A Relatedness Analysis of Government Regulations using Domain Knowledge and Structural Organization

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

Abstract

The complexity and diversity of government regulations make understanding and retrieval of regulations a non-trivial task. One of the issues is the existence of multiple sources of regulations and interpretive guides with differences in format, terminology and context. This paper describes a comparative analysis scheme developed to help retrieval of related provisions from different regulatory documents. Specifically, the goal is to identify the most strongly related provisions between regulations. The relatedness analysis makes use of not only traditional term match but also a combination of feature matches, and not only content comparison but also structural analysis.

Regulations are first compared based on conceptual information as well as domain knowledge through feature matching. Regulations also possess specific organizational structures, such as a tree hierarchy of provisions and heavy referencing between provisions. These structures represent useful information in locating related provisions, and are therefore exploited in the comparison of regulations for completeness. System performance is evaluated by comparing a similarity ranking produced by users with the machine-predicted ranking. Ranking produced by the relatedness analysis system shows a reduction in error compared to that of Latent Semantic Indexing. Various pairs of regulations are compared and the results are analyzed along with observations based on different feature usages. An example of an e-rulemaking scenario is shown to demonstrate capabilities and limitations of the prototype relatedness analysis system.

Keywords

Relatedness analysis, e-rulemaking, structural analysis, feature matching

1 Introduction

Government regulations are an important asset of the society. They extend the laws governing the country with specific guidance for corporate and public actions. Ideally regulations should be readily available and retrievable by the general public. However, the extensive volume of regulations, heavy referencing between provisions and non-trivial definitions of legal terminologies hinder public understanding of the regulations. Besides the difficulties in locating and understanding a particular regulation, the existence of multiple jurisdictions means that often many documents need to be consulted and their provisions satisfied. Sections dealing with the same or similar conceptual ideas sometimes impose conflicting requirements by different jurisdictions. Hence, it is a difficult task to locate all of the relevant provisions.

In the United States, government regulations are typically specified by Federal as well as State governmental bodies and are amended and regulated by local counties or cities. In addition, non-profit organizations sometimes publish codes of practice. These multiple sources of regulations tend to complement and modify each other; at times, the provisions of two applicable codes are in direct conflict. The regulations, amending provisions and interpretive manuals together create a massive volume of semi-structured documents with possible differences in formatting, terminology and context. This results in a loss of productivity and efficiency, and the identified problem is not confined to the US. Rissland et al. (Rissland et al. 2003) observed that in the European Union there is a great need for sharing and reusing of knowledge to harmonize legislation across the polyglot countries. The problem is manifested in multinational companies who must comply with multiple jurisdictions across continents (Bender 2004, Raskopf and Bender 2003). A survey revealed that “widely divergent legal restrictions present a growing obstacle to multinational companies. … A surprisingly large amount of companies are still “solving” this problem by ignoring it (Raskopf and Bender 2003).”

1.1 An Example of the Complexity of Regulations

The following example, drawn from an interpretive guidebook for California accessibility regulations (Gibbens 2000), will put the above-described complexity into context. In the domain of accessibility regulations, Gibbens documented several “controversial issues between the [California] state and federal guidelines.” There are instances where one regulation is less restrictive than another. There are cases where two provisions are in direct conflict; for instance, Figure 1 shows an example from the California Building Code (*CBC* 1998) and the Americans with Disabilities Act Accessibility Guidelines (*ADAAG* 1999). The conflict is due to the fact that the intents of the California and Federal codes are different – the California code (this provision is from the 1998 version) addresses the mobility of the visually impaired when using a cane, while the Federal standard focuses on wheelchair traversal. Gibbens pointed out that “when a state or local agency requires you to construct the California required ½ inch beveled lip, they are requiring you to break the federal law,” and this clearly deserves attentions from industry planners, designers and affected individuals.

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 Fig. 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.4.4 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: Example of Two Conflicting Provisions

The above example illustrates that it is indeed a non-trivial task to search through multiple codes with multiple terms to locate related provisions, if there is any. Nonetheless, it is crucial to identify as much relevant information as possible, since the cost of missing relevant information is growing in the legal system (Berman and Hafner 1989). There is a need for an analysis tool to provide a reliable measure of relatedness among pairs of provisions, and to recommend similar sections of a selected provision based on a similarity measure.

1.2 Computational Properties of Regulations

It is worth noting that legal documents are different from typical documents found in generic free-form text corpora. Any form of analysis on a generic free-form text corpus requires deep understanding of the underlying computational properties of language structure, which is often difficult and possibly subjective. However, focusing on a semi-structured text corpus reduces the problem to a more tangible one. Regulatory documents possess three main characteristics that are not found in generic text, which makes them interesting to analyze.

- Regulations assume a deep tree hierarchy as illustrated in Figure 2. They are semi-structured documents that are organized into a tree structure; for example, Section 4.7.4 can be interpreted as a subpart or a child node of Section 4.7, which makes it a sibling of Section 4.7.3 as well. This regulatory structure is crucial in understanding contextual information between sections.

- Sections are heavily cross-referenced within one regulation. For instance, Section 4.7.4 can refer to Section 4.5 for compliance requirements under other conditions. In analyzing and comparing provisions, this type of linkage information is important, since rules prescribed in one section is only complete with the inclusion of references.

- Important terms used in a particular regulation are usually defined in a relatively early "definition" chapter of that regulation. For instance, in the domain of accessibility, the term "signage" is defined as "verbal, symbolic, tactile, and pictorial information (*UFAS* 1997)." Term definitions clearly add semantic information to domain-specific phrases and help understanding of regulations. Computationally, term definitions can be useful in linguistic analysis between different phrases that share similar definitions.

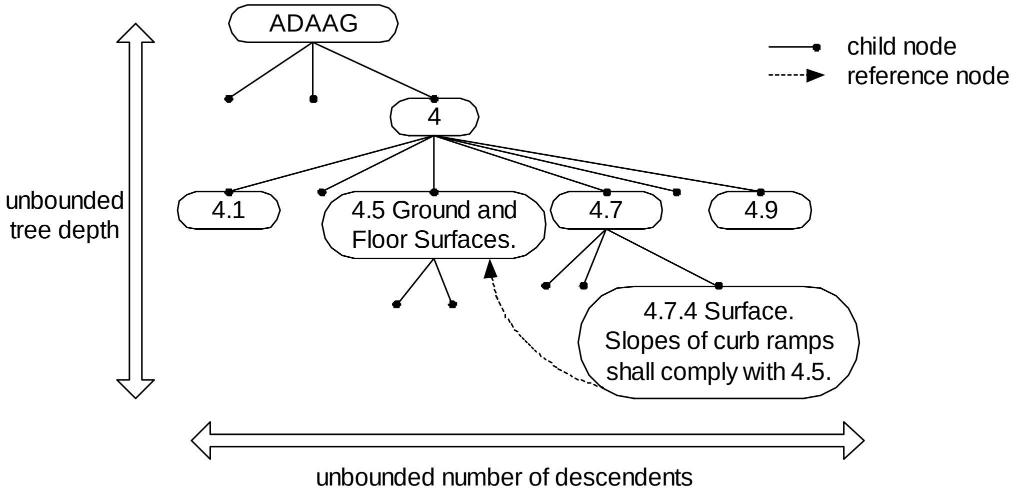

ADAAG
child node
reference node
4
unbounded tree depth
4.1
4.5 Ground and Floor Surfaces.
4.7
4.9
4.7.4 Surface. Slopes of curb ramps shall comply with 4.5.
unbounded number of descendents

Figure 2: Regulation Structure Illustrated with Selected Sections from the ADAAG

The first two properties are *structural* properties of regulations, while the third can be interpreted as a *feature* of regulations. We define *feature* to be the non-structural characteristics found in document contents that are specific to a corpus. In particular, since we are interested in comparing regulatory documents, features in our system can be defined as evidences that identify similarity or relatedness between provisions. Another example of feature is domain knowledge from industry experts as well as legal professionals and practitioners. This is because regulations are domain-centered; for instance, Title 40 from the US Code of Federal Regulations (*CFR* 2002) focuses exclusively on environmental protection. Domain experts from the field of environmental protection might identify other computational properties, such as a list of potential drinking water contaminants published by the US Environmental Protection Agency (*EPA* 2003), that they want to annotate for purpose of understanding as well as analysis in a regulatory infrastructure.

In this work, we will focus on accessibility regulations, whose intent is to provide the same or equivalent access to a building and its facilities for disabled persons. Two US Federal documents are incorporated in our corpus: the Americans with Disabilities Act Accessibility Guidelines (*ADAAG* 1999) and the Uniform Federal Accessibility Standards (*UFAS* 1997). We also include two European accessibility regulations: British

Standard BS 8300 (2001) and a selected part from the Scottish Technical Standards (2001). This paper discusses the development and results of a proposed relatedness analysis framework for regulations. The goal is to identify the most strongly related provisions across different regulation trees using not only a traditional term match but also a combination of feature matches, and not only content comparison but also structural analysis. We will exploit different computational properties of regulations, such as available domain knowledge, the tree organization and the referential structure of provisions, to perform a comprehensive comparison between regulations.

This paper is organized as follows. Section 2 reviews the literature of legal informatics, document comparisons and hyperlink topology. Section 3 defines the similarity score and the computation of relatedness between two provisions from two different regulation trees. The analysis starts with a base similarity score computation introduced in Section 3.1. The base score represents a linear combination of feature matching. We introduce a traditional Boolean matching model and a non-Boolean matching model to incorporate domain knowledge. Score refinements based on the structure of regulations are presented in Section 3.2. We address the natural hierarchical structure of regulations through a process termed neighbor inclusion. The referential structure of regulations is incorporated in the analysis through reference distribution. The final similarity score combines the base score with the score refinements so that similarities based on node content comparison as well as similarities from both neighbors and references are accounted for. Preliminary results and an application to e-rulemaking are presented in Section 4.

2 Related Work

This paper examines the use of a combination of feature and structural matching for a relatedness analysis for regulatory documents. There has been a great deal of work done in this area, and thus literature review is divided into three parts. Section 2.1 gives a brief description of related research in legal informatics. Section 2.2 examines different techniques for textual comparisons, such as the Vector Model and Latent Semantic Indexing. One of the computational properties of regulations is the hierarchical and referential organization of provisions; therefore, we will review citation analysis and different work based on hyperlink structure of the Web in Section 2.3.

2.1 Legal Informatics

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. For instance, CalDAG is one of many reference books written for compliance guidance with the accessibility code in California (Gibbens 2000). Unlike the long existence of interpretive guidelines, the introduction of Information Technology (IT) to aid legal interpretation is rather new.

Recently, governments are putting more information on the Internet, but information still remains difficult to locate and access (Baru et al. 2000). The emergence of e-government (dg.o 2001, dg.o 2002, dg.o 2003) has created a lot of research potential as a new application domain for IT, such as law enforcement (Lin et al. 2003) and e-rulemaking (Coglianese 2003). Some focus on regulation guidance using existing IT tools; for instance, the Business Gateway[1] project, a presidential e-government initiative, aims to reduce the burden of business by making it easy to find, understand, and comply with relevant laws and regulations (Small Business Administration 2002). Others focus on enhancing the search and browse aspect of legal corpus, whose targeted users are legal practitioners. Merkl and Schweighofer suggested that “the exploration of document archives may be supported by organizing the various documents into taxonomies or hierarchies that have been used by lawyers for centuries (Merkl and Schweighofer 1997).” Examples of long-existing legal resource vendors based on this paradigm include LexisNexis[2] and Westlaw[3].

Some researchers have applied Information Retrieval (IR) techniques to the domain of law. Schweighofer et al. attempted a content-based clustering and labeling of European

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

[2] LexisNexis online legal research system can be accessed at http://www.lexisnexis.com.

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

law, taking into account the importance of different terms (Schweighofer et al. 2001). 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 (Osborn and Sterling 1999). A similar approach is used in the SALOMON project that identified and extracted relevant information from case laws, such as keywords and summaries (Moens et al. 1997).

The use of available technologies from the field of Artificial Intelligence (AI) to aid the understanding of the law has been an active research topic for years (Rissland et al. 2003, Zeleznikow and Hunter 1994). The abstraction (Moens et al. 1997), representation (Bench-Capon 1991, Brüninghaus and Ashley 2001), classification (Schweighofer et al. 2001, Thompson 2001) and retrieval (Al-Kofahi et al. 2001) of case laws are widely studied. Earlier research focused on building expert system for law (Sergot et al. 1986, Zeleznikow and Hunter 1994). Case-based and rule-based systems are developed (Branting 1991, Daniels and Rissland 1997, Rissland and Skalak 1991). A logic-based reasoning tool has been prototyped to perform an automated compliance check on regulations (Kerrigan 2003, Kerrigan and Law 2003, Lau et al. 2003). Due to the complexity of the legal language, Natural Language Processing (NLP) techniques have been considered inappropriate to represent legal cases (Brüninghaus and Ashley 2001). For instance, one of the complexities of legal language is its open texture property, or in other words, the incomplete definition of many legal predicates. Some believe that rule makers favor phrases that are intentionally or unintentionally arguable in meaning (Berman and Hafner 1989), which results in the difficulty of modeling law using AI techniques (Rissland et al. 2003). Nevertheless, Brüninghaus and Ashley suggested that "recent progress in NLP has yielded tools that measure up to some of the complexities of legal texts (Brüninghaus and Ashley 2001);" for example, the open texture problem is well addressed in (Gardner 1984).

2.2 Document Comparisons

Text document comparison, in particular similarity analysis between a user query and documents in a generic corpus, is widely studied in the field of Information Retrieval. 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 locate the best match between user queries and documents, such as the Boolean model and the Vector model[4] (Salton 1971, Salton and McGill 1983). Most of these techniques are bag-of-word type of analysis, which means that they are word order insensitive (Baeza-Yates and Ribeiro-Neto 1999).

In the Vector model, each index term *i* is assigned a positive and non-binary weight $w_{i,M}$ in each document *M*. A document is represented as a *n*-entry vector $\vec{d}_M = (w_{1,M}, w_{2,M}, \ldots, w_{n,M})$, where *n* is the total number of index terms in the corpus. The Vector model proposes to evaluate the degree of similarity between two documents as the correlation between the two document vectors. By taking the correlation between two vectors as the degree of similarity, the Vector model assumes a Boolean matching between index terms, or in other words, term axes are mutually independent. For instance, the cosine of the angle between the two document vectors can be used as a correlation measure (Baeza-Yates and Ribeiro-Neto 1999):

$$f_v = \frac{\vec{d}_M \bullet \vec{d}_N}{|\vec{d}_M| \times |\vec{d}_N|} = \frac{\sum_{i=1}^{n} w_{i,M} \times w_{i,N}}{\sqrt{\sum_{i=1}^{n} w_{i,M}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,N}^2}}$$

where f_v is the similarity between documents *M* and *N* based on the Vector model. $|\vec{d}|$ denotes the norm of the document vector, which provides a normalization factor in the document space. Since cosine similarity is normalized, it always produces a score between 0 and 1.

There are a variety of algorithms to compute the index term weight *w*, and a general review can be found in (Salton and Buckley 1988). A simple approach is to use the count of term appearance as the term weight. One of the more popular algorithms is the *tf*×*idf*

[4] The Vector model is also called the Vector space model.

approach (Dumais 1991, Salton and Buckley 1988), which stands for the term frequency (*tf*) multiplied by the inverse document frequency (*idf*). Term frequency (*tf*) measures the term density in a document, whereas the inverse document frequency (*idf*) measures the term rarity across the corpus. Apparent from the name, *tf* represents the frequency count of term appearance in documents. The *idf* component is often computed as $log(k/k_i)$, where k is the total number of documents, and k_i is the number of documents in which the particular index term i appears. The *log* formula implements the intuition that a frequently-used term is not useful in distinguishing similarities between documents. Essentially, *tf* represents the intra-cluster similarity, while *idf* accounts for the inter-cluster dissimilarity.

Without the help of thesauri, this type of models cannot capture synonyms which can potentially convey important information. The Latent Semantic Indexing (LSI) model aims to fill the gap between terms and concepts (Deerwester et al. 1990). LSI uses an algorithm called Singular Value Decomposition (SVD) (Golub and Van Loan 1983) to reduce the dimension of term space into concept space as well as to perform noise reduction. The claim is that synonyms that represent the same concept are mapped onto the same concept axis through a dimension reduction. There are some investigations into improving the LSI, such as the Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999). In general, bag-of-word based approaches, such as the LSI or PLSA, are criticized for their lack of deep semantic understanding and their limitation to identifying only surface similarity (Crouch et al. 2002). As an alternative, work has been done in the area of linguistic analysis and ambiguity resolutions (Crouch et al. 2002, Everett et al. 2002) to detect redundant documents, on a very focused document set.

2.3 Hyperlink Topology

Due to the evolution of the World Wide Web, there has been a lot of research work related to academic citation analysis (Garfield 1995). For instance, CiteSeer is a scientific literature digital library that provides academic publications indexed with their citations (Bollacker et al. 1998). Different types of hyperlink topology and fitting models are examined extensively for different purposes (Calado et al. 2003, Gurrin and Smeaton

1999, Silva et al. 2000). One of the examples is Google's PageRank algorithm which ranks the importance of web pages by simulating the navigation pattern of Web users (Brin and Page 1998, Page et al. 1998). In this model, importance of web pages propagates through the hyperlink structure of the World Wide Web, with some random jumping behavior subsumed.

Aside from simulating Web surfers' behavior, the HITS (Hypertext Induced Topic Search) algorithm exploits the hyperlink structures to locate authorities and hubs on the Internet (Kleinberg 1998). Authorities are pages that have many citations pointing to them, whereas hubs represent pages that have a lot of outgoing links. It is a two-way feedback system where good hubs point to important authorities, and vice versa. Based on HITS, work has been done to infer Web communities and the breadth of topics in different disciplines from link analysis (Gibson et al. 1998). 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 slightly different from the Web – citation analysis assumes a pool of documents citing one another, while regulations are separate islands of information. Within an island of regulation, provisions are highly referenced; across islands, they are seldom cross-referenced.

3 Relatedness Analysis

The goal of a comparative analysis among regulations and supplementary documents is to identify materials that are alike in substance and/or connected by reason of a discoverable relation. Although the term *relatedness* appears more appropriate in this sense, the phrase "similarity score" has been used in the field of Information Retrieval (IR) traditionally. Therefore, we will use the terms *similarity* and *relatedness* interchangeably to represent the desired outcome of the above-defined comparative analysis in a legal domain. The phrase "similarity score" will be used to denote the comparison metric of *relatedness* between two provisions.

The proliferation of the Internet has led to an extensive amount of research on retrieving relevant documents based on a keyword search (Berry and Browne 1999). Well-established techniques such as query expansions (Ide 1971, Rocchio 1971) have been deployed to increase retrieval accuracy, with a significant amount of subsequent developments (Attar and Fraenkel 1977, Crouch and Yang 1992, Qiu and Frei 1993, Xu and Croft 1996) to improve performance. Thus, most repositories are equipped with a search and browse capability for viewing and retrieval of documents. It is reasonable to assume the following in a regulatory repository: at least one relevant document will be located by the user either with keyword search or by browsing through an ontology. Starting from a piece of correctly identified material, related documents are suggested to the user by our system, which is designed to incorporate special characteristics of regulations into comparisons between the identified material and the rest of the corpus. In essence, we focus on refining the back end comparison technique for documents rather than matching queries at the front end.

Besides the goal and assumptions of the analysis, we shall define the unit of comparison as well. Since a typical regulation can easily exceed thousands of pages, a comparison between a full set of regulation and another is meaningless (Branting 1991). 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 (1999) and Section 1127B.4.4 in CBC (1998) as in the example shown in Figure 1. There is one terminological clarification – we use the terms "section" and "provision" interchangeably to represent the unit of comparison. The actual and official terminology differs from regulation to regulation. For example, Section 4 (in our terminology) could be termed Part 4, Section 4.3 could be referred to as Subpart 4.3 and Section 4.3(a) could be called Provision 4.3(a). We will use the terms "section" and "provision" to represent all of the above indistinguishably.

To help define the terminologies for the basis in our comparison, we show below an illustration of two partial regulation trees: the Americans with Disabilities Act Accessibility Guidelines (*ADAAG* 1999) and the Uniform Federal Accessibility Standards (*UFAS* 1997). As shown in Figure 3, we take Section *A* from the ADAAG and Section *U* from the UFAS as our interested point of comparison. The immediate neighbors of a node, i.e., the parent, siblings and children of a provision, are collectively termed the *psc* of that particular provision. In other words, the *psc* operation on a node returns the set of nodes defined as the immediate neighbors. The references from a provision are collectively termed the *ref* of that particular provision, as shown as set *ref*(*U*) for Section *U* in Figure 3. Here, two different regulation trees are shown as an example, which is the intent of our analysis. A self-comparison, which is defined as a comparison among provisions in the same regulation tree, can also be performed using the same analysis.

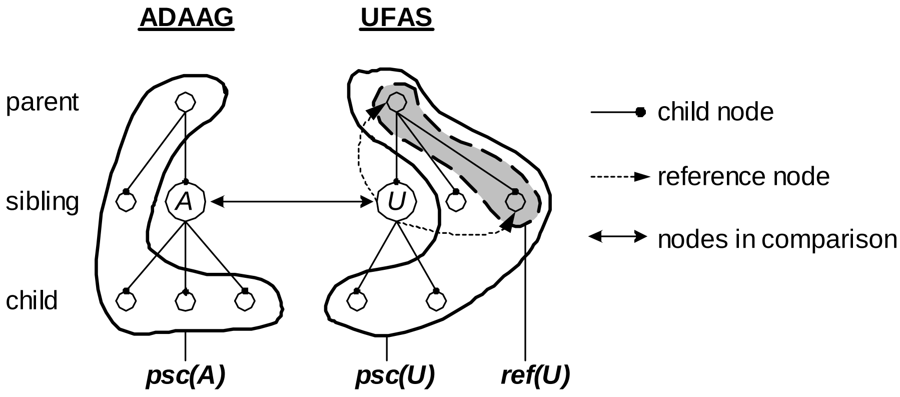


Figure 3: Immediate Neighboring Nodes and Referenced Nodes in Regulation Trees

After defining the goal, the unit and the operators of our analysis, we will introduce the measure we use for comparison – a similarity score. A similarity score measures the *degree of similarity* between two documents, and is defined on a relatedness measurement interval that ranges from 0 to 1, with 0 representing unrelated materials and 1 being the most related or identical materials. The similarity score is denoted by *f*(*A*, *U*) ∈ [0, 1] per pairs of provisions, for example, pair (*A*, *U*) with Section *A* from the ADAAG and Section *U* from the UFAS. The comparison should be commutative as well, that is, a comparison between Sections *A* and *U* should produce the same result as a comparison between Sections *U* and *A*. In other words, we have *f*(*A*, *U*) = *f*(*U*, *A*).

A schematic is shown below in Figure 4 for the similarity analysis core. The input to the system is a set of refined XML regulations tagged with the associated features as well as any user-provided domain knowledge. The system produces as a result a list of the most related pairs of provisions across different regulations. The dissimilar pairs are discarded while the most related pairs are returned to interested users. Starting from a wellprepared repository such as one described in (Lau et al. 2003), we employ a combination of IR techniques and document structure analysis to extract related provisions based on a similarity measure. The goal of the similarity analysis core is to produce a similarity score *f* per pairs of provisions as defined above. As shown in Figure 4 and will be discussed in Section 3.1, a base score is first computed based on different feature matching, which incorporates domain knowledge if available. Section 3.2 will introduce the subsequent refinements of the base score to account for the structure of regulations. The resulting final score represents a combination of feature and structural matching between provisions in different regulation trees.

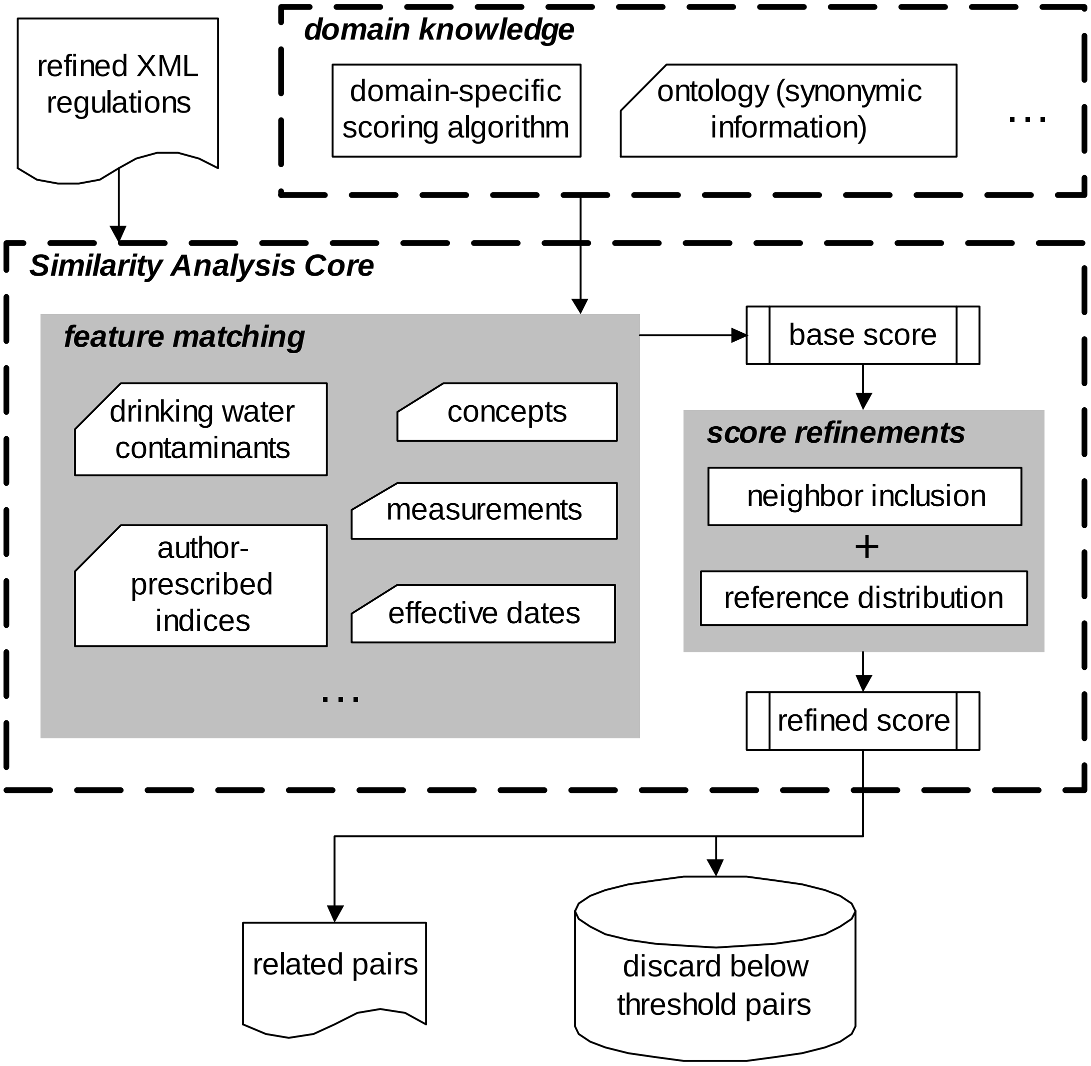


Figure 4: Similarity Analysis Core Schematic

3.1 Base Score Computation

The base score represents the direct content comparison of provisions based on different feature matching. As defined earlier, feature represents the evidence of relatedness between two provisions, which could be domain-specific information. Feature matching is the comparisons of non-structural characteristics from regulations. As shown in Figure 4, there are generic features that are common across all domains of regulations, such as exceptions, definitions and concepts. The second type of features are domain-specific ones, such as glossary terms defined in engineering handbooks, author-prescribed indices at the back of reference books, measurements found in both accessibility and environmental regulations, and chemicals and effective dates specific to environmental regulations. The example of two conflicting provisions, shown in Figure 1 (Gibbens 2000), best illustrates the reason for including both types of features. The conflict is capture by the clash between the term *flush* and the measurement *½ inch lip at 45 degrees*. The example demonstrates the need to extract conceptual information, e.g., key phrases in the corpus, as well as domain-specific information, such as measurements in this case, for a complete regulatory analysis.

The base score is a linear combination of the scores obtained using different feature matching, which allows for a combination of generic features, such as concepts, as well as 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 (Salton 1971), 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. Scoring schemes for other features are developed based on a similar idea.

Here, our usage of the Vector model differs from generic applications in two ways. Our comparison is on extracted features, such as measurements, but not index terms; in addition, we have a much more selective collection of documents, namely regulations in certain domains rather than a general-purpose corpus. If one desires to incorporate domain knowledge, axis independence no longer holds. For instance, some features are characterized by ontologies to define synonyms. Some features simply cannot be modeled as Boolean term matches due to their inherent non-Boolean property, such as measurements, (As an example, 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 supplemented with feature dependency information defined by knowledge experts, who do not necessarily agree with a Boolean definition. It is unrealistic to assume that the world can be modeled as a Boolean match, and as a result, domain knowledge is potentially non-Boolean. In essence, the degree of match between two features is no longer limited to only 0% or 100%.

To accommodate a non-Boolean degree-of-match algorithm, we propose a vector space transformation based on the Vector model. For features with defined synonyms or a non-Boolean matching scheme, the feature vectors are mapped onto a different vector space before a cosine comparison. A linear transformation in the form of $\vec{m}' = D\vec{m}$, where D denotes the transformation matrix, is employed to account for axis dependencies introduced by user-defined partial match algorithms. In other words, D captures available domain knowledge, and projects the feature vector $\vec{m}$ onto an alternate space where the resultant vector $\vec{m}' = D\vec{m}$ represents the consolidated feature frequencies. Details and proofs of the formulation are given in (Lau 2004). The transformation is shown to produce consistent results when synonymic information are modeled using two different spaces, namely the original n-dimensional space and a reduced vector space with the synonymic feature axes collapsed into one.

3.2 Score Refinements

The base score is subsequently refined by utilizing the tree structure of regulations. As shown in Figure 4, there are two types of score refinement: neighbor inclusion and reference distribution. In neighbor inclusion, 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 3, the immediate neighbors of provision A, i.e., the parent, siblings and children, are collectively termed the *psc*(A) of node A. The set of nodes in *psc*(*A*) is related to node *A* through a parent, sibling or child relationship. As defined earlier, similarity analysis aims to reveal entities that are "connected by reason of an established or discoverable relation"; therefore, we utilize the *psc* relationships between nodes to refine the comparison in an attempt to discover more similarity relationships.

Neighbor inclusion assumes *diffusion* of similarity between clusters of nodes in the tree; Figure 5 best illustrates the idea. The similarity between *psc*(A_1) and *psc*(U_1), represented by clusters shaded in dark gray, diffuses to nodes A_1 and U_1. Likewise, the dissimilarity between *psc*(A_2) and *psc*(U_2), shown using lightly-shaded clusters, spreads to nodes A_2 and U_2. In other words, neighbor inclusion implies that there exist clusters of related nodes when comparing two trees. A tree-structured regulation should theoretically support this assumption, since the purpose of such structured regulation is to organize relevant materials into coherent provisions and sub-provisions. A matrix representation is developed, where a neighbor structure matrix is defined to codify the neighbor relationship in a regulation tree (Lau 2004). The similarity score between *psc*(*A*) and *psc*(*U*) contributes to the final similarity score between nodes *A* and *U*, which implements the intuition that similarities between the immediate neighbors imply similarity between the interested pair.

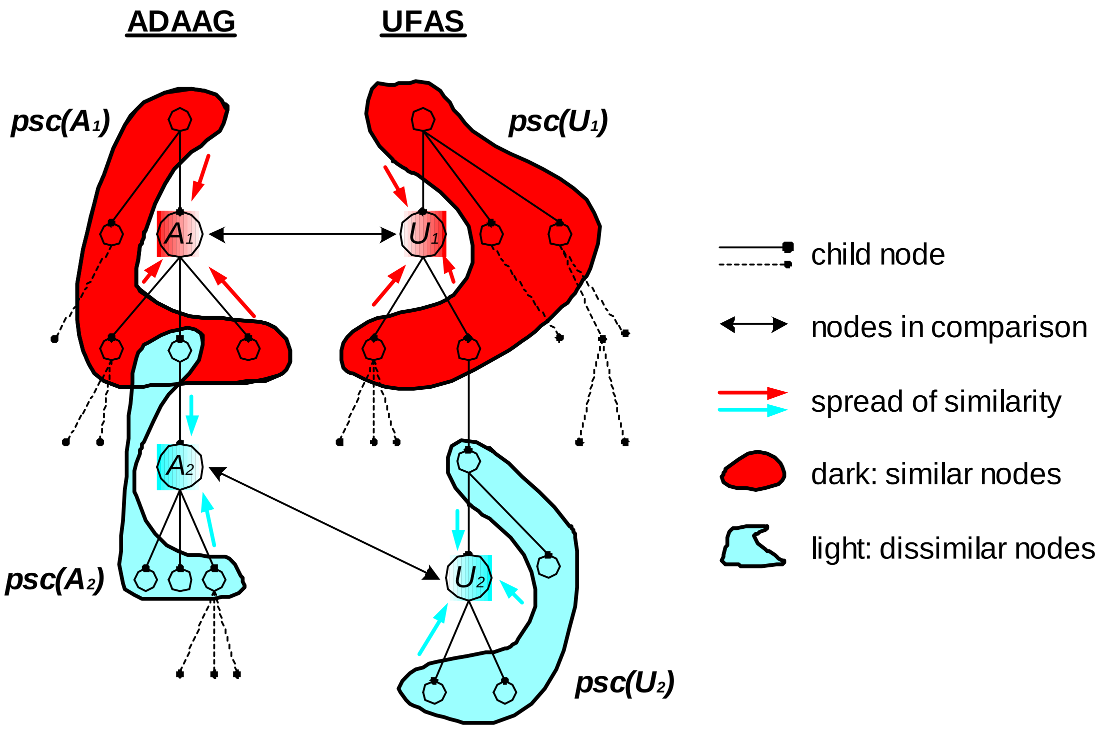


Figure 5: Diffusion of Similarity among Clusters of Nodes Introduced by Neighbor Inclusion

The referential structure of regulations is handled in a similar manner, based on the assumption that similar sections often reference similar sections. 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. Analogous to neighbor inclusion, a reference structure matrix is introduced to represent the citations among nodes in a regulation tree, which results in a concise matrix notation of the computation (Lau 2004). Referring to Figure 3, the out-references from provision *A* are termed the *ref*(*A*) of node *A*. Taking Section *A* from the ADAAG (1999) and Section *U* from the UFAS (1997) as an example, *ref*(*A*) is compared to *ref*(*U*) just as *psc*(*A*) versus *psc*(*U*) in neighbor inclusion. After successive refinements, similarities from both near-tree neighbors and references are identified, and related provisions can be retrieved based on the final similarity scores.

3.3 Final 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 these two interested Sections *A* and *U*. Any available domain knowledge is also captured in the base score. Neighbor inclusion infers similarity between Sections *A* and *U* based on their shared clusters of neighbors in their 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 comparison and structural comparison.

As a result of a relatedness analysis, related provisions can be retrieved and recommended to users based on the resulting scores. Different combinations of features, different weighting schemes, as well as different combinations of techniques such as a base score computation and neighbor inclusion only, can be experimented. The next section gives an overview of results obtained using different combinations of parameters. Potential application of the developed analysis will be demonstrated through an e-rulemaking scenario.

4 Results and Applications

In this section, we will show some performance evaluation, examples of comparison results and an application on e-rulemaking. Section 4.1 compares our system to LSI through the use of a user survey. Section 4.2 gives examples of results based on comparisons among different sources of regulations. To illustrate possible extended application of the system, we will discuss an e-rulemaking scenario in Section 4.3.

4.1 Comparisons to a Traditional Retrieval Model

Precision and recall are the general metric to evaluate the performance of a textual similarity analysis system. The problem with a precision and recall measure is that such benchmark does not always exist. A legal corpus is particularly difficult. For example, contextual information cannot be neglected since provisions are not usually self-contained. Terminological differences need to be anatomized with technical and domain-specific terms defined. Therefore, it is rather an impossible task for any individual to thoroughly understand each provision and determine whether it is related to other provisions.

To assess the performance of our system, we devise a user survey as an evaluation metric. Users are asked to rank the similarity of ten randomly chosen provisions from the ADAAG (1999) and ten from the UFAS (1997). The ranking is chosen as the metric since similarity scores are a relative measure. Ten surveys are collected, and the average ranking is taken to be the "correct" answer. As a benchmark, we compare the accuracy of our system with that of a traditional retrieval model. Traditional techniques, such as the Vector model (Salton 1971), simply compare the frequency counts of index terms between documents. A popular alternative is Latent Semantic Indexing (LSI) (Deerwester et al. 1990), which uses Singular Value Decomposition (SVD) to reduce the dimension of term space into concept space by keeping only a portion of the largest singular values. We compare our system with Latent Semantic Indexing, since LSI claims to form concept axes instead of term axes, which shares a similar goal as for our feature comparison.

To compute the error of machine rankings with respect to human rankings, we compare the ADAAG with the UFAS and sections are ranked according to the scores produced by our system as well as by LSI. The Root Mean Square Error (RMSE) is used to compute the ranking prediction error, i.e., the difference between the "correct" ranking and the machine predicted ones, based on the survey results as the "correct" answer. The RMSE is the root of the Residual Sum of Squares (RSS) normalized according to the number of observations, which is 100 in our case. We compute the RMSE for a LSI implementation using the 300 largest singular values. Overall, our system outperforms the LSI - RMSEs of our system and LSI are 22.9 and 27.4, respectively. Individual combinations of features and structural matching produce 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 the importance of 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 $ref(\cdot)$ 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.

4.2 Comparisons among Different Sources of Regulations

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, and some improvement is observed. For instance, Section 4.1.6(3)(d) in the ADAAG (1999) is concerned with doors, while Section 4.14.1 in the UFAS (1997) 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 the neighbors in the tree. The related accessible elements, namely door and entrance[5], are identified indirectly through neighbor inclusions.

ADA Accessibility Guidelines
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 beveled edge on each side, they may
  remain.
```

Uniform Federal Accessibility Standards
4.14.1 Minimum Number

```
  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.
```

Figure 6: Related Provisions Identified Through Neighbor Inclusion

To illustrate the similarity between American and British standards, we compare the UFAS (1997) with the BS8300 (2001). Figure 7 and Figure 8 show a sub-tree 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. In addition, even a dictionary would not be able to identify the esoteric phrases “door hardware” and “door furniture” as relevant. However, similarities in neighboring nodes, in particular the parent and siblings, implied a higher similarity between Section 4.13.9 of UFAS and

[5] Definitions of “*door*” in WordNet (Miller et al. 1993) 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.”

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.

Uniform Federal Accessibility Standards
4.13.9 Door Hardware

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

British Standard 8300
12.5.4.2 Door Furniture

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.

Figure 7: Terminological Differences Between the UFAS and the BS8300

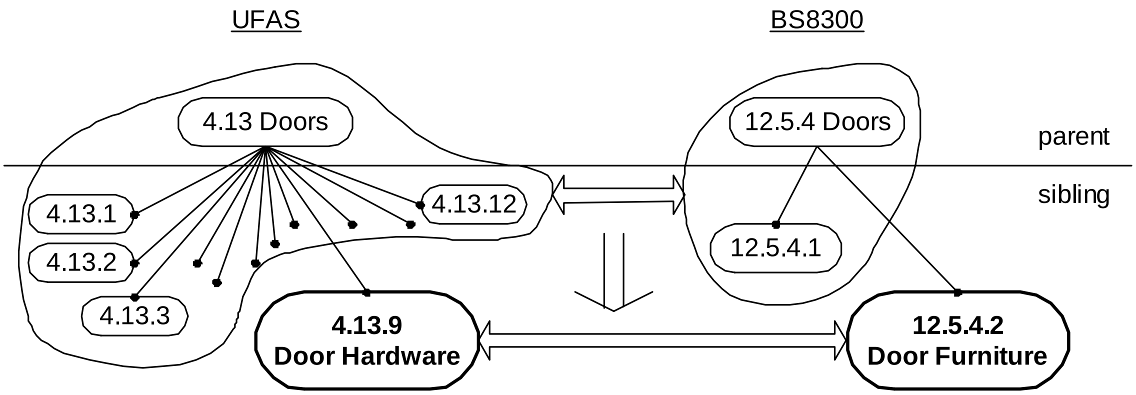


Figure 8: Similarities Between Neighbors Imply Similarities Between Section 4.13.9 from the UFAS and Section 12.5.4.2 from the BS8300

The UFAS (1997) is compared with the Scottish code (2001) in addition to the BS8300 (2001). An observation based on the comparisons between the UFAS and the Scottish code is given below in Figure 9, where reference distribution contributes to revealing hidden similarities between provisions. As shown in Figure 9, both sections from the UFAS and the Scottish code 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. 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.

Uniform Federal Accessibility Standards
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 9: Related Elements “Stairs” and “Ramp” Revealed Through Reference Distribution

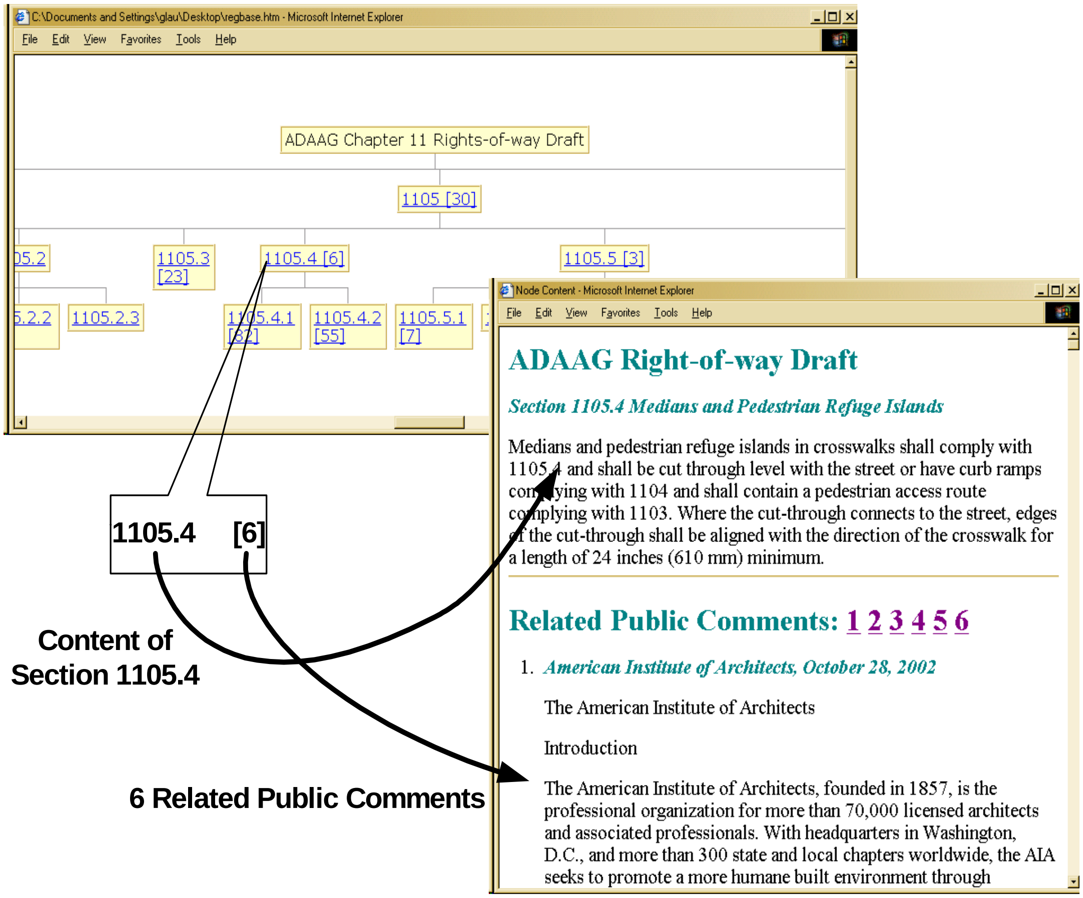
4.3 Application on E-rulemaking

Apart from the intended application on comparisons of regulatory documents and to demonstrate system scalability and extensibility, we have applied the prototype system to the domain of electronic-rulemaking (e-rulemaking). The making of government regulations represents an important communication between the government and citizens. During the process of rulemaking, government agencies are required to inform and to invite the public to review a proposed rule. Interested and affected citizens then submit comments accordingly. E-rulemaking redefines this process of rule drafting and commenting to effectively involve the public in the making of regulations. The electronic media, such as the Internet, is used as the means to provide a better environment for the public to comment on proposed rules and regulations. For instance, email has become one popular communication channel for comment submission. Based on the review of the received public comments, government agencies revise the proposed rules.

The process of e-rulemaking easily generates a large amount of electronic data, i.e., the public comments, that needs to be reviewed and analyzed along with the drafted rules. With the proliferation of the Internet, it becomes a growing problem for government agencies to handle a growing amount of data from the public. In order to help screening and filtering of public comments, we applied our system to this domain by comparing the drafted rules with the associated comments. Our source of data is from the US Access Board, who released a newly drafted chapter (2002) for the ADAAG (1999), 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 totaled around 10 Megabytes in size, with some comments longer than the draft itself. To facilitate understanding of the comments with reference to the draft, a relatedness analysis is performed on the drafted chapter and the comments.

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. Domain-specific features, such as measurements, do not add much value here since comments coming from the general public tend to be less technical. However, commenters often follow similar terminologies found in the regulation, and therefore generic features, such as concepts, are still representative of comments. As for structural matching, we are essentially performing a single-tree (only the regulation tree but not the comments) structural analysis, since comments are not hierarchically organized. Nevertheless, neighbors and references in the draft regulation should not be overlooked.

The results of a relatedness analysis are related pairs between the provision from the draft and individual comment. Figure 10 shows the developed framework where users are given an overview of the draft along with related comments. Industry designers, planners, policy makers as well as interested and affected individuals are potential users who can benefit from the exploration of relevant provisions and comments provided by this framework. As shown Figure 10, 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 demonstrates how a regulatory comparison system can also be useful in an e-rulemaking situation where one needs to review drafted rules based on a large pool of public comments.


C:\Documents and Settings\glau\Desktop\regbase.htm - Microsoft Internet Explorer
File Edit View Favorites Tools Help
ADAAG Chapter 11 Rights-of-way Draft
1105 [30]
05.2
1105.3 [23]
1105.4 [6]
1105.5 [3]
5.2.2
1105.2.3
1105.4.1 [82]
1105.4.2 [55]
1105.5.1 [7]
Node Content - Microsoft Internet Explorer
File Edit View Favorites Tools Help
1105.4 [6]
Content of Section 1105.4
6 Related Public Comments

ADAAG Right-of-way Draft

Section 1105.4 Medians and Pedestrian Refuge Islands

Medians and pedestrian refuge islands in crosswalks shall comply with 1105.4 and shall be cut through level with the street or have curb ramps complying with 1104 and shall contain a pedestrian access route complying with 1103. Where the cut-through connects to the street, edges of the cut-through shall be aligned with the direction of the crosswalk for a length of 24 inches (610 mm) minimum.

Related Public Comments: 1 2 3 4 5 6

1. *American Institute of Architects, October 28, 2002*

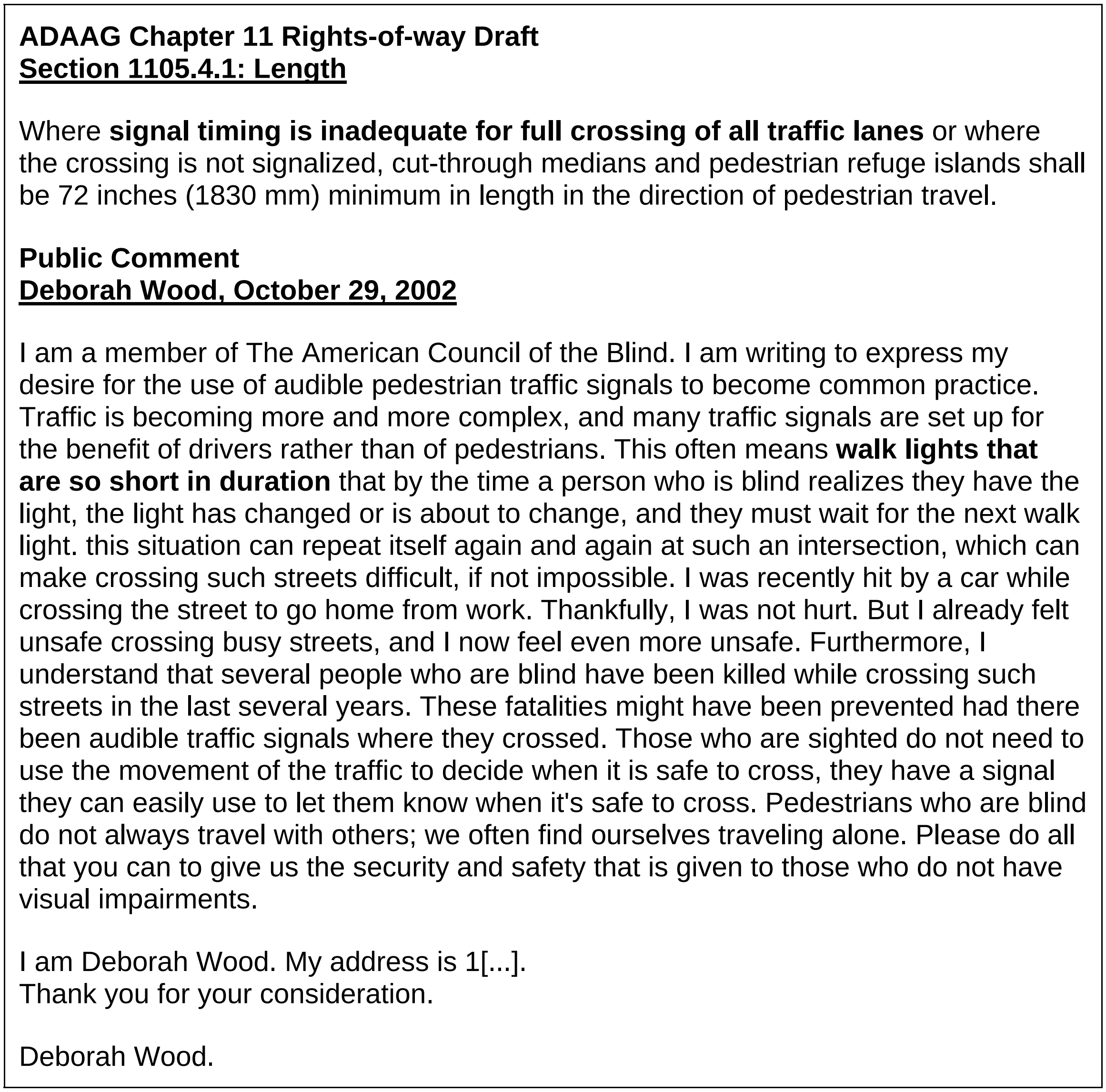
   The American Institute of Architects

   Introduction

   The American Institute of Architects, founded in 1857, is the professional organization for more than 70,000 licensed architects and associated professionals. With headquarters in Washington, D.C., and more than 300 state and local chapters worldwide, the AIA seeks to promote a more humane built environment through

Figure 10: Comparisons of Drafted Rules with Public Comments in E-Rulemaking

Two interesting results are presented here to illustrate the potential impacts as well as limitations of the use of a comparison framework on rulemaking. Figure 11 shows a typical pair of drafted section and its identified related public comment. Section 1105.4.1 in the draft discusses about situations when "signal timing is inadequate for full crossing of traffic lanes." Indeed, one of the reviewers complained about the same situation, where in the reviewer's own words, "walk lights that are so short in duration" should be investigated. This example illustrates that our system correctly retrieves related pairs of drafted section and public comment, which is useful to aid user understanding of the draft. 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 clearly needed as it would otherwise require a lot of human effort to perform a full content comparison to sort through piles of comments.



ADAAG Chapter 11 Rights-of-way Draft
Section 1105.4.1: Length

Where **signal timing is inadequate for full crossing of all traffic lanes** or where the crossing is not signalized, cut-through medians and pedestrian refuge islands shall be 72 inches (1830 mm) minimum in length in the direction of pedestrian travel.

Public Comment
Deborah Wood, October 29, 2002

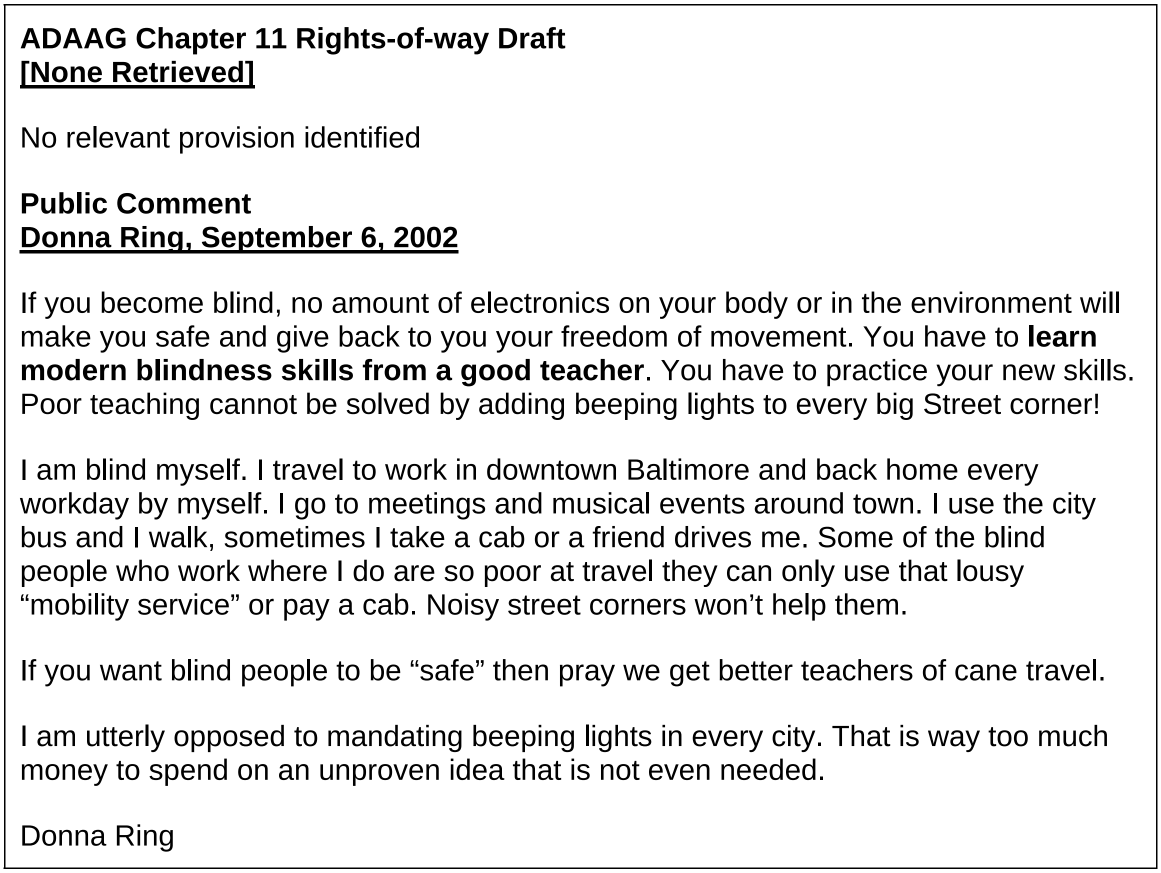
I am a member of The American Council of the Blind. I am writing to express my desire for the use of audible pedestrian traffic signals to become common practice. Traffic is becoming more and more complex, and many traffic signals are set up for the benefit of drivers rather than of pedestrians. This often means **walk lights that are so short in duration** that by the time a person who is blind realizes they have the light, the light has changed or is about to change, and they must wait for the next walk light. this situation can repeat itself again and again at such an intersection, which can make crossing such streets difficult, if not impossible. I was recently hit by a car while crossing the street to go home from work. Thankfully, I was not hurt. But I already felt unsafe crossing busy streets, and I now feel even more unsafe. Furthermore, I understand that several people who are blind have been killed while crossing such streets in the last several years. These fatalities might have been prevented had there been audible traffic signals where they crossed. Those who are sighted do not need to use the movement of the traffic to decide when it is safe to cross, they have a signal they can easily use to let them know when it's safe to cross. Pedestrians who are blind do not always travel with others; we often find ourselves traveling alone. Please do all that you can to give us the security and safety that is given to those who do not have visual impairments.

I am Deborah Wood. My address is 1[...].
Thank you for your consideration.

Deborah Wood.

Figure 11: Related Drafted Rule and Public Comment

A different type of comment screening is shown in Figure 12. 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 installment of electronic devices on streets 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 the relatedness analysis system. As shown in the two examples, by segmenting the pool of comments according to their relevance to individual provisions, our system can potentially save rule makers significant amount of time in reviewing public comments in regard to different provisions in the drafted regulations.



ADAAG Chapter 11 Rights-of-way Draft
[None Retrieved]

No relevant provision 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**. You have to practice your new skills. Poor teaching cannot be solved by adding beeping lights to every big Street corner!

I am blind myself. I travel to work in downtown Baltimore and back home every workday by myself. I go to meetings and musical events around town. I use the city bus and I walk, sometimes I take a cab or a friend drives me. Some of the blind people who work where I do are so poor at travel they can only use that lousy "mobility service" or pay a cab. Noisy street corners won't help them.

If you want blind people to be "safe" then pray we get better teachers of cane travel.

I am utterly opposed to mandating beeping lights in every city. That is way too much money to spend on an unproven idea that is not even needed.

Donna Ring

Figure 12: A Piece of Public Comment Not Related to the Draft

5 Summary

The advance in Information Technology has provided us with tools to streamline the development of regulatory policy and to facilitate understanding of regulations. One important aspect is to integrate rules with other laws, such as using IT to "link all the traces of a rule's history, both back to the underlying statues and back to past or related rules, facilitating improved understanding of legal requirements (Coglianese 2004)." In this paper, we introduce a relatedness analysis system that links relevant provisions to one another.

Based on a well-parsed repository for regulations (Lau et al. 2003), the development of a comparative analysis between regulatory provisions is presented. The goal is to identify relatedness or similarity among different sources of regulations. The computational properties of regulations are identified and used in the proposed analysis. Specifically, the hierarchical and referential structures of regulations as well as available domain knowledge are incorporated into the comparison model.

We start with the computation of a base score, which represents the degree of similarity between two provisions based on a pure content comparison. The base score is a linear combination of scores from each feature matching. 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. Both Boolean and non-Boolean feature matching algorithms are discussed, where domain knowledge is accounted for in the model.

The base score is subsequently refined by utilizing the tree 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. 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. 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 comparison and structural comparison.

Performance evaluation is conducted through a user survey, where results obtained using our system are compared with results from a traditional retrieval model. Different groups of regulations are compared and examples are given to illustrate the use of different features and structures of regulations. To demonstrate system capability, we applied the developed tool to the e-rulemaking domain where drafted rules are compared with their associated public comments. Results and applications showed that our system successfully identify pairs of related elements in a regulatory domain.

The development of a relatedness analysis framework is only the beginning of many potential applications of IT to aid the making of law. For instance, regulations are frequently updated by agencies to reflect environmental changes and new policies. However, the desynchronized updating of regulations seems to be problematic, especially when different regulations reference one another. We observe that there is a need for consistency check among multiple sources of regulations citing each other as references. For instance, in the domain of accessibility, Balmer pointed out that the "ADAAG references the A17.1 elevator code for conformance. Since 2000 there has been no section of the A17 that references lifts for the disabled. Therefore ADAAG references a non-existent standard (Balmer 2003)." Extending on the developed reference extraction tool, cross citations can be automatically located and checked for consistency. Such kind of tool is valuable for rule makers to validate regulations during the drafting process.

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 (*Shepard's Federal Citations* 1990). If we include cases in our corpus, citations from cases to provisions can potentially help to identify related provisions. Case citations can then be incorporated into the computation analogous to reference distribution.

In the e-rulemaking application, our system currently is limited to compare drafted provisions with public comments. We observed that this approach would miss comments that are not *directly* related to any particular provision in the draft. Sometimes, 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 help classify this type of opinions. We also observed that commenters frequently suggest rewording and revisions of the drafted provisions directly in their reviews. To precisely locate revisions embedded in the comments, one can perform linguistic analysis to compute differences between the drafted version and the suggested version. This is assuming that the suggested revision does not differ significantly from the draft, so that patterns can still be matched.

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.

6 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.

7 Bibliography

Al-Kofahi K, Tyrrell A, Vachher A and Jackson P (2001) 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, pp. 88-93.

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

Attar R and Fraenkel AS (1977) Local Feedback in Full-Text Retrieval Systems. *Journal of the ACM, 24 (3)*: 397-417.

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

Balmer DC (2003) Trends and Issues in Platform Lift. In: *Proceedings of Space Requirements for Wheeled Mobility Workshop*, Buffalo, NY.

Baru C, Gupta A, Papakonstantinou Y, Hollebeek R and Featherman D (2000) "Putting Government Information at Citizens' Fingertips," *EnVision, 16 (3),* pp. 8-9.

Bench-Capon TJM (1991) *Knowledge Based Systems and Legal Applications*. Academic Press Professional, Inc., San Diego, CA.

Bender D (2004) 2003 Data Protection Survey: Cross-Border Transfer of Personal Data in 22 Major Jurisdictions. In: *Proceedings of the 3rd Annual Law Firm C.I.O. Forum 2004*, San Francisco, CA, pp. 95-122.

Berman DH and Hafner CD (1989) The Potential of Artificial Intelligence to Help Solve the Crisis in Our Legal System. *Communications of the ACM, 32 (8)*: 928-938.

Berry MW and Browne M (1999) *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Bollacker KD, Lawrence S and Giles CL (1998) 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, pp. 116-123.

Branting LK (1991) Reasoning with Portions of Precedents. In: *Proceedings of the 3rd International Conference on Artificial Intelligence and Law (ICAIL 1991)*, Oxford, England, pp. 145-154.

Branting LK (1991) Building Explanations from Rules and Structured Cases. *International Journal of Man-Machine Studies, 34 (6)*: 797-837.

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

British Standard 8300 (2001) British Standards Institution (BSI), London, UK.

Brüninghaus S and Ashley KD (2001) 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, pp. 42-51.

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

California Building Code (CBC) (1998) California Building Standards Commission, Sacramento, CA.

Code of Federal Regulations (CFR) (2002) Title 40, Parts 141 - 143, US Environmental Protection Agency, Washington, DC.

Coglianese C (2003) *E-Rulemaking: Information Technology and Regulatory Policy,* Technical Report, Regulatory Policy Program, Kennedy School of Government, Harvard University, Cambridge, MA.

Coglianese C (2004) Information Technology and Regulatory Policy. *Social Science Computer Review, 22 (1)*: 85-91.

Crouch CJ and Yang B (1992) Experiments in Automatic Statistical Thesaurus Construction. In: *Proceedings of the 15th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, pp. 77-88.

Crouch R, Condoravdi C, Stolle R, King T, de Paiva V, Everett J and Bobrow D (2002) Scalability of Redundancy Detection in Focused Document Collections. In: *Proceedings of the 1st International Workshop on Scalable Natural Language Understanding (ScaNaLU-2002)*, Heidelberg, Germany.

Daniels JJ and Rissland EL (1997) 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, pp. 325-336.

Deerwester S, Dumais ST, Furnas GW, Landauer TK and Harshman R (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science, 41 (6)*: 391-407.

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

Dumais ST (1991) Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers, 23 (2)*: 229-236.

Everett JO, Bobrow DG, Stolle R, Crouch R, de Paiva V, Condoravdi C, Berg Mvd and Polanyi L (2002) Making Ontologies Work for Resolving Redundancies Across Documents. *Communications of the ACM, 45 (2)*: 55 - 60.

Gardner A (1984) *An Artificial Intelligence Approach to Legal Reasoning*, Ph.D. Thesis, Computer Science, Stanford University, Stanford, CA.

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

Gibbens MP (2000) *CalDAG 2000: California Disabled Accessibility Guidebook.* Builder's Book, Canoga Park, CA.

Gibson D, Kleinberg J and Raghavan P (1998) Inferring Web Communities from Link Topology. In: *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, PA, pp. 225-234.

Golub GH and Van Loan CF (1983) *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD.

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

Hofmann T (1999) Probabilistic Latent Semantic Indexing. In: *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, pp. 50-57.

Ide E (1971) "New Experiments in Relevance Feedback," In G. Salton (Eds.), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Inc., Englewood Cliffs, NJ.

Kerrigan S (2003) *A Software Infrastructure for Regulatory Information Management and Compliance Assistance*, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA.

Kerrigan S and Law K (2003) Logic-Based Regulation Compliance-Assistance. In: *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003)*, Edinburgh, Scotland, pp. 126-135.

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

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

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

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

Lau G, Law K and Wiederhold G (2003) A Framework for Regulation Comparison with Application to Accessibility Codes. In: *Proceedings of the National Conference on Digital Government Research*, Boston, MA, pp. 251-254.

Lin C, Hu PJ, Chen H and Schroeder J (2003) Technology Implementation Management in Law Enforcement: COPLINK System Usability and User Acceptance Evaluations. In: *Proceedings of the National Conference on Digital Government Research*, Boston, MA, pp. 151-154.

Merkl D and Schweighofer E (1997) En Route to Data Mining in Legal Text Corpora: Clustering, Neural Computation, and International Treaties. In: *Proceedings of the 8th International Workshop on Database and Expert Systems Applications*, Toulouse, France, pp. 465-470.

Miller GA, Beckwith R, Fellbaun C, Gross D and Miller K (1993) *Five Papers on WordNet*, Technical Report, Cognitive Science Laboratory, Princeton, NJ.

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

Osborn J and Sterling L (1999) 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, pp. 173-181.

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

Potential Drinking Water Contaminant Index (2003) US Environmental Protection Agency, Washington, DC.

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

Proceedings of the National Conference on Digital Government Research (dg.o 2001) Los Angeles, CA.

Proceedings of the National Conference on Digital Government Research (dg.o 2002) Los Angeles, CA.

Proceedings of the National Conference on Digital Government Research (dg.o 2003) Boston, MA.

Qiu Y and Frei H-P (1993) Concept Based Query Expansion. In: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, pp. 160-169.

Raskopf RL and Bender D (2003) Cross-Border Data: Information Transfer Restrictions Pose a Global Challenge. *New York Law Journal.*

Rissland EL, Ashley KD and Loui RP (2003) AI and Law: A Fruitful Synergy. *Artificial Intelligence, 150 (1-2)*: 1-15.

Rissland EL and Skalak DB (1991) CABARET: Rule Interpretation in a Hybrid Architecture. *International Journal of Man-Machine Studies, 34 (6)*: 839-887.

Rocchio JJ (1971) "Relevance Feedback in Information Retrieval," In G. Salton (Eds.), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Inc., Englewood Cliffs, NJ.

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

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

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

Schweighofer E, Rauber A and Dittenbach M (2001) 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, pp. 78-87.

Sergot MJ, Sadri F, Kowalski RA, Kriwaczek F, Hammond P and Cory HT (1986) The British Nationality Act as a Logic Program. *Communications of the ACM, 29 (5)*: 370-386.

Shepard's Federal Citations (1990). Shepards/Mcgraw-Hill, Colorado Springs, CO.

Silva I, Ribeiro-Neto B, Calado P, Moura E and Ziviani N (2000) 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, pp. 96-103.

Technical Standards (2001) Scottish Executive, Edinburgh, Scotland, UK.

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

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

Xu J and Croft WB (1996) Query Expansion Using Local and Global Document Analysis. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 4-11.

Zeleznikow J and Hunter D (1994) *Building Intelligent Legal Information Systems: Representation and Reasoning in Law*. Kluwer Law and Taxation Publishers, Deventer, The Netherlands.