Legal Information Retrieval and Application to E-Rulemaking

Gloria T. Lau
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
glau@stanford.edu

Kincho H. Law
Stanford University
Dept. of Civil & Environmental Eng.
Stanford, CA 94305-4020
law@stanford.edu

Gio Wiederhold
Stanford University
Computer Science Dept.
Stanford, CA 94305-9040
gio@db.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; the latter are often independent of governing bodies. This work aims to develop an information infrastructure for legal information retrieval with applications to electronic-rulemaking. The pilot study focuses on accessibility regulations from the US Federal government, private organizations and European agencies. 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. Handcrafted rules and a text mining tool are developed to extract the important features, such as concepts, measurements, effective dates and so on, and to incorporate them into the corpus.

To compare and locate related provisions from different regulatory documents, we employ Information Retrieval techniques to combine generic features with domain knowledge. Structural information from regulations, such as the hierarchical organization of provisions and heavy referencing among provisions, are used to help improve the relatedness analysis. Results are obtained to illustrate the use of regulatory structure and domain knowledge in provision comparisons. Application to an e-rulemaking scenario for a rights-of-way drafted regulation is shown to demonstrate extended capabilities of the prototype system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models*, I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *law*.

General Terms

Algorithms, Languages, Legal Aspects.

ICAIL '05, June 6-11, 2005, Bologna, Italy.

Keywords

Relatedness Analysis, E-Rulemaking, Regulatory Comparison, Structural Analysis.

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. The regulations, amending provisions and interpretive manuals together create a massive volume of semi-structured documents with potentially similar content but possible differences in format, terminology and context. An information infrastructure that can consolidate, compare and contrast different regulatory documents will greatly enhance and aid the understanding of regulations.

To motivate the problem, Figure 1 shows a classic example of such complexity and conflict found across different regulations [24]. Both Federal and California regulations provide design requirements of a curb ramp; however, the Federal regulation [2] focuses on wheelchair traversal, which is in conflict with the California regulation (this provision is from the 1998 version) [16] 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 analysis is much desired to alert users of related information across multiple sources.

This paper describes a research prototype system that combines text mining and Information Retrieval (IR) techniques to help better understand and analyze regulatory documents. First, some related works on regulatory information retrieval, repository development and similarity analysis are reviewed in Section 2. Section 3 presents the development of a legal corpus with multiple sources of regulatory documents consolidated into a unified format. Extraction of important features, e.g., concepts, measurements, references and so on, is also described in Section 3. Section 4 discusses the work on applying IR and structural matching techniques to perform a relatedness analysis between provisions, with results to illustrate the identification of hidden relatedness of the compared provisions. Potential application of

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

relatedness analysis for aiding the electronic-rulemaking (e-rulemaking) process is shown in Section 5. A brief summary and discussion on future works are given in Section 6.

2. RELATED WORK

Guidance in the interpretation of government regulations has existed as long as regulatory documents. Reference materials and handbooks are merely the byproducts of the many sources of regulatory agencies and the ambiguity of regulations. The example of conflicting provisions shown in Figure 1 is drawn from CalDAG [24], which is one of many reference books written for compliance guidance with the accessibility code in California. The introduction of information technology (IT) to aid regulation exploration follows naturally. For instance, in the US, the Business Gateway[1] project aims to reduce the burden of business by making it easy to find, understand, and comply with relevant laws and regulations [36]. In Europe, many participated in the standardization of legislative texts, in particular, using the eXtensible Markup Language (XML) [7, 8, 21].

The use of available technologies from the field of Artificial Intelligence to aid the understanding of law has been an active research topic for years [38, 52]. The abstraction [33], representation [5, 14], classification [44, 50] and retrieval [1] of case laws are widely studied. Earlier research focused on building expert system for law [46, 52]. Case-based and rule-based systems are developed [10, 17, 39]. In addition, there are many research efforts in applying IR techniques to a legal corpus. Data mining techniques, in particular, text mining algorithms, are sought to perform automated classifications on legal documents. Schweighofer et al. attempted a content-based clustering and labeling of European law, taking into account the importance of different terms [44]. Besides clustering of regulations, work has been done on improving the search experience in a legal corpus. Information extraction techniques are used to aid legal case retrieval based on a "concept" search, where "concepts" are defined to be the headnotes, heading section, case name, court name, judge, etc [34]. A similar approach is used in the SALOMON project that identified and extracted relevant information from case laws, such as keywords and summaries [33]. Finally, natural language search capabilities are supported by various online legal research services such as Westlaw[2].

In repository development, feature extraction is an important step when the data is voluminous. One of the motivations for feature extraction is to avoid the curse of dimensionality [4]. The goal is to reduce data dimensions by including only the important features. It is a form of pre-processing, for example, combining input variables to form a new variable. Often features are constructed by hand based on some understanding of the particular problem being tackled [6]. Automation of this process is also possible. In the field of IR, software tools exist to fulfill "the task of feature extraction … to recognize and classify significant vocabulary items [6]." Taking key phrases as an example of feature, IBM's Intelligent Miner for Text [19] and Semio Tagger [45] are examples of fully automated key phrase extraction tools. Most commercial tools use a combination of linguistic heuristics, pattern matching and lexical analysis for this task.

Text document comparison, in particular, similarity analysis between a user query and documents in a generic corpus, is widely studied. User queries are mostly treated as a pseudo-document containing very few keywords from user input. As a result, determining the similarity between documents and user query (which can be modeled as a short document) can be modeled as document comparisons. Different techniques are developed to compute the match between user queries and documents, such as the Boolean model and the Vector model [41, 43]. Most of these techniques are bag-of-word analyses on the index terms [3]. There are a variety of algorithms to compute index term weights, and a general review can be found in [42]. Our work follows a simple approach, which is to use the count of term appearance as the term weight.

In the relatedness analysis of regulations, we introduce the notion of structural comparisons based on the hierarchical and referential organization of provisions. Among case-based systems, structural mapping is traditionally performed to build mapping of actors and objects in law [23, 31]. Aside from citations used in law, due to the evolution of the World Wide Web, there has been a lot of research work related to academic citation analysis [22]. For instance, CiteSeer is a scientific literature digital library that provides academic publications indexed with their citations [9]. Different types of hyperlink topology and fitting models are examined extensively for different purposes [15, 26, 48]. While Google's PageRank algorithm simulates web surfers' behavior [12, 35], the HITS (Hypertext Induced Topic Search) algorithm exploits the hyperlink structures to locate authorities and hubs on the Internet [28]. In our work, the heavily referenced nature of regulations provides extra information about provisions similar to the link topology of the Web. Our domain is different from the Web - citation analysis assumes a pool of documents citing one another, whereas regulations resemble separate islands of information. Within an island of regulation, provisions are highly referenced; across islands, they are seldom cross-referenced.

[1] The Business Gateway project, a presidential e-government initiative, is formerly called the Business Compliance One-Stop project. The web address for this portal is http://www.business.gov.

[2] Westlaw online legal research service can be accessed at http://www.westlaw.com.

3. DEVELOPMENT OF AN XML REGULATORY REPOSITORY

In order to develop a prototypic system, this work focuses on accessibility regulations, whose intent is to provide the same or equivalent access to a building and its facilities for disabled persons. Our corpus currently includes two US Federal documents: the Americans with Disabilities Act Accessibility Guidelines (ADAAG) [2] and the Uniform Federal Accessibility Standards (UFAS) [51]. In addition, Chapter 11 of the International Building Code [27], titled Accessibility, is included to reflect the similarity and dissimilarity between federal and private agency mandated regulations. Related sections from the British Standard BS8300 [13] and the Scottish Technical Standards [49] are also included for comparisons between American and European regulations. These five documents that we have chosen from the domain of accessibility create a small corpus for prototyping (122,000 words and 1854 provisions). We have also tested our system on a much larger domain of drinking water standards from environmental regulations. Parts from the US code of federal regulations and California state regulations on drinking water are included in the corpus, which totaled 360,000 words and 5310 provisions. In this paper, we will focus on the domain of accessibility.

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. In consolidating regulations into XML format, provisions can be first encapsulated as an XML node. The tree hierarchy of regulations can be captured by properly structuring these XML nodes. Features, including domain-specific information, can be easily added as extra XML elements as well. Figure 2 shows a schematic of the repository development process.

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, 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 is thus 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 organization. For the task of extracting and reconstructing the tree structure in XML format, pattern matching is used. Apart from the tree organization of regulations, the shallow parser extracts referential structures, such as the explicit reference from Section 4.7.4 to Section 4.5 shown in Figure 3, through pattern matching as well. An example of a reference XML tag is shown in Figure 4.

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 [41], 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 4, 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. Concept extraction is performed with the help of the software tool Semio Tagger [45]. Primarily based on co-occurrence relationship of noun phrases and other linguistic analysis techniques, the Tagger identifies a list of noun phrases, or concepts, that are central to the corpus. If we take the ADAAG and the UFAS as an example, they generate just over a thousand concepts together. For other features such as measurements and dates, handcrafted rules are implemented to automatically match them in provisions [30]. The corpus of documents is refined with the extracted features tagged as additional XML elements in provisions where they appear. Figure 4 shows excerpts from a provision and its refined XML version that includes several features such as concept, index term and measurement.

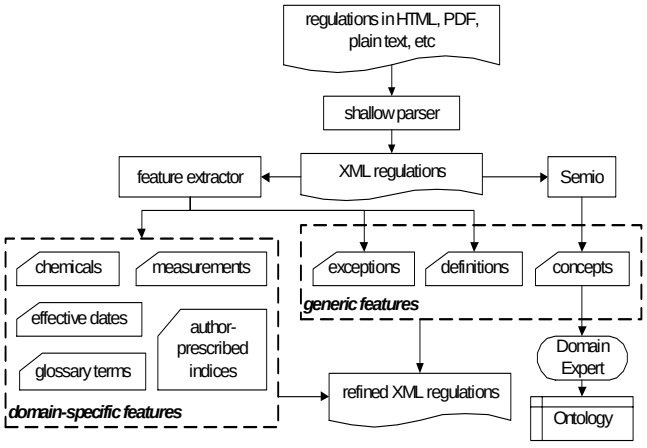


Figure 2: Repository development with feature extraction

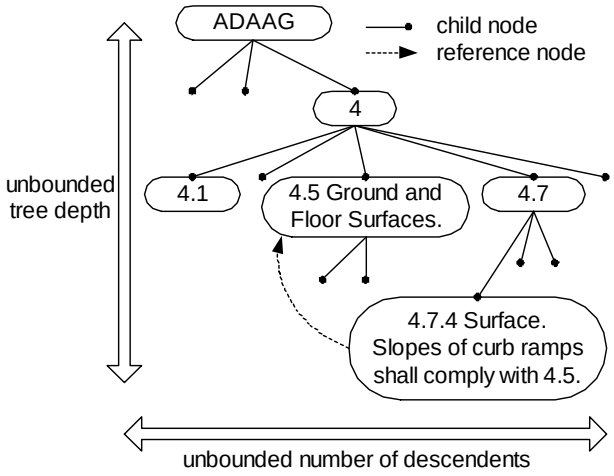


Figure 3: Hierarchical tree structure of regulations

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"
  num="1" />
  <reference name="ufas.4.3" num="1" />
  …
  <regText> Parking spaces for disabled people ...
  </regText>
  <exception> If accessible parking spaces ...
  </exception>
</regElement>
```

Figure 4: Example of XML structure and extracted features

4. RELATEDNESS ANALYSIS OF PROVISIONS IN REGULATIONS

Starting from a well-prepared repository as described in Section 2, 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 a typical regulation is massive in size, a comparison between a full set of regulation and another is meaningless [11]. Instead, a section from one set of regulation is compared with another section from another set, such as a comparison between Section 4.7.2 in ADAAG [2] and Section 1127B.4.4 in CBC [16] as in the example shown in Figure 1. Regulations are represented as trees in the analysis; thus the unit of comparison is pairs of nodes in regulation trees, such as nodes A and U shown in Figure 5. 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. To this end, our system first compares 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 tree hierarchy of regulations and the referential structure in Figure 3. These structures also represent useful information in locating related provisions, and are therefore incorporated into the analysis for a more comprehensive comparison.

A base score is first computed between two provisions by matching extracted features such as those shown in Figure 2. This allows for a combination of generic features, such as concepts, as well as domain knowledge, such as measurements in accessibility regulations. This design provides the flexibility to add on features and different feature 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

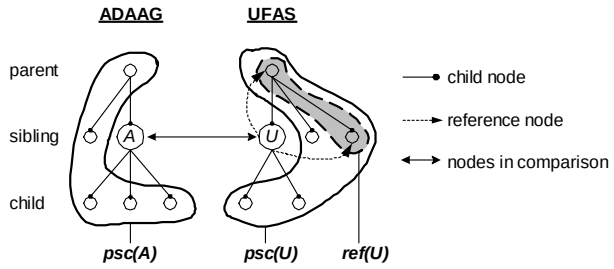


Figure 5: Immediate neighboring nodes and referenced nodes in regulation trees

sections based on that particular feature. For instance, concept matching is done similarly to the index term matching in the Vector model [41], where the degree of similarity of documents is evaluated as the correlation between their index term vectors. Under the Vector model, a cosine similarity between the two concept vectors would represent the degree of similarity between the two provisions based on a concept match. Scoring schemes for other features are developed using the same idea.

Some features are associated with ontologies to define synonyms, which cannot always be modeled as Boolean term matches. As an example, a domain expert can potentially define a measurement of "12 inches maximum" as 75% similar to a measurement of "12 inches." Here, domain knowledge results in a non-Boolean index, or a soft index, which resembles the idea of a dimension that contributes to weaknesses and strengths of a legal case [37]. Therefore, these feature vectors are mapped onto a different vector space before comparison to account for synonyms and non-Boolean matching [29].

The base score is subsequently refined by utilizing the tree structure of regulations. The parent, siblings and children of the interested sections are compared to include similarities between the interested sections that are not previously accounted for based on a direct comparison. Referring to Figure 5, the immediate neighbors of provision A, i.e., the parent, siblings and children, are collectively termed the psc(A) of node A. 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 each other. Two sections referencing similar sections are more likely to be related and should have their similarity score raised. The process of reference distribution essentially utilizes the heavily self-referenced structure of the regulation to further refine the similarity score. Figure 5 shows the out-references from provision A as the ref(A) of node A. Taking Section A from the ADAAG [2] and Section U from the UFAS [51] as an example, psc(A) is compared to psc(U) as well as ref(A) versus ref(U) in score refinements. After successive refinements, similarities from both near-tree neighbors and references are identified, and related provisions are retrieved based on the resulting scores.

Preliminary results obtained from the comparisons between different regulations are documented in [29]. A user survey is conducted to rank the similarity of ten randomly chosen

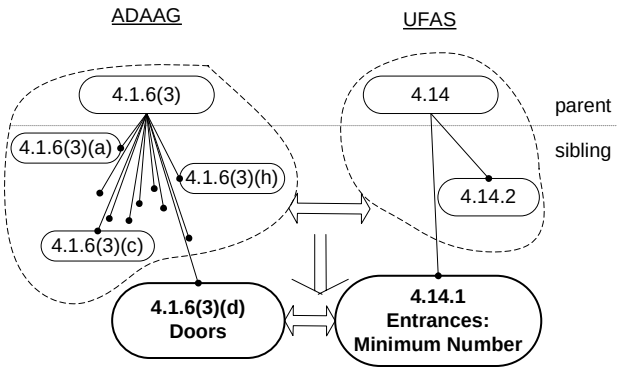
provisions from the ADAAG [2] and ten from the UFAS [51]. The relatedness analysis system is compared with Latent Semantic Indexing (LSI) [18], as LSI claims to reduce the dimension of term space into concept space based on Singular Value Decomposition (SVD) [25], which shares a similar goal as our feature extraction. The Root Mean Square Error (RMSE) is used to compute the ranking prediction error based on the survey results as the "correct" answer. Overall, our system outperforms the LSI with RMSE of 22.9 and 27.4 respectively. Individual combinations of features and structural matching produce prediction errors ranging from 12.0 to 29.1; majority of which are smaller than the error produced by a LSI implementation. Among the features implemented in an accessibility domain, such as concepts, measurements and author-prescribed indices, the use of measurement features results in far reduced errors such as 12.0. This reinforces our belief in domain knowledge, especially in this case, when both the ADAAG and the UFAS prescribe heavily quantified requirements that can only be captured by measurement features.

On the other hand, structural matching does not seem to affect the error in any noticeable trend. This is possibly due to the fact that the ten randomly selected pairs of provisions happen to be not very much referenced – the ref(·) operation returned mostly empty sets. Another explanation is that the "correct" answers do not make use of the structures either. The users are not given with much contextual (psc nodes) and referential (ref nodes) information in the survey for a complete understanding of the two regulations in comparison. Since this survey is only conducted using accessibility regulations, it is difficult to generalize the results to claim that the use of domain knowledge produces superior results compared to analysis performed without domain knowledge in other domains. However, the results do indicate that domain knowledge has its values in enhancing the understanding of provisions, as is apparent in the domain of accessibility based on the survey.

To justify for the proposed score refinements, we compare results obtained using the base score with results from neighbor inclusion and reference distribution. The first example shown in Figure 6 illustrates the use of neighbor inclusion, where we compare the base score with the refined score. Here, Section 4.1.6(3)(d) in the ADAAG [2] is concerned with doors, while Section 4.14.1 in the UFAS [51] deals with entrances. As expected, a pure concept match could not identify the relatedness between door and entrance, thus resulting in a zero base score. However, with non-zero similarities between their psc nodes, the system is able to infer some relatedness between the two sections from their neighbors in the tree. The related accessible elements, namely door and entrance[3], are identified indirectly through neighbor inclusions.

To illustrate the similarity between American and British standards, we compare the UFAS [51] with the BS8300 [13]. Figure 7 shows 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 illustrates 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.

Apart from neighbor inclusion, reference distribution also contributes to revealing hidden similarities between provisions. For instance, as shown in Figure 8, both sections from the UFAS [51] and the Scottish code [49] 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 out-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.



ADAAG

4.1.6(3)(d) ***Doors***

(i) Where it is technically infeasible to comply with clear opening width requirements of 4.13.5, a projection of 5/8 in maximum will be permitted for the latch side stop. (ii) If existing thresholds are …

UFAS

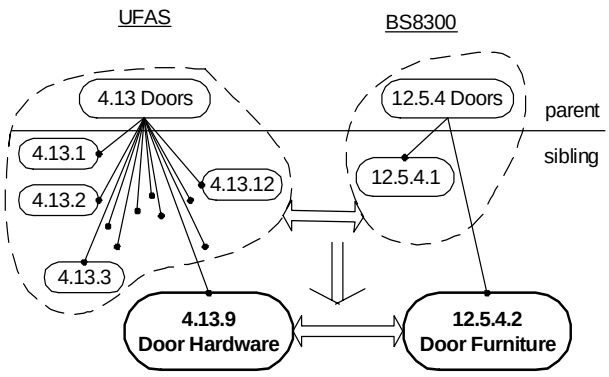
4.14 Entrances

4.14.1 Minimum Number

Entrances required to be accessible by 4.1 shall be part of an accessible route and shall comply with 4.3. Such entrances shall be connected by an accessible route to public transportation …

Figure 6: Related provisions identified through neighbor inclusion

[3] Definitions of "*door*" in WordNet [32] include "the *entrance* (the space in a wall) through which you enter or leave a room or building" and "a swinging or sliding barrier that will close the *entrance* to a room or building or vehicle."



UFAS
4.13 Doors
4.13.1 General
…
4.13.9 ***Door Hardware***
Handles, pulls, latches, locks, and other operating devices on accessible doors shall have a shape that is easy to grasp with one hand and does not require tight grasping, tight pinching, or twisting of the wrist to operate …
…
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 in accessible bedrooms should be easy to grip and operate by a wheelchair user or ambulant disabled person (see 6.5). Handles fixed to hinged and sliding doors of furniture and fittings in bedrooms should be easy to grip and …

Figure 7: Example of a similarity analysis between American and British regulations

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 8: Related elements stair and ramp identified

5. APPLICATION TO 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 e-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 recent scenario is as follows: the US Access Board released a newly drafted chapter [20] for the ADAAG, 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.

The relatedness analysis framework compares each provision from the drafted chapter with each of the 1,400 public comments. To compare provisions with comments, a similarity score is computed per pairs of provisions and comments based on the computational properties, including feature matching and structural matching as defined in the previous section. The results of a relatedness analysis are related pairs between the provision from the draft and individual comments. Figure 9 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 browse through the tree of drafted provisions, and follow the links to view the content of the selected provision along with its retrieved relevant public comments. This prototype shows how a regulatory comparison system can help improve the e-rulemaking process 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 9 represents a typical pair of drafted section and its identified public comment. Section 1105.4.1 discusses about inadequate signal timing for pedestrian crossing of traffic lanes, while one of the reviewers complains 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. Another observation from this example is that a full content comparison between provisions and comments is necessary, since title phrases, such as "length" in this case, are not always illustrative of the content. Automation is needed as it would otherwise require a lot of human effort to perform a full content comparison for the large number of comments.

A different type of comment screening is shown in the lower box in Figure 9. It is an even more 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. This reviewer commented on how a visually impaired person should practice "modern blindness skills from a good teacher" instead of relying on government installed electronic devices on streets to help. This opinion is not represented in the drafted document from the Access Board, which explains why this comment is not related to any provision according to the relatedness analysis system. As shown in the two examples, by segmenting the pool of comments according to their relevance to individual provisions, the relatedness analysis can potentially save rule makers significant amount of time in reviewing public comments in regard to different provisions in the drafted regulations.

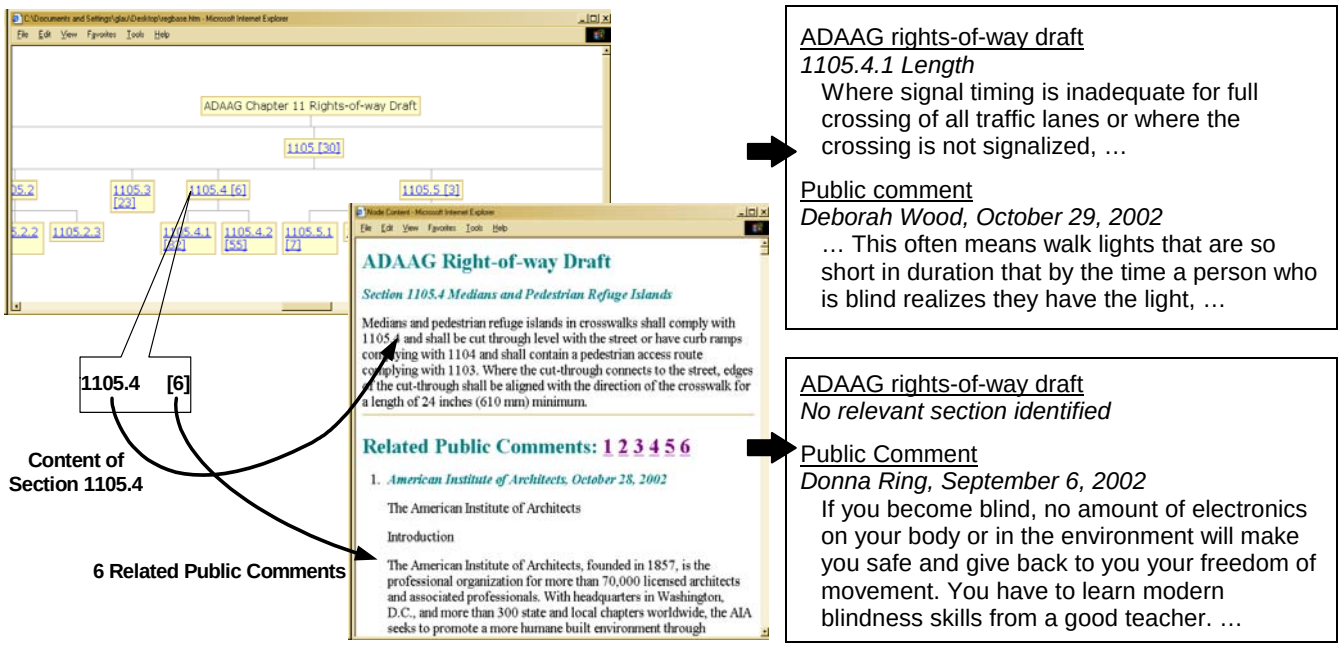


Figure 9: Application of similarity analysis to e-rulemaking

6. CONCLUSIONS AND FUTURE TASKS

This paper has presented the development of a legal corpus, its associated relatedness analysis with an application to e-rulemaking. A regulation repository is developed using XML as the standard, and our prototype includes several accessibility regulations. The tree hierarchy of regulations and its referential structure are preserved by properly structuring XML elements. Tools have been developed for extracting generic as well as domain-specific features which include concepts, measurements, effective dates and so on. These features are encapsulated in XML elements whenever they appear in provisions. Relatedness analysis combines domain knowledge with corpus-specific document structural information, such as provision hierarchy and inter-section referencing. It is shown to provide a reliable measure of similarity between pairs of provisions, based on their shared features, neighbors or references. Potential application of our system to the e-rulemaking process is demonstrated to help identify related drafted provisions and public comments.

Limitations of the current prototype system include mismatches between provisions that use same phrases with different meanings in relatedness analysis. There are also provisions written using different terminologies where our existing features and structural analysis would fail to capture their relatedness. Different linkages and citation signals used in law might help to improve the system; for instance, Shepardizing is standard practice in legal research where cases and statutes are validated through previous citations [47]. Links between legal theories and cases are traced for retrieval tasks [40]. If we include cases in our corpus, citations from cases to provisions can potential help to identify related provisions. Case citations can then be incorporated into the computation analogous to reference distribution. In an e-rulemaking application, we also observed that a comparison between provisions and comments might not be enough. Sometimes, there are comments that are not directly related to any provision in the draft; instead, commenters tend to support another organization's position on the general direction and intent of the draft. Clustering of comments with external documents and references can potentially help classify this type of opinions.

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.

7. 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.

8. REFERENCES

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

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

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

[4] Bellman, R.E. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

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

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

[7] Boer, A., Hoekstra, R., and Winkels, R. METALex: Legislation in XML. In *Proceedings of Jurix 2002: 15th Annual International Conference on Legal Knowledge and Information Systems* (London, UK, 2002). IOS Press, 2002, 1-10.

[8] Bolioli, A., Dini, L., Mercatali, P., and Romano, F. For the Automated Mark-Up of Italian Legislative Texts in XML. In *Proceedings of Jurix 2002: 15th Annual International Conference on Legal Knowledge and Information Systems* (London, UK, 2002). ISO Press, 2002, 21-30.

[9] 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, 1998, 116-123.

[10] Branting, L.K. Building Explanations from Rules and Structured Cases. *International Journal of Man-Machine Studies*, *34*, 6 (1991), 797-837.

[11] Branting, L.K. Reasoning with Portions of Precedents. In *Proceedings of the 3rd International Conference on Artificial Intelligence and Law (ICAIL 1991)* (Oxford, England, 1991). ACM Press, 1991, 145-154.

[12] 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.

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

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

[15] Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E., and Silva, I. Local versus Global Link Information in the Web. *ACM Transactions on Information Systems (TOIS)*, *21*, 1 (2003), 42 - 63.

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

[17] Daniels, J.J., and Rissland, E.L. What You Saw Is What You Want: Using Cases to Seed Information Retrieval. In *Proceedings of the 2nd International Conference on Case-Based Reasoning (ICCBR-97)* (Providence, RI, 1997), 1997, 325-336.

[18] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, *41*, 6 (1990), 391-407.

[19] Dörre, 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, 1999, 398-401.

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

[21] Engers, T.M.v., and Vanlerberghe, R.A.W. The POWER-Light Version: Improving Legal Quality under Time Pressure. In *Proceedings of EGOV 2002: the 1st International Conference on Electronic Government* (Aix-en-Provence, France, 2002), 2002, 75-83.

[22] Garfield, E. New International Professional Society Signals the Maturing of Scientometrics and Informetrics. *The Scientist*, *9*, 16 (1995).

[23] Gentner, D., and Markman, A.B. Structure Mapping in Analogy and Similarity. *American Psychologist*, *52*, 1 (1997), 45-56.

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

[25] Golub, G.H., and Van Loan, C.F. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, 1983.

[26] Gurrin, C., and Smeaton, A.F. A Connectivity Analysis Approach to Increasing Precision in Retrieval from Hyperlinked Documents. In *Proceedings of Text REtrieval Conference (TREC)* (Gaithersburg, MD, 1999), 1999.

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

[28] Kleinberg, J. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms* (San Francisco, CA, 1998), 1998, 668-677.

[29] Lau, G. *A Comparative Analysis Framework for Semi-Structured Documents, with Applications to Government Regulations*. Ph.D. Thesis, Stanford University, Stanford, CA, 2004.

[30] 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, 2003, 111-117.

[31] McLaren, B.M. Extensionally Defining Principles and Cases in Ethics: an AI Model. *Artificial Intelligence*, *150*, 1-2 (2003), 145-181.

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

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

[34] Osborn, J., and Sterling, L. JUSTICE: A Judicial Search Tool Using Intelligent Concept Extraction. In *Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL 1999)* (Oslo, Norway, 1999), 1999, 173-181.

[35] Page, L., Brin, S., Motwani, R., and Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report, Stanford University, Stanford, CA, 1998.

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

[37] Rissland, E.L. Dimension-Based Analysis of Hypotheticals from Supreme Court Oral Argument. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Law (ICAIL 1989)* (Vancouver, Canada, 1989). ACM Press, 1989, 111-120.

[38] Rissland, E.L., Ashley, K.D., and Loui, R.P. AI and Law: A Fruitful Synergy. *Artificial Intelligence*, *150*, 1-2 (2003), 1-15.

[39] Rissland, E.L., and Skalak, D.B. CABARET: Rule Interpretation in a Hybrid Architecture. *International Journal of Man-Machine Studies*, *34*, 6 (1991), 839-887.

[40] Rissland, E.L., Skalak, D.B., and Friedman, M.T. BankXX: A Program to Generate Argument Through Case-Base Research. In *Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL 1993)* (Amsterdam, The Netherlands, 1993). ACM Press, 1993, 117-124.

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

[42] Salton, G., and Buckley, C. Term-Weighting Approaches in Automatic Retrieval. *Information Processing and Management*, *24*, 5 (1988), 513-523.

[43] Salton, G., and McGill, M. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.

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

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

[46] Sergot, M.J., Sadri, F., Kowalski, R.A., Kriwaczek, F., Hammond, P., and Cory, H.T. The British Nationality Act as a Logic Program. *Communications of the ACM*, *29*, 5 (1986), 370-386.

[47] *Shepard's Federal Citations*. Shepards/Mcgraw-Hill, Colorado Springs, CO, 1990.

[48] Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., and Ziviani, N. Link-Based and Content-Based Evidential Information in a Belief Network Model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece, 2000), 2000, 96-103.

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

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

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

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