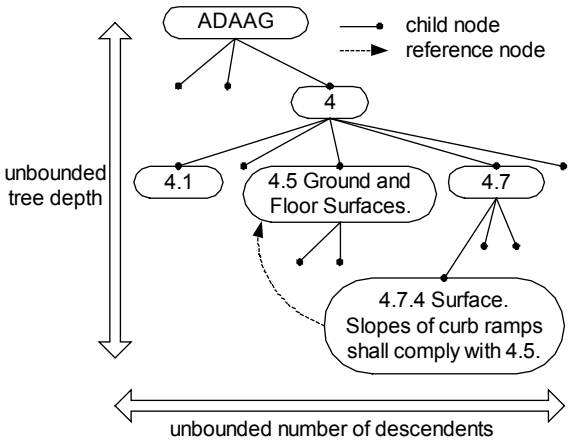


Figure 3: Hierarchical tree structure of regulations

In addition to properly preserving the hierarchy of regulations in XML, our system also extracts referential structures, such as the explicit reference from Section 4.7.4 to Section 4.5 in Figure 3, through a context-free parsing system [19]. We develop a reference parsing system using a context-free grammar and a semantic representation/interpretation system that is capable of tagging regulation provisions with the list of references they contain. Tabular parsing is performed to build parse trees to identify regulation references, such as the example parse tree shown in Figure 5. Our system is shown to correctly identify both simple references, for example, "as stated in 40 CFR section 262.14(a)(2)", and complex references, for example, "the requirements in subparts G through I of this part". When appropriately rendered and linked, references provide users with additional but crucial information to a complete understanding of regulations. The usage of references along with other domain-specific features, which will be introduced below, is illustrated in Figure 6.

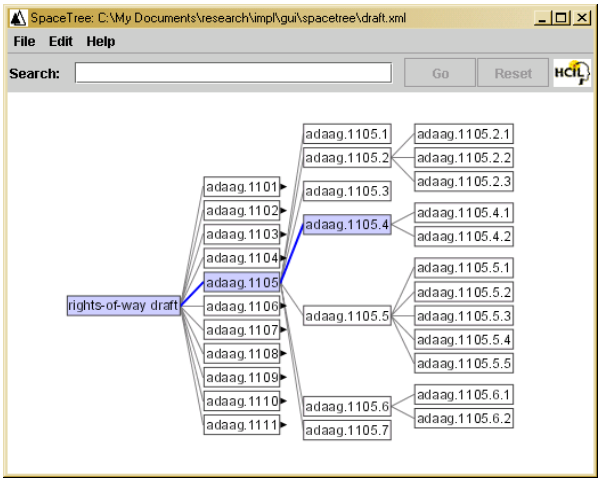


Figure 4: XML regulation rendered as a tree

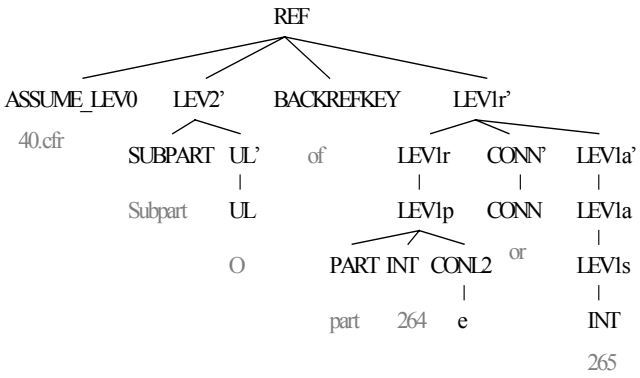


Figure 5: Example parse tree for reference identification

The example shown in Figure 1, where two provisions are in direct conflict, clearly demonstrates the need for a comparison system that brings together related sections in regulations. It further amplifies the importance of conceptual information, such as key phrases in the corpus (e.g., "free of abrupt changes"), as well as domain-specific information, such as measurements (e.g., ½ inch lip), for deep comparisons between provisions. However, traditional textual comparison techniques that employ simple term matching, such as the Vector model [25], lack conceptual understanding of documents. They also suffer from the inflexibility to incorporate domain-specific information. Therefore, our comparison system, which is discussed in Section 5, combines conceptual information with domain knowledge. To enable this deeper comparison, the repository is refined with the extraction of features.

The process of feature extraction identifies the important features from the corpus that signal similarity or relatedness. As shown in Figure 2, there are two types of features: generic features that are

applicable on all areas of law, and domain-specific features. An example of generic feature is concepts, or important noun phrases in the corpus. Concept extraction is performed with the help of the software tool Semio Tagger [26], which is also used for a semi-automated concept ontology generation as shown in Figure 7 to help document retrieval. The ontology is developed by a knowledge engineer based on the list of concepts extracted, and provisions are classified according to the ontology. Users can click through the structure to view relevant sections classified according to concepts.

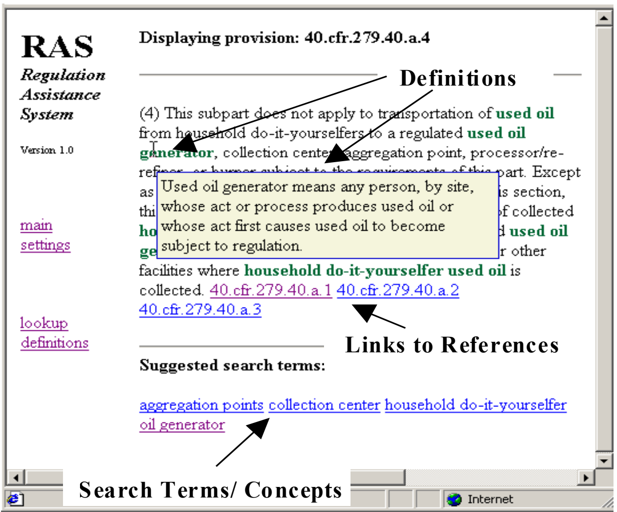


Figure 6: Usages of extracted features

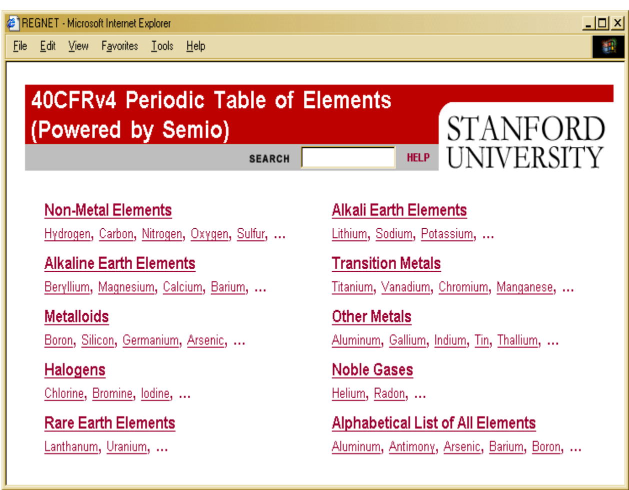


Figure 7: Ontology based on environmental regulations

For other features such as measurements and definitions, handcrafted rules are implemented to automatically match them in provisions where they appear [22]. The corpus of documents is refined with the extracted features tagged as additional XML elements. The underlying XML regulation produced by the shallow parser is showed in Figure 8, which includes excerpts from a provision and its refined XML version that includes several features such as concept, index term and measurement.

Potential usages of these additional tagged features are shown in Figure 6, where a provision is 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. These are all supported by the refined XML regulation framework.

Original Section 4.6.3 from the UFAS

4.6.3 Parking Spaces

Parking spaces for disabled people shall be **at least 96 in** (2440 mm) wide and … shall be part of an **accessible route** to the building or facility entrance and shall comply with **4.3** …
EXCEPTION: If accessible parking spaces for vans ...

Refined Section 4.6.3 in XML format

```
<regElement name="ufas.4.6.3" title="parking spaces">
  <concept name="accessible route" num="1" />
  <indexTerm name="accessible circulation route" num="1" />
  <measurement unit="inch" size="96" quantifier="min" />
  <reference name="ufas.4.3" num="1" />
  …
  <regText> Parking spaces for disabled people ... </regText>
  <exception> If accessible parking spaces ... </exception>
</regElement>
```

Figure 8: Example of XML structures and extracted features

4. LOGIC-BASED COMPLIANCE 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 checking 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 rules [20].

Section 40.cfr.279.20.b.3 from 40CFR

Generators who burn off-specification used oil for energy recovery, except under the on-site space heater provisions of §279.23, must also comply with subpart G of this part.

Regulation logic metadata in XML format

```
<logicSentence>
  all _client _oil ((generator(_client) & usedOil(_oil) &
  burnsForEnergy(_client, _oil) & -satisfied(40_cfr_279_23)) ->
  provApplies(40_cfr_279_G)).
</logicSentence>
```

Figure 9: Logic representation of a provision

Three types of XML metadata are implemented: regulation logic metadata, user interface metadata and control processing metadata. Regulation logic metadata represents a rule or concept that must be followed for an entity to be in compliance with the regulations. An example XML logic representation is shown in Figure 9 where a used oil specification is translated into regulation logic metadata. Apart from regulation logic metadata, user interface metadata also uses FOPC logic sentences to

represent compliance questions and a list of possible user answers to those questions as shown in Figure 10. Control processing metadata provides information about what provisions of a regulation need to be checked for compliance. An example XML entity of such kind is <goto target="40.cfr.279.20.b.3" />, which introduces a new provision to be checked for compliance. Each type of logic or control processing metadata can be associated with any provision in the regulation.

```
User interface metadata in XML format
<logicOption>
  <question>
    Is the used oil used as a dust suppressant?
  </question>
  <logicOpt answer = "yes">
    <logicAns>
      (usedOil(oil1) AND dustSuppressant(oil1)).
    </logicAns>
  </logicOpt>
  <logicOpt answer = "no">
    <logicAns>
      (usedOil(oil1) AND (-(dustSuppressant(oil1)))).
    </logicAns>
  </logicOpt>
</logicOption>
```

Figure 10: Logic representation of a set of compliance question and answers

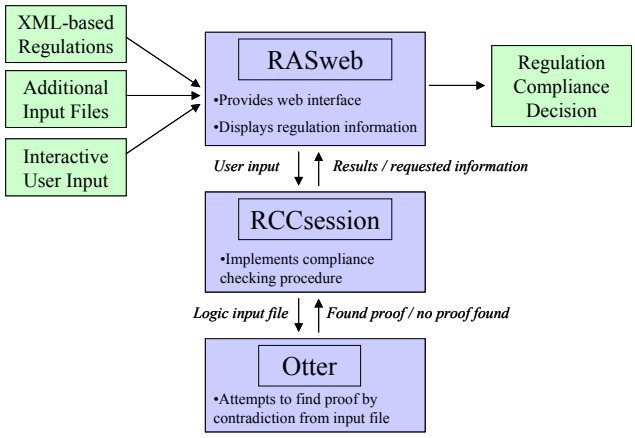


Figure 11: Structure of the regulation assistance system

A regulation assistance system (RAS) is developed based on the XML metadata implemented in the regulations. As shown in Figure 11, the RAS functionality is implemented by a web interface that communicates with a regulation compliance checking (RCC) system. The RCC system parses the XML-structured regulation to extract the information necessary to run a compliance check. Because the performance of FOPC theorem provers decreases rapidly as the number of logic sentences used for reasoning increases, the RCC system properly scopes the metadata to reduce the amount of extraneous data passed to the reasoning system. In essence, only the logic and control processing metadata necessary for the compliance-checking session are acquired and dynamically loaded into the reasoning system.

With the appropriate logic metadata extracted along with interactive user input, the RCC system interacts with a theorem prover for compliance check. The system design is such that any FOPC theorem prover can be used to perform the logic checks. We employ Otter, a publicly available FOPC theorem prover developed at the Argonne National Laboratory [24], for this purpose. As a result, users are provided with a regulation compliance decision based on their input to our system through a web interface, where they will learn the resulting conflicts or compliances with provisions.

For the purpose of demonstration, a used oil regulation (40 CFR 279) has been manually tagged with regulation logic metadata, with user-interface logic metadata, and with control processing metadata. An example scenario of use is given in Figure 12, where an interested user locates a vehicle maintenance shop guide online, from which the user may access information on specific materials or processes, such as used oil. Regulatory requirements for used oil are provided on this guideline, which points to a specific section (40 CFR 279.23) in the Federal Regulation [13] for further compliance information. Here, the online guide can link to our regulation assistance system, where users can check for compliance with the referenced used oil regulation provision or connect to the document repository to look for related supplementary documents. The compliance-checking session is explained below.

A web interface asks users questions based on the user interface metadata embedded in the XML regulation. 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

Figure 12: From industry-specific guides to the regulation assistance system

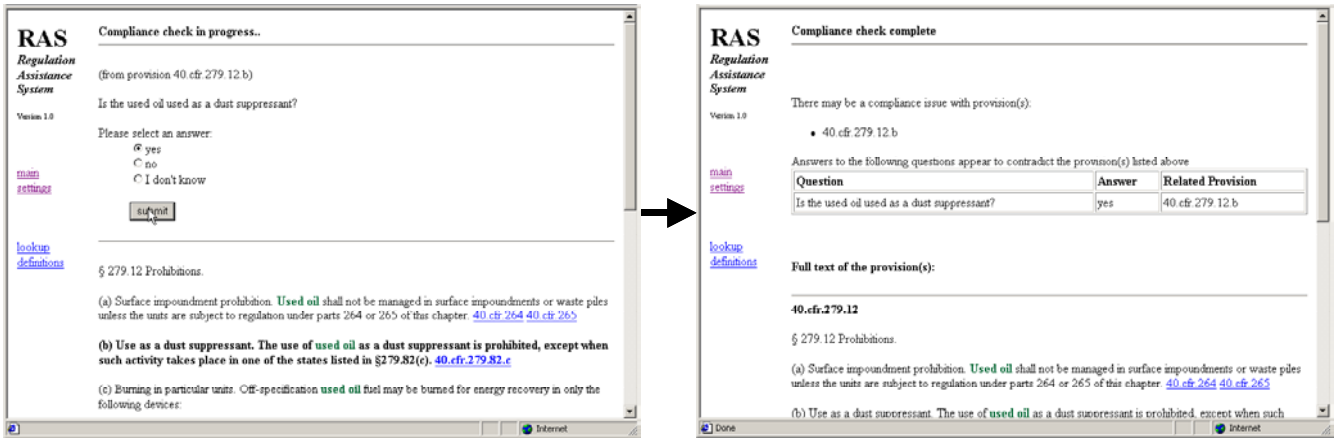


Figure 13: Example compliance-checking session

system then checks user answers against the implemented regulation logic sentences encapsulated in the regulation logic metadata. Control processing metadata mandates the flow of compliance check, for example, following referenced provisions for specific compliance requirements. When the system completes a check against the 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 use of and the results produced by the system are illustrated in Figure 13. 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.

5. AUTOMATED EXTRACTION OF RELATED PROVISIONS

Apart from compliance checking and assistance, another capability of our prototype system is relatedness analysis across different sources of regulations. Starting from a well-prepared repository such as one described in Section 3, we employ a combination of IR techniques and document structure analysis to extract related provisions based on a similarity measure. The degree of relatedness is defined to be 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. Legal documents also possess specific structures, such as the tree hierarchy of regulations in Figure 3 and the referential structure in Figure 5. These structures represent useful information in locating related provisions, and are therefore incorporated into our analysis for a more accurate comparison. A schematic is shown in Figure 14.

We first compute a base score between two provisions by matching extracted features such as those shown in Figure 14. The base score represents a similarity computation based on a combination of generic features, such as concepts, and domain knowledge, such as drinking water contaminants in environmental regulations. 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. For instance, concept matching is done similar to the index term matching in the Vector model [25], where the degree of similarity of documents is evaluated as the correlation between their index term vectors. Using this vector model, we take the cosine similarity between the two concept vectors as the similarity score based on a concept match.

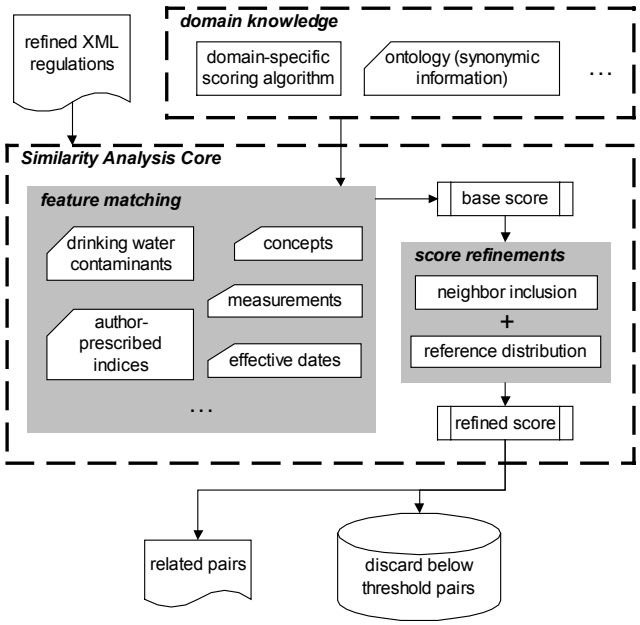


Figure 14: Relatedness analysis

Some features, such as the list of drinking water contaminants in environmental regulations, come with ontologies to define synonyms. Some features simply cannot be modeled as Boolean term matches due to their inherent non-Boolean property, such as measurements (e.g., a domain expert can potentially define a measurement of "12 inches maximum" as 75% similar to a measurement of "12 inches"). Some domain-specific features are provided with feature dependency information defined by

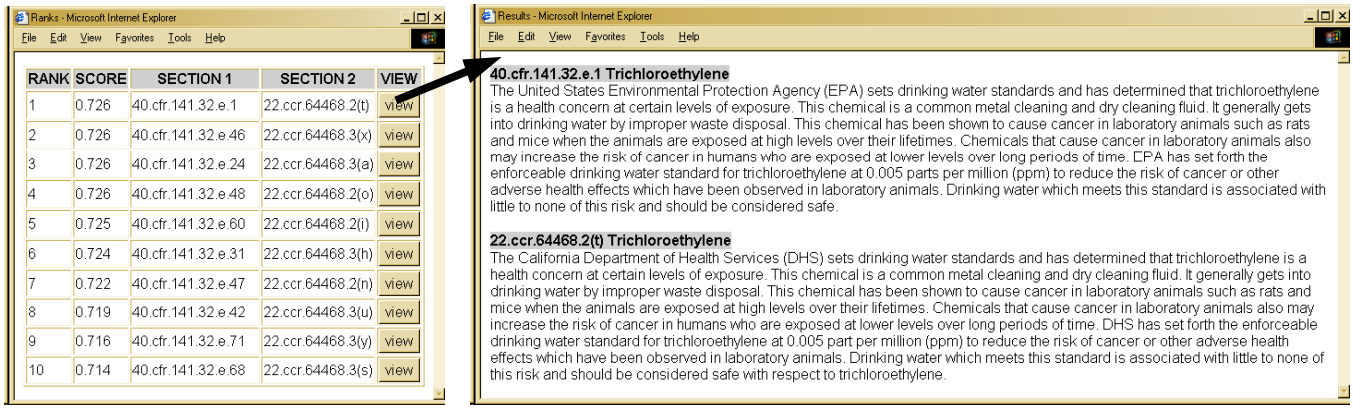


Figure 15: Top ranked related provisions in 40 CFR and 22 CCR

knowledge experts, who do not necessarily agree with a Boolean definition. The limitation of the Vector model is observed: axes are assumed to be mutually independent. Therefore, we modify the Vector model to accommodate dependency information, such as synonyms and non-Boolean matches, via a vector space transformation. In other words, feature vectors are mapped onto an alternate space before cosine comparisons.

The base score is subsequently refined by utilizing the structure of regulations. There are two types of score refinement: neighbor inclusion and reference distribution. In neighbor inclusion, 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, which defines the basis of neighbor inclusion. The referential structure of regulations is handled in a similar manner, based on the assumption that similar sections often reference similar sections. Reference distribution utilizes the heavily self-referenced structure of the regulation to further refine the similarity score.

The final similarity score is a linear combination of the base score, the score obtained from neighbor inclusion as well as reference distribution. We can interpret the base score as a basis of relatedness analysis formed on the shared clusters of similar features between two interested nodes: Sections A and U. Neighbor inclusion infers similarity between Sections A and U based on their shared clusters of neighbors in their respective regulation trees. On the other hand, reference distribution infers similarity through the shared clusters of references from Sections A and U. In essence, the potential influence of the near neighbors are accounted for in neighbor inclusion, while the potential influence of the not-so-immediate neighbors in the tree are incorporated into the analysis through reference distribution. Thus, the final similarity score represents a combination of node content comparisons and structural comparisons.

As a result of a relatedness analysis, related provisions can be retrieved and recommended to users based on the final scores. Results obtained from the comparisons between different regulations are briefly illustrated in Figure 15 to Figure 18, and described in [21]. Figure 15 shows a top ranked pair of related provisions on drinking water control from the 40 CFR [13] and 22 CCR [12]. This pair of provisions, ranked as number one in similarity between the 40 CFR and 22 CCR, indeed is identical in text except the subject of governing agency changes between Environmental Protection Agency (EPA) and California Department of Health Services (DHS). It is not uncommon that one agency directly adopts provisions issued by another agency. Indeed, in the domain of disabled access, our system identified a lot of identical provisions when comparing the ADAAG [3] with the UFAS [28]; however, this is more or less expected since both are Federal regulations.

UFAS
4.13 Doors
4.13.1 General
…
4.13.9 ***Door Hardware***
Handles, pulls, latches, locks, and other …
…
4.13.12 Door Opening Force

BS8300
12.5.4 Doors
12.5.4.1 Clear Widths of Door Openings
12.5.4.2 ***Door Furniture***
Door handles on hinged and sliding doors …

Figure 16: A Comparison between Section 4.13.9 in UFAS and Section 12.5.4.2 in BS8300

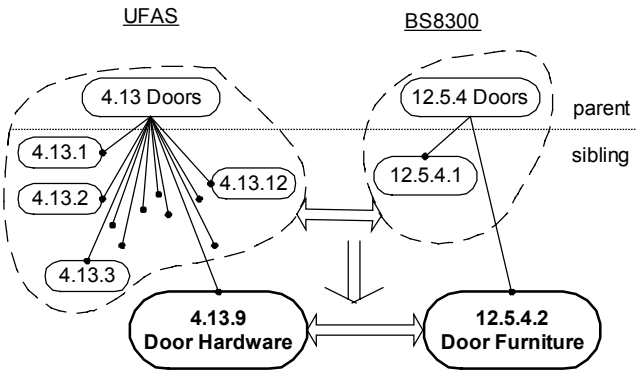


Figure 17: Illustration of a comparison between American and British regulations

To illustrate the similarity between American and British standards, we compare the UFAS [28] with the BS8300 [10]. Figure 16 and Figure 17 show a subtree of provisions from the two regulations both focusing on doors. Given the relatively high similarity score between Sections 4.13.9 of UFAS and 12.5.4.2 of BS8300, they are expected to be related, and in fact they are. Due to the differences in American and British terminologies ("door hardware" versus "door furniture"), a simple concept comparison, i.e., the base score, cannot identify the match between them. However, similarities in neighboring nodes, in particular the parent and siblings, implied a higher similarity between Section 4.13.9 of UFAS and Section 12.5.4.2 of BS8300. This example shows how structural comparison, such as neighbor inclusion, is capable of revealing hidden similarities between provisions, while a traditional term-matching scheme is inferior in this regard.

UFAS
4.1.2 Accessible Buildings: New Construction
(4) Stairs connecting levels that are not connected by an elevator shall comply with 4.9.

Scottish Technical Standards
3.17 Pedestrian Ramps
A ramp must have (a) a width at least the minimum required for the equivalent type of stair in S3.4; and (b) a raised kerb at least 100mm high on any exposed side of a flight or landing, except – a ramp serving a single dwelling.

Figure 18: Related elements "stair" and "ramp" identified

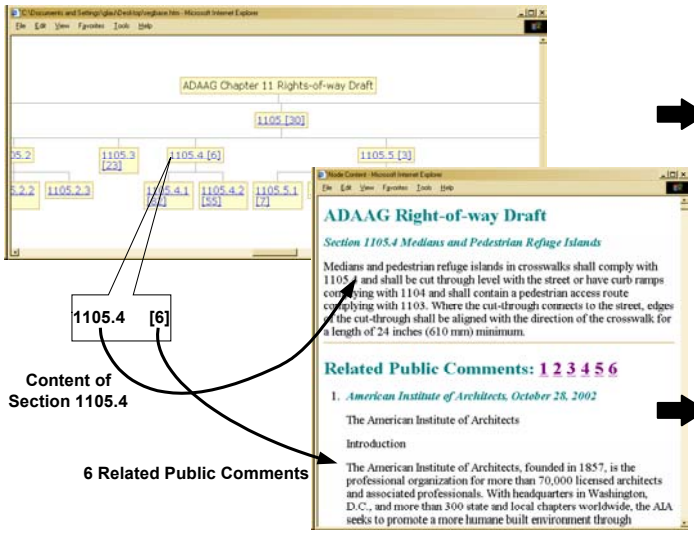
Apart from neighbor inclusion, reference distribution also contributes in revealing hidden similarities between provisions. For instance, as shown in Figure 18, both sections from the UFAS [28] and the Scottish code [27] are concerned about pedestrian ramps and stairs which are related accessible elements. However, even with neighbor inclusion, these two sections show a relatively low similarity score, which is possibly due to the fact that a pure term match does not recognize stairs and ramps as related elements. In this case, after considering reference distribution, these two provisions show a significant increase in similarity based on similar references. Again, this example shows how structural matching, such as reference distribution, is important in revealing hidden similarities which will be otherwise neglected in a traditional term match.

6. APPLICATION ON E-RULEMAKING

Apart from the intended application on comparisons between regulatory documents, we have applied the prototype system to other domains as well, such as electronic-rulemaking. E-rulemaking defines the process in which the electronic media, such as the Internet, is used to provide a better environment for the public to comment on proposed rules and regulations. An example of a real scenario is as follows: the US Access Board recently released a newly drafted chapter [15] for the ADAAG [3], titled "Guidelines for Accessible Public Rights-of-way." This draft is less than 15 pages long. However, over a period of four months, the Board received over 1400 public comments which total around 10 Megabytes in size. Based on the review of these public comments, the Board revises the proposed rules. As a result, the process of e-rulemaking generates a huge amount of data, i.e., the public comments, that needs to be reviewed and analyzed together with the drafted rules.

We applied our system on this domain by comparing the drafted rules with the associated public comments. Figure 19 below shows the generated output, where the drafted regulation appears in its natural tree structure with each node representing sections in the draft. Next to the section number on the node, for example, Section 1105.4, is a bracketed number that shows the number of related public comments identified. Users can follow the link to view the content of the selected section in addition to its retrieved relevant public comments. This prototype shows how a regulatory comparison system can be very useful in an e-rulemaking situation where one needs to review drafted rules based on a large pool of public comments.

Two sample results are observed and presented here. The upper box in Figure 19 represents a typical pair of drafted section and its



ADAAG rights-of-way draft
1105.4.1 Length
Where signal timing is inadequate for full crossing of all traffic lanes or where the crossing is not signalized, …

Public comment
Deborah Wood, October 29, 2002
… This often means walk lights that are so short in duration that by the time a person who is blind realizes they have the light, …

ADAAG rights-of-way draft
No relevant section identified

Public Comment
Donna Ring, September 6, 2002
If you become blind, no amount of electronics on your body or in the environment will make you safe and give back to you your freedom of movement. You have to learn modern blindness skills from a good teacher. …

Figure 19: Application of relatedness analysis on e-rulemaking

identified related public comment. Section 1105.4.1 discusses about inadequate signal timing for pedestrian crossing of traffic lanes, while one of the reviewers complained about the same situation that needs to be dealt with; this illustrates that our system correctly retrieves relevant pairs of drafted section and public comment. It potentially saves rule-makers a tremendous amount of time in reviewing public comments in regard to different provisions among the draft.

The lower box in Figure 19 shows an interesting result in which a particular piece of public comment is not latched with any drafted section. Indeed, this reviewer's opinion is not shared by the draft; she commented on how a visually impaired person should practice "modern blindness skills from a good teacher" instead of relying on government installment of electronic devices on the environment to help. Clearly, the opinion is not shared by the drafted document from the Access Board, which explains why this comment is not related to any provision according to our system.

7. CONCLUSIONS AND FUTURE TASKS

In this paper, we present the development of a legal corpus, its associated similarity analysis, and a compliance assistance framework. A regulation repository is developed using XML as the standard, and our prototype includes several accessibility regulations as well as environmental regulations and supplementary documents. The tree hierarchy of regulations and its referential structure are preserved by properly structuring XML elements. Tools have been developed to extract generic as well as domain-specific feature information which include concepts, measurements, definitions and so on. These features are encapsulated in XML elements whenever they appear in provisions.

An interactive compliance assistance tool is developed by incorporating FOPC logic sentences and control elements to the XML structure. The compliance assistance system guides users through provisions and its implicit references as well as logging the answers for future reference. Relatedness analysis, which combines IR techniques with corpus-specific document structure information, is shown to provide a reliable measure of similarity between pairs of provisions. We show a potential application of our system on the e-rulemaking process to help identify related drafted provisions and public comments. Limitations of our system include mismatches between provisions that use same phrases with different meanings in similarity analysis, and scalability issues that involve vocabulary consolidation in logic implementation for compliance check.

The goal of this research project is to develop an information infrastructure to aid regulation management and understanding in e-government. Due to the existence of multiple sources of regulations and the potential conflicts between them, conflict identification becomes the natural next step to a complete regulatory document analysis. We plan to study the formal representation derived from structured texts to perform an automated analysis of overlaps, completeness and conflicts.

8. ACKNOWLEDGMENTS

This research project is sponsored by the National Science Foundation, Contract Numbers EIA-9983368 and EIA-0085998. The authors would like to acknowledge an equipment grant from Intel Corporation. We would also like to acknowledge the support by Semio Corporation in providing the software for this research.

9. REFERENCES

[1] *Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL 1999)* (Oslo, Norway, 1999). ACM Press, New York, NY, 1999.

[2] *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001)* (St. Louis, Missouri, 2000). ACM Press, New York, NY, 2001.

[3] *Americans with Disabilities Act (ADA) Accessibility Guidelines for Buildings and Facilities*. US Architectural and Transportation Barriers Compliance Board (Access Board), Washington, DC, 1999.

[4] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*. ACM Press, New York, NY, 1999.

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

[6] Berman, D.H., and Hafner, C.D. The Potential of Artificial Intelligence to Help Solve the Crisis in Our Legal System. *Communications of the ACM*, *32*, 8 (1989), 928-938.

[7] Bishop, C. *Neural Networks for Pattern Recognition*. Oxford University Press; Clarendon Press, New York, NY, 1995.

[8] Bollacker, K.D., Lawrence, S., and Giles, C.L. CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. In *Proceedings of the 2nd International Conference on Autonomous Agents* (Minneapolis, MN, 1998). ACM Press, New York, NY, 1998, 116-123.

[9] Brin, S., and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference* (Brisbane, Australia, 1998), 1998, 107-117.

[10] *British Standard 8300*. British Standards Institution (BSI), London, UK, 2001.

[11] *California Building Code (CBC)*. California Building Standards Commission, Sacramento, CA, 1998.

[12] *California Code of Regulations (CCR)*. California Office of Administrative Law, Sacramento, CA, 2003.

[13] *Code of Federal Regulations (CFR)*. US Environmental Protection Agency, Washington, DC, 2002.

[14] Dorre, J., Gerstl, P., and Seiffert, R. Text Mining: Finding Nuggets in Mountains of Textual Data. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, CA, 1999). ACM Press, New York, NY, 1999, 398-401.

[15] *Draft Guidelines for Accessible Public Rights-of-Way*. US Architectural and Transportation Barriers Compliance Board (Access Board), Washington, DC, 2002.

[16] Gibbens, M.P. *California Disabled Accessibility Guidebook 2000*. Builder's Book, Canoga Park, CA, 2000.

[17] Grosjean, J., Plaisant, C., and Bederson, B. SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation. In *Proceedings of IEEE Symposium on Information Visualization* (Boston, MA, 2002). IEEE, Inc., Piscataway, NJ, 2002, 57-64.

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

[19] Kerrigan, S. *A Software Infrastructure for Regulatory Information Management and Compliance Assistance*. Ph.D. Thesis, Stanford University, Stanford, CA, 2003.

[20] Kerrigan, S., and Law, K. Logic-Based Regulation Compliance-Assistance. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Law (ICAIL 2003)* (Edinburgh, Scotland, 2003). ACM Press, New York, NY, 2003, 126-135.

[21] Lau, G., Kerrigan, S., and Law, K. An Information Infrastructure for Government Regulations. In *Proceedings of the 13th Workshop on Information Technology and Systems (WITS'03)* (Seattle, WA, 2003), 2003, 37-42.

[22] Lau, G., Law, K., and Wiederhold, G. Similarity Analysis on Government Regulations. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, 2003). ACM Press, New York, NY, 2003, 111-117.

[23] McCarty, T. Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning. *Harvard Law Review*, *90* (1977), 837-893.

[24] McCune, W.W. *Otter 3.0 Reference Manual and Guide*. Technical Report, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1994.

[25] Salton, G. *The Smart Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[26] *Semio Tagger.* Semio Corporation, 2002. http://www.semio.com.

[27] *Technical Standards*. Scottish Executive, Edinburgh, Scotland, UK, 2001.

[28] *Uniform Federal Accessibility Standards (UFAS)*. US Architectural and Transportation Barriers Compliance Board (Access Board), Washington, DC, 1997.

[29] Wahlgren, P. *Automation of Legal Reasoning*. Kluwer Law and Taxation Publishers, Deventer, the Netherlands, 1992.

[30] Zeleznikow, J., and Hunter, D. *Building Intelligent Legal Information Systems*. Kluwer Law and Taxation Publishers, Deventer, the Netherlands, 1994.