A regulatory information infra-structure with application to accessibility codes

Gloria T Lau and Kincho H Law
Stanford University, USA
glau@stanford.edu law@stanford.edu

Bimal Kumar
Glasgow Caledonian University, UK
B.Kumar@gcal.ac.uk

Keywords
Disabled access; information retrieval; similarity analysis and conflict .

Abstract
This paper describes a research project that addresses the difficulties in dealing with regulatory documents such as national and regional codes. These documents tend to be voluminous, heavily cross-referenced, possibly ambiguous and even conflicting at times. There are often multiple documents that need to be consulted and satisfied; however it is a difficult task to locate all of the relevant provisions, and also sections dealing with the same or similar conceptual ideas sometimes lay down conflicting requirements. We propose a framework for regulation representation, analysis and comparison with emphasis on the extraction of similarities between provisions. We focus on accessibility regulations, whose intent is to provide the same or equivalent access to a building and its facilities for disabled persons.

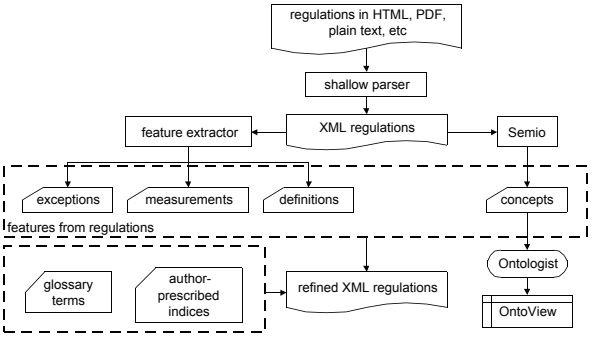
Introduction
Government regulations are an important asset of the society; ideally they should be readily available and retrievable by the general public. Curious citizens are entitled to and thus should be provided with the means to better understand government regulations. However, to locate a particular regulation of interest is no easy matter: neither a generic bookstore nor a public library has a good chance of possessing such a regulation. In addition, even with the hardcopy in hand, an average citizen is likely to soon get lost in the jungle of definitions of legal terms and references.

Besides the difficulties in locating and understanding a particular regulation, the inherent nature of multiple coding agencies also deserves attention. Regulations are typically specified by Federal as well as State governmental bodies and are amended and regulated by local counties or cities. These multiple sources of regulations sometimes compliment and modify each other, while at times the provisions of two applicable codes are in direct conflict. Thus there is a need for a tool to compare and contrast regulations from different sources, with possible differences in formatting,

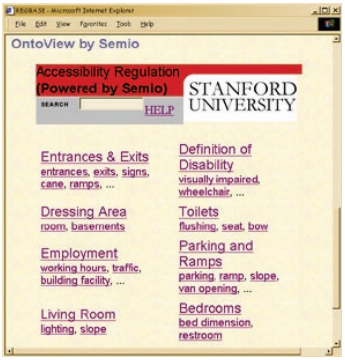
terminology and context, that regulate over the same topic.

The advance in technology has provided some tools to mitigate some of the above problems. It is now easier to locate and search through a particular regulation; for instance, the Americans with Disabilities Act Accessibility Guidelines (ADAAG), (The Access Board, 1998) and the Uniform Federal Accessibility Standards (UFAS), (The Access Board, 1996) are both available online in Hypertext Markup Language (HTML). With the recent development of the eXtensible Markup Language (XML), one can consolidate different regulations and represent codes in a semi-structured format, which is briefly discussed in Section 2. To enhance understanding of regulations, textual comparison techniques from the field of Information Retrieval can be deployed; details of the proposed similarity analysis technique are given in Section 3. Tentative results are presented in Section 4. A brief summary and discussion of future work are described in Section 5. The purpose of this paper is to demonstrate that information technology can help solve some of the problems in locating, reasoning and comparing different sets of regulations.

Repository development



1. (a) Repository Development Schematic



(b) OntoView by Semio

The first phase of the project is to develop an online repository of regulatory documents, and a schematic of the process is shown in Figure 1a. The key issue here is to extend the usability of digital information: a standard format for interoperable information exchange is needed to consolidate regulations from different places. XML is chosen as the communication model because of its expressiveness to represent the organization of provisions and its ability to format semi-structured data; for instance, the tree hierarchy of regulations can be captured by properly structuring XML tags. Repository development starts with a shallow parser to consolidate different formats of regulations into XML format. This is necessary since regulations originally come in different formats, e.g. our corpus currently includes the HTML-originated ADAAG and UFAS, both of which are Federal documents that provide prescriptive measures on disabled access in the US. To illustrate the difference between Federal codes and codes of practice in design, Chapter 11, Accessibility, of the International Building Code (International conference of building officials, 2000) is also included in our corpus. Further, to show the similarity between regulations across continents, the British Standard BS8300 (British Standards institution, 2001) and Section S from the Scottish Technical Standards (Scottish executive, 2001) are incorporated as well.

After the documents are parsed into XML format, Semio Tagger, a commercial tool for text categorization, is used to help develop a taxonomy for regulations. A list of important noun phrases, or concepts in Semio terms, is first identified by the Tagger, with a knowledge engineer to create a taxonomy accordingly. Figure 1b

a taxonomy generated using Semio where users can click through the structure to view relevant provisions classified based on concepts. Here the repository is complete for provision retrieval after data cleaning, consolidation and categorization. However, features, or evidences, that signal related or similar sections need to be extracted before we can start to analyze and compare provisions.

Feature extraction

Feature extraction may be best explained using examples; here excerpts from the original provision are shown followed by the complete set of XML mark-ups. In the first example, features like <concept>, <indexTerm>, <measurement>, <ref> and <exception> are extracted and illustrated. In the second example, a <definition> tag is shown.

The first feature in Example 1 is <concept>, or key phrase extracted from the corpus by Semio Tagger. Concepts help to identify similarity and conflicts due to its ability to capture sequencing information from phrases and its simplicity compared to traditional index terms. Another source of potentially important terms comes from author-prescribed indices (<indexTerm>) at the back of reference books or even the regulation itself; human-written information such as this can sometimes be more valuable than machine-generated phrases. Particular to accessibility provisions, measurements (<measurement>) play a very important role in which they define most of the conflicts, e.g. one provision might ask for a clear width of 10 to 12 inches, while another one might require 13 inches. Therefore, it is crucial to identify measurements and the associated quantifiers if there is any. Finally, out references to other provisions and exceptions amending the body text of the provision are extracted and captured correspondingly in the <ref> and <exception> tags as shown below.

Example 1
Original provision from the UFAS
4.6.3 PARKING SPACES. Parking spaces for disabled people shall be at least 96 in...
EXCEPTION: If accessible parking spaces for vans designed for handicapped persons...
Refined provision in XML format

```
<regElement name="ufas.4.6.3" title="parking spaces">
        <concept name="access aisle" num="3" />
        <indexTerm name="accessible circulation route" num="1" />
        <measurement unit="inch" size="96" quantifier="min" />
        <ref name="ufas.4.5" num="1" />
        <regText> Parking spaces for disabled people... </regText>
        <exception> If accessible parking spaces for... </exception></regElement>
```

The second example shows a common practice in regulations: a designated section in an early chapter that defines the important terminologies used in the code, such as Section 3.5 in the ADAAG. Again, these human-generated terms (<definition>) are more likely to convey key concepts than machine extracted ones such as Semio concepts; also, the definition of a term gives the meaning to a term, which is useful in comparison. In addition, although not shown in the examples here, engineering handbooks also define the important glossary terms (<glossaryDef>) used in the field. The primary difference between <definition> and <glossaryDef> is that <definition> comes from the regulation itself, while <glossaryDef> comes from sources other than the regulation.

Example 2
Original provision from the ADAAG
3.5 DEFINITIONS. …
ACCESSIBLE
Describes a site, building, facility, or portion thereof that complies with these guidelines. …
Refined provision in XML format

```
<regElement name="adaag.3.5" title="definitions" asterisk="0">
        <definition>
                <term> accessible </term>
                <definedAs> Describes a site, building, facility, or portion thereof
                that complies with these guidelines. </definedAs>
        </definition> </regElement>
```

Similarity analysis

With the repository fully developed, users can browse and search through the regulations easily. However, upon finding a relevant provision, to search through multiple codes with multiple terms to locate related provisions, if there is any, is more difficult. Our goal is to provide a reliable measure of relatedness of pairs of provisions, and to recommend similar sections of a selected provision based on the similarity measure. Here since a typical regulation can easily go over thousands of pages, we do not attempt to compare a full set of regulation against one another; rather a section from one set of regulation is compared with another section from another set, such as a comparison between section 4.3(a) in ADAAG and section 3.12 in UFAS.

The comparison core takes two regulations and produces a similarity score, denoted by f Œ (0, 1). It first computes a base score f0 for each pair of provisions by matching extracted features. The scoring scheme for each of the features essentially reflects how much resemblance can be inferred between the pair of sections based on that particular feature. For instance, concept matching is done exactly like index term matching in the vector model (G Salton, 1971), where the degree of similarity of documents is evaluated as the correlation between their index term vectors that represent the weights for each index term in the document. For two sections, the similarity score fco obtained by comparing concepts is given by the cosine similarity between the two concept vectors. Since cosine similarity is normalized, it always produces a score between 0 and 1. Scoring schemes for other features follow the same idea.

After obtaining the initial base score between pairs of sections, it is subsequently refined by taking into account the hierarchical organization of regulations as a tree structure. In the score refinement process, we use the term 'psc' collectively to denote the parent, siblings and children of a provision, i.e. its immediate neighbours. For example, let's take two sections A and B for discussion. To refine the score f0(A, B), section A itself is first compared with psc(B), and vice versa, to produce the score fs-psc. The next refinement takes into account the comparison between psc(A) and psc(B), which gives the score fpsc-psc. Therefore, the similarities in immediate neighbouring sections that are not apparent from (A, B) are identified.

The final refinement frd comes from the not-so-immediate surroundings of nodes A and B through a process called Reference Distribution. The intuition behind reference distribution is to take into account that regulations are heavily self-referenced and cross-referenced documents, which contribute to the difficulty in reading and understanding them. Here the hypothesis is that two sections referencing similar

sections are more likely to be related and should have their similarity score raised; in other words, the referencing structure of regulatory documents is used to further refine similarity scores. Therefore, after successive score refinements, similarities from both near-tree neighbours and not-so-immediate surroundings are accounted for, and a stable ranking of the most related sections is produced as a result.

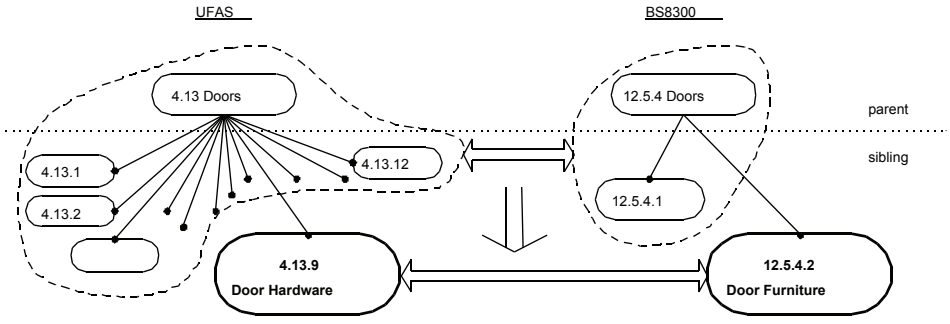
Results and examples

To assess the performance of our system, sections from different regulations are randomly selected for comparison. First, to justify for score refinements and neighbour inclusions within our system, results from f0 and fs-psc are compared and some improvement is observed. For instance, example 3 below shows that section 4.1.6(3)(d) in ADAAG concerns about door, while section 4.14.1 in UFAS deals with entrance. As expected, concept match in f0 could not identify the similarity between door and entrance, thus f0 = 0. With fs-psc, the system is able to infer some relatedness between the two sections from the neighbours in the tree, and thus results in a nonzero score for fs-psc.

Example 3
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 3/4 in high or less, and have (or are modified to have) a bevelled edge on each side, they may remain.
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 stops, to accessible parking and passenger loading zones, and to public streets or sidewalks if available (see 4.3.2(1)). They shall also be connected by an accessible route to all accessible spaces or elements within the building or facility.

To illustrate the similarity between American and British standards, UFAS is compared with BS8300. Example 4 below shows sections from the UFAS and BS8300 both focusing on doors. Given the relatively high similarity score between sections 4.13.9 and 12.5.4.2 (f0 = 0.425), they are expected to be related, and in fact they are; section 4.13.9 from the American code is titled Door Hardware while section 12.5.4.2 from the British standard is titled Door Furniture. As the American and British phrasing is different, concept comparison does not pick up the match between 'door hardware' and 'door furniture'; however, by comparing the neighbours of the sections, a higher similarity score is observed (fpsc-psc = 0.471). As explained in Figure 2, similarities in neighbouring nodes in the regulation trees imply a higher similarity between the compared sections 4.13.9 and 12.5.4.2.

Example 4
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. Lever-operated mechanisms, push-type mechanisms, and U-shaped handles are acceptable designs. When sliding doors are fully open, operating hardware shall be exposed and usable from both sides. In dwelling units, only doors at accessible entrances to the unit itself shall comply with the requirements of this paragraph. Doors to hazardous areas shall have hardware complying with 4.29.3. Mount no hardware required for accessible door passage higher than 48 in (1220 mm) above finished floor.
…
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 manipulate. They should conform to the recommendations in 6.5 for dimensions and location, and the minimum force required to manipulate them.
Consideration should be given to the use of electronic card-activated locks and electrically powered openers for bedroom entrance doors.
COMMENTARY ON 12.5.4.2. Disabled people with a weak hand grip or poor co-ordination, find that using a card to open a door lock is easier than turning a key. A wide angle viewer should be provided in doors to accessible bedrooms at two heights, 1050 mm and 1500 mm above floor level to allow viewing by a person from a seated position and a person standing.
Door furniture should contrast in colour and luminance with the door.



2. Score refinement based on neighboring nodes in tree

Summary and future works

This project aims to develop an information infrastructure for regulation management and comparative analysis. A repository is built by transforming regulations into XML format because of its capability to handle semi-structured data. After all regulations are in a unified format, features, or evidences, are extracted from the set of regulations automatically or by a knowledge engineer, in addition to features from reference materials such as engineering handbooks. A taxonomy is developed on top of the concepts identified by an information retrieval tool, such as Semio, to enable categorized document retrieval following the hierarchy.

To allow for easy retrieval of relevant sections in regulations, a similarity analysis is performed on the documents. The similarity analysis core first computes a base score between pairs of provisions by combining similarity scores from each of the features. The base score is refined by taking into account neighbouring sections. Reference distribution is performed to further tune the scores according to the reference structure in the regulations. A list of the most related sections is produced as a result. Preliminary results are obtained and examples show that our system is capable of deducing similarities from the structure of the regulations. We also anticipate the development of conflict analysis in which both feature extraction from repository development and similarity analysis have laid its groundwork. The proposed conflict analysis core will identify potential conflicts among the most related sections, assuming that sections must be related to have conflicting ideas over certain subject areas, and thus the process of similarity analysis.

Acknowledgments

This research project is sponsored by the National Science Foundation, Contract Numbers EIA-9983368 and EIA-0085998. The authors would like to acknowledge a Technology for Education 2000 equipment grant from Intel Corporation. We would also like to acknowledge the support by Semio Corporation in providing the software for this research.

References

The Access Board (1998) ADA Accessibility Guidelines for Buildings and Facilities

The Access Board (1986) Uniform Federal Accessibility Standards (UFAS)

British Standards Institution (2001) British Standard BS8300

International Conference of Building Officials (2000) International Building Code 2000

Salton G (1971) The Smart Retrieval System - Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice Hall Inc

Scottish Executive (2001) Scottish Technical Standards