

A COMPARATIVE ANALYSIS FRAMEWORK FOR SEMI-
STRUCTURED DOCUMENTS, WITH APPLICATIONS TO
GOVERNMENT REGULATIONS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
CIVIL AND ENVIRONMENTAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Gloria T. Lau

August 2004

© Copyright by Gloria T. Lau 2004

All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Kincho H. Law
(Principal Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Gio Wiederhold
(Co-Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Hans C. Bjornsson

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Cary Coglianese

Approved for the University Committee on Graduate Studies.

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. In this work, an information infrastructure is proposed for regulation management and analysis, which includes a consolidated document repository and tools for similarity analysis. The corpus covers accessibility and environmental regulations from the US Federal government, California state government, non-profit organizations and some European agencies.

The regulatory repository is to be populated with regulations in XML format. XML is chosen as the representation format because it is well suited for handling semi-structured data such as legal documents. A shallow parser is developed to consolidate regulations published in different formats, for example, PDF or HTML, into XML. The shallow parser also extracts important features, such as concepts, measurements, definitions and so on, and incorporates them into the XML structure.

Having a well-formed regulatory repository, analysis tools are developed to help retrieval of related provisions from different domains of regulations. The theory and implementation of a relatedness analysis framework is presented. The goal is to identify the most strongly related provisions using not only a 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 a combination of feature matching. Regulations also possess specific structures, such as a tree hierarchy of provisions and the referential structure. These structures represent useful information in locating related provisions, and are therefore exploited in the analysis for a complete comparison.

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 of the prototype system.

Acknowledgments

My personal quest for a doctoral degree would have led a very lonely path if I haven't had the support from my family. I thank my parents for their continuous trusts, encouragements and beliefs in me through the ups and downs of my graduate work. I thank my brother for keeping my parents accompanied while I am selfishly abroad pursuing my dream. I am forever indebted to my wonderful partner Kevin Wong, who has taken good care of me and grown with me since our undergrad days at UCLA.

My deepest gratitude goes to my advisor and friend, Professor Kincho Law, for his selfless sharing of knowledge in research and in life in general. I thoroughly enjoyed the nurturing experience in his group, where I am given as much freedom as I wanted in this work, and at the same time, help and guidance from him whenever I needed it. From him, I learned to be both an independent thinker and a team player.

I extend my gratitude to the rest of my thesis committee. I am grateful for Professor Gio Wiederhold's help on the project and on my writing, especially his careful revisions on my papers. I thank Professor Hector Garcia-Molina for chairing my defense. I thank Professor Hans Bjornsson for agreeing to be on my committee on a short notice with his hectic schedule. I thank Professor Cary Coglianese from Harvard University for sharing his insights on this project, as well as for working out the technicalities to participate remotely in my defense.

Brainstorming and discussions with fellow researchers and colleagues represent an integral part of my work at Stanford University. I thank Professor Prabhakar Raghavan, who sat through coffee breaks with me to solve difficult problems. I truly appreciate the support from members of the Engineering Informatics Group, particularly Jun Peng, Jerry Lynch, David Liu, Chuck Han, Shawn Kerrigan, Haoyi Wang, Charles Heenan, Liang Zhou, Pooja Trivedi, Yang Wang and Jim Cheng.

I am indebted to my dearest friends who provided tremendous emotional supports throughout this work, especially Christy McBride, Gee Liek Yeo, Chika Ando, Amy Wang, Candy Wong, Bonnie Leung, Miiko Man and Saki Tong.

This research project is sponsored by the National Science Foundation, Contract Numbers EIA-9983368 and EIA-0085998. I would also like to acknowledge the support by Semio Corporation in providing the software for this research.

Table of Contents

Abstract	iv
Acknowledgments	vi
List of Tables	xii
List of Figures	xiii
List of Notations	xvi
1 Introduction	1
1.1 Problem Statement	5
1.2 Related Research in Legal Informatics	8
1.2.1 Retrieval of Legal Documents	9
1.2.2 Artificial Intelligence and Law	11
1.3 Thesis Outline.....	12
2 Repository Development	15
2.1 Related Work.....	17
2.2 The Need for Structure and Feature Identification in a Regulatory Repository	19
2.2.1 Overview of the Current Standard of Digital Publication of Regulations	21
2.2.2 Computational Properties of Regulations	23
2.3 XML Representation of Regulations.....	25
2.3.1 Basic Structure of XML Regulations.....	27

2.3.2	The Shallow Parser for Transformation into XML	29
2.4	Feature Extraction for Comparative Analysis	32
2.4.1	Concepts.....	34
2.4.2	Author-Prescribed Indices	35
2.4.3	Definitions and Glossary Terms	35
2.4.4	Exceptions.....	36
2.4.5	Measurements	37
2.4.6	Chemicals – Drinking Water Contaminants	38
2.4.7	Effective Dates.....	40
2.5	Results	41
2.5.1	Examples with Complete Set of XML Markups	42
2.5.2	Natural Tree View of Regulations	44
2.5.3	Concept Ontology	47
2.6	Summary	49
3	Relatedness Analysis	51
3.1	Related Work.....	53
3.1.1	Information Extraction, Retrieval and Mining.....	53
3.1.2	Document Comparisons.....	54
3.1.3	Hyperlink Topology.....	57
3.2	Relatedness Analysis Measure	58
3.2.1	Basis of Comparison.....	60
3.2.2	Similarity Score	62
3.3	Base Score Computation	65
3.3.1	Boolean Feature Matching.....	66
3.3.1.1	Comparisons of Concepts and Author-Prescribed Indices	66
3.3.1.2	Comparisons of Drinking Water Contaminants	67
3.3.2	Non-Boolean Feature Matching.....	68
3.3.2.1	Comparisons of Measurements.....	69
3.3.2.2	Comparisons of Effective Dates	80

3.3.3	Discussions of Other Feature Comparisons.....	83
3.4	Score Refinement Based on Regulation Structure	85
3.4.1	Neighbor Inclusion	86
3.4.1.1	<i>Psc</i> Vs. <i>Psc</i>	87
3.4.1.2	<i>Self</i> Vs. <i>Psc</i>	91
3.4.1.3	Combination of Both Analyses.....	94
3.4.2	Reference Distribution.....	95
3.5	Summary	102
4	Performance Evaluation Models, Results and Applications	106
4.1	Related Work.....	108
4.2	Comparisons to a Traditional Retrieval Model Using a User Survey.....	109
4.2.1	User Survey and the Metric	110
4.2.2	Root Mean Square Error (RMSE)	112
4.2.2.1	General Observations.....	114
4.2.2.2	Results With and Without Domain Knowledge	115
4.3	Comparisons Among Different Sources of Regulations	115
4.3.1	General Observations.....	118
4.3.2	Group 1 Comparison: ADAAG Vs. UFAS	120
4.3.3	Group 2 Comparison: UFAS Vs. IBC11	124
4.3.4	Group 3 Comparison: UFAS Vs. BS8300/STS	126
4.3.5	Group 4 Comparison: 40CFRdw Vs. 22CCRdw.....	129
4.3.6	Group 5 Comparison: 40CFRdw Vs. IBC9	132
4.4	Electronic Rulemaking.....	134
4.5	Summary	144
5	Conclusions and Future Works	148
5.1	Summary	148
5.2	Future Directions.....	150
5.2.1	Applications of A Semi-Structured Document Analysis Tool	150
5.2.2	Improving the Analysis.....	151

5.2.3 Impacts on the Making of Regulations	152
Bibliography	154

List of Tables

<i>Number</i>	<i>Page</i>
Table 4.1: RMSE of Different Combinations of β and α Parameters.....	113
Table 4.2: Average Similarity Scores Among Comparisons Using Different Feature Sets	118

List of Figures

<i>Number</i>	<i>Page</i>
Figure 1.1: Example of One Provision More Stringent Than Another	4
Figure 1.2: Example of Two Conflicting Provisions	4
Figure 2.1: Formatting Differences between Regulations in HTML	22
Figure 2.2: Source Code of the HTML Table from Figure 2.1(a)	23
Figure 2.3: Repository Development Schematic.....	26
Figure 2.4: Regulation Structure Illustrated with Selected Sections from the ADAAG...	27
Figure 2.5: XML Representation of Regulation Structure.....	28
Figure 2.6: A Schematic of the Shallow Parser	29
Figure 2.7: Example of Two Conflicting Provisions	33
Figure 2.8: Ontology Developed on Drinking Water Contaminants	39
Figure 2.9: Concept, Definition and Index Tags.....	42
Figure 2.10: Measurement and Exception Tags.....	43
Figure 2.11: Drinking Water Contaminant and Effective Date Tags.....	44
Figure 2.12: XML Regulation Rendered in Internet Explorer without a Stylesheet.....	45
Figure 2.13: Tree View of XML Regulation Rendered with an XSL Stylesheet	45
Figure 2.14: Content of a Provision Obtained by Clicking on a Node in Figure 2.13.....	46
Figure 2.15: SpaceTree Display of an XML Regulation	47
Figure 2.16: An Ontology Based on Environmental Regulations.....	48
Figure 3.1: Immediate Neighboring Nodes and Referenced Nodes in Regulation Trees .	62

Figure 3.2: Similarity Analysis Core Schematic	64
Figure 3.3: Pseudo-Code for a User-Defined Measurement Matching Algorithm	72
Figure 3.4: Illustration of an Example of a User-defined Measurement Comparison Algorithm	72
Figure 3.5: Pseudo-Code of a User-Defined Effective Date Matching Algorithm.....	81
Figure 3.6: Illustration of an Example of a User-defined Effective Date Comparison Algorithm	82
Figure 3.7: Pseudo-Code to Populate the E Matrix Using Ontology Information.....	84
Figure 3.8: Diffusion of Similarity among Clusters of Nodes Introduced by a $p_{sc-p_{sc}}$ Comparison	89
Figure 3.9: Pseudo-Code of an $f_{p_{sc-p_{sc}}}$ Computation.....	89
Figure 3.10: A Neighbor Structure Matrix to Represent Tree Structure.....	90
Figure 3.11: Diffusion of Similarity among Cluster of Nodes Introduced by an $s-p_{sc}$ Comparison	92
Figure 3.12: Pseudo-Code of an $f_{s-p_{sc}}$ Computation	93
Figure 3.13: Illustrations of an $s-ref$ and a $ref-ref$ Comparison	96
Figure 3.14: Pseudo-Code of an f_{s-ref} Computation.....	97
Figure 3.15: Pseudo-Code of an $f_{ref-ref}$ Computation.....	97
Figure 3.16: A Reference Structure Matrix to Represent References among Nodes.....	98
Figure 3.17: Illustrations of $In-In$ and $In-Out$ Reference Comparisons	101
Figure 4.1: Precision and Recall	108
Figure 4.2: User Survey	111
Figure 4.3: Related Provisions Identified Through Neighbor Inclusion.....	122
Figure 4.4: Related Provisions Identified Through Reference Distribution	123
Figure 4.5: Almost Identical Provisions Prescribed by the UFAS and the IBC	125
Figure 4.6: Mid-Ranked Related Provisions from the UFAS and the IBC.....	125
Figure 4.7: Terminological Differences Between the UFAS and the BS8300	127
Figure 4.8: Similarities Between Neighbors Imply Similarities Between Section 4.13.9 from the UFAS and Section 12.5.4.2 from the BS8300.....	128

Figure 4.9: Related Elements “Stairs” and “Ramp” Revealed Through Reference Distribution.....	129
Figure 4.10: Direct Adoption of Provisions Across Federal and California State on the Topic of Drinking Water Standards	131
Figure 4.11: Terminological Differences Between Federal and State Regulations on the Topic of Drinking Water Standards	132
Figure 4.12: Remotely Related Provisions Identified From a Drinking Water Regulation and a Fire Code.....	133
Figure 4.13: Comparisons of Drafted Rules with Public Comments in E-Rulemaking .	136
Figure 4.14: Related Drafted Rule and Public Comment.....	138
Figure 4.15: A Piece of Public Comment Not Related to the Draft.....	139
Figure 4.16: Comment Intended for a Single Provision Only	141
Figure 4.17: Suggested Revision of Provision in Comment	142
Figure 4.18: Comment on the General Direction of Draft	143

List of Notations

<i>Notation</i>	<i>Meaning</i>
A, U	Regulations
A	A single provision from regulation A
U	A single provision from regulation U
n	Number of unique entities identified in the corpus for a feature
psc	A set of parents, siblings and children
ref	A set of references
p, q	Number of provisions in a regulation
x, y	Number of nodes in a ref or psc set
I	Identity matrix
K	Term-document association matrix in LSI
s	Largest singular values to remain in LSI
D	Vector space transformation matrix
E	Matching matrix
N	Neighbor structure matrix
R	Reference structure matrix
\bar{R}	Reference structure matrix to model “in references”
λ, μ	Elements in a structure matrix
\vec{d}	Document vector
\vec{c}	Provision-concept vector

\vec{i}	Provision-index vector
\vec{m}	Provision-measurement vector
\vec{m}'	Consolidated frequency vector for measurements in a provision
\vec{d}	Provision-date vector
\vec{t}	Provision-contaminant vector
$w_{i,j}$	Weight of term i in document j
f	Similarity score
f_v	Similarity score based on the Vector model
F	Similarity score based on one feature
f_0	Base score
f_{s-psc}	Similarity score based on an $s-psc$ comparison
$f_{psc-psc}$	Similarity score based on a $psc-psc$ comparison
f_{s-ref}	Similarity score based on an $s-ref$ comparison
$f_{ref-ref}$	Similarity score based on a $ref-ref$ comparison
Φ	Similarity score matrix
Φ_0	Base score matrix
Φ_{s-psc}	Similarity score matrix based on an $s-psc$ comparison
$\Phi_{psc-psc}$	Similarity score matrix based on a $psc-psc$ comparison
Φ_{s-ref}	Similarity score matrix based on an $s-ref$ comparison
$\Phi_{ref-ref}$	Similarity score matrix based on a $ref-ref$ comparison
Φ_{final}	Final similarity score matrix
β_i	Weighting coefficient for feature i
α_0	Weighting coefficient for a base score computation
α_{s-psc}	Weighting coefficient for an $s-psc$ comparison
$\alpha_{psc-psc}$	Weighting coefficient for a $psc-psc$ comparison
α_{s-ref}	Weighting coefficient for an $s-ref$ comparison
$\alpha_{ref-ref}$	Weighting coefficient for a $ref-ref$ comparison

Chapter 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. Curious and affected citizens are entitled to and thus should be provided with the means to better understand regulations. 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 multiple documents need to be consulted and their provisions satisfied. Sections dealing with the same or similar conceptual ideas sometimes impose conflicting requirements. 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, and users often have to choose the more restrictive provision as the safest route. However, there are instances where the provisions of two applicable codes are in direct conflict. In the engineering

industry, designers often turn for resolution to reference handbooks that are produced by organizations that are independent of governing bodies. For example, for disabled access in buildings, an engineer may consult the California Disabled Accessibility Guidebook (CalDAG) by Gibbens [50]. 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. For instance, the permitting process in the design and construction industry is significantly prolonged by the increasing complexity of regulations and codes of practice. Building designers and contractors, although more knowledgeable than the general public in the domain, have yet to search through the continuously changing provisions and locate the relevant sections related to the project, then sort through potential ambiguities in the provisions. Inspectors have to go through a similar evaluation process before a permit can be approved.

The existence of multiple sources of regulations is not confined to the US. Rissland et al. [89] observed that in the European Union there is a great need for sharing and reusing of knowledge to harmonize legislation across the polyglot countries. It becomes a global challenge for companies involved in cross-border transfer, for example, between the US and EU, who must comply with multiple jurisdictions across continents [11, 88]. A survey on the cross-border data-protection laws in a number of jurisdictions in the world suggests the following:

“Widely divergent legal restrictions present a growing obstacle to multinational companies ... The more prudent multinationals want to comply with data protection laws in an efficient and coordinated manner. It’s just not obvious to them how to do it. The laws vary from jurisdiction to jurisdiction, they are constantly changing, and sometimes difficult to understand ... a surprisingly large amount of companies are still “solving” this problem by ignoring it [88].”

While multinational corporations *want* to comply with laws from different jurisdictions, most small companies simply do not have the resources to check for compliance with multiple regulations. The volume of regulations from different governing bodies makes it difficult for small businesses to locate relevant information. This in turn hinders the growth of such companies that have to devote their already-limited resources on compliance checks or budgets for violation penalties. Therefore, a tool for regulation analysis could help individuals to locate related provisions, and thus makes *understanding* of regulations easier. In addition, tools that group together related provisions could help shorten the process of compliance check against the complicated set of regulations.

The following two examples, drawn from Gibbens' interpretive guidebook [50], will put the above-described complexity into context. In the domain of disabled access regulations, Gibbens documented several "controversial issues between the [California] state and federal guidelines." He claimed that "for those of you who have been told by state or local agencies that you have nothing to worry about because California has a more stringent set of guidelines than the [Federal] ADA, I can assure you that this is not the situation." Figure 1.1 shows the first example where the California Building Code [25] is less restrictive than the Americans with Disabilities Act (ADA) Accessibility Guidelines [1]. The California code allows certain types of curb ramps encroaching into accessible parking stall access aisles, while the Federal guideline disallows encroachment into any portion of the stall.

The second example in Figure 1.2 shows two provisions that are in direct conflict. The conflict is due to the fact that the intents of the California and Federal codes are different – the California code 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 industry designers' attentions.

<p>ADA Accessibility Guidelines A4.6.3 Parking Spaces</p> <p>... The parking access aisle must either blend with the accessible route or have a curb ramp complying with 4.7. Such a curb ramp opening must be located within the access aisle boundaries, not within the parking space boundaries. Unfortunately, many facilities are designed with a ramp that is blocked when any vehicle parks in the accessible space. Also, the required dimensions of the access aisle cannot be restricted by planters, curbs or wheel stops.</p> <p>California Building Code 1129B.4.3 [No Title]</p> <p>... Pedestrian ways which are accessible to persons with disabilities shall be provided from each such parking space to related facilities, including curb cuts or ramps as needed. Ramps shall not encroach into any parking space. EXCEPTIONS: 1. Ramps located at the front of accessible parking spaces may encroach into the length of such spaces when such encroachment does not limit the capability of a person with a disability to leave or enter a vehicle, thus providing equivalent facilitation...</p>

Figure 1.1: Example of One Provision More Stringent Than Another

<p>ADA Accessibility Guidelines 4.7.2 Slope</p> <p>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.</p> <p>California Building Code 1127B.4.4 Beveled Lip</p> <p>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.</p>

Figure 1.2: Example of Two Conflicting Provisions

As such, there is a need for an information infrastructure for regulation analysis and comparisons. As suggested above, the difficulty in understanding regulations, especially the existence of multiple jurisdictions, leads to inefficiency and loss in productivity. An analysis tool that allows easy access to and help retrieval of related legal documents is beneficial to both multinational corporations and small businesses. We have provided two examples to demonstrate some of the motivations of this work in the domain of accessibility. We will formally define the problem statement and scope of this research in Section 1.1. Section 1.2 reviews the relevant literature in the field of legal informatics. An outline of this thesis is given in Section 1.3.

1.1 Problem Statement

This research addresses the difficulties in dealing with regulatory documents such as national and regional codes. In order to develop a prototype system, we focus on the limited domain of accessibility regulation. The prototype is then applied to another area of regulations, namely drinking water standards, to demonstrate its potential application to other domains. Our initial corpus includes five different 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) [1] and the Uniform Federal Accessibility Standards (UFAS) [101]. Part of a non-profit organization mandated code, the International Building Code (IBC) [63], is included as well. The remaining two accessibility regulations are the British Standard BS 8300 [21] and a selected part from the Scottish Technical Standards [97]. In the domain of drinking water standards, our corpus contains a Federal and a State regulation. Several parts are selected from Title 40, “Protection of the Environment,” of the US Code of Federal Regulations [28] and from Title 22, “Social Security,” of the California Code of Regulations [26]. Our corpus also contains a fire code from the IBC [63] to illustrate the dissimilarity

between different domains of regulations. In order to show an application on electronic rulemaking, we include in our corpus a newly drafted chapter for the ADAAG [37] and its associated public comments.

The scope of work includes the implementation of a regulatory repository, the development of a relatedness analysis system, and the evaluation of system performance and applications. Our focus is on the comparative analysis of hierarchically structured and domain-centered regulations that are heavily self-referenced. The hierarchy, domain knowledge and references together define the computational properties of regulations that differ from generic text corpora. We observe that provisions in regulations follow a parent and child hierarchy, which resemble a tree structure. Each regulation tends to be domain-specific, such as the ADAAG, which is focused on disabled access requirements. Finally, provisions frequently reference other provisions in the same regulation, but cross-regulation references are relatively few. Based on the observed computational properties, we develop a data representation format for the repository and a relatedness analysis framework.

Our repository development starts with a review of current digital publication format of regulations. Although some regulations are still only available as hardcopies, most regulations are gradually migrating to digital format; for instance, the International Building Code (IBC) [63] is available on CD-ROMs with the provisions in Portable Document Format (PDF) [77]. Some are available online in HyperText Markup Language (HTML) [61]. In this framework, the eXtensible Markup Language (XML) [41] is chosen to be the data representation format for its capability and flexibility to handle semi-structured documents. A shallow parser, which analyzes texts independent of their linguistic structures, is implemented to consolidate different formats of regulations into XML. The hierarchical and referential structures of regulations are preserved through a proper construction of XML elements. Feature extraction is performed to capture generic as well as domain-specific features in our corpus, such as concept phrases, measurements and effective dates. We use a combination of

handcrafted rules and text mining tools, in addition to input from knowledge experts, to semi-automate the process of feature extraction. A regulatory repository is complete with hierarchical and referential structures reconstructed and features identified in an integrated XML framework.

Built upon the XML framework, relatedness analysis combines feature comparisons with structure matching in order to perform a complete analysis between regulations. We define the unit of comparison to be a provision in a regulation, where the relatedness measure is a pair-wise similarity score between two provisions from different regulation trees. Existing techniques of similarity comparisons between documents, which are not specific to the computational properties assumed in regulations, could overlook important evidences of similarity. Thus, we propose to compute relatedness using not only a traditional term matching but also the incorporation of domain-knowledge through feature matching, not only a pure content comparison but also a structural matching.

Feature matching defines the computation of relatedness between two provisions based on their shared features. Available domain knowledge is incorporated in feature matching, where we propose a vector space transformation to handle potentially non-Boolean domain knowledge. The importance of domain knowledge is best illustrated with an example. In the area of accessibility, domain expert Balmer¹ clarified that “the terms “lift” and “elevator” although synonymous in definition in normal English usage have evolved into specific references in North America [6].” It is clear that domain knowledge is irreplaceable by common sense or dictionary knowledge.

Apart from feature comparisons, structural matching aims to reveal potential hidden similarities that are embedded in the structure of regulation trees. The hierarchical and referential structures of regulations are incorporated into the relatedness analysis. Neighboring provisions are compared to identify similarities that are not apparent

¹ Mr. David Balmer is a representative of the Accessibility Equipment Manufacturers Association (AEMA).

through a direct provision-to-provision comparison. Referenced provisions are compared using an analogous approach. Feature comparisons, hierarchical and referential structure matching together define the basis of our proposed relatedness analysis for regulations.

We compare our relatedness analysis system with traditional document retrieval techniques for performance evaluation. A user survey is conducted to obtain the *true* similarity between provisions. We compare the machine-predicted results with human-generated results by computing the root mean square error. Results from different types of regulations are drawn to identify potential hidden similarity through feature and structural matching. Finally, an application on electronic-rulemaking is demonstrated by comparing drafted rules with their received public comments.

1.2 Related Research in 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 [50] is one of many reference books written for compliance guidance with the accessibility code in California. It is intended to “sort out and explain the differences between the ADA & Title 24 that all California professionals must understand and apply to comply with both laws [50].” Such reference books are updated periodically to reflect the ongoing changes in the regulations. Unlike the long existence of interpretive guidelines, the introduction of information technology to aid legal interpretation is rather new. Nonetheless, the advance in technology has provided us with such tools to mitigate some of the problems mentioned earlier in this chapter.

Recently, there is a growing interest in digital government research [81-83], which brings together different communities interested in various aspects of digital government, such as Information Technology (IT) professionals, social scientists and government officials.

A variety of disciplines are covered; for example, law enforcement, government data access and management, electronic-rulemaking and so on. Among all of the digital government projects, a few focus on regulation guidance using existing IT tools. For instance, the Business Gateway² 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 [80].

The emergence of e-government has created a lot of research potential as a new application domain for IT. A few topics are suggested above, such as law enforcement [69] and e-rulemaking [29]. As this thesis is focused on regulatory analysis, we will briefly survey some related research work in this area. Section 1.2.1 reviews some literature on the retrieval of legal documents, such as data representation, retrieval based on an ontology, natural language search on case laws and so on. Section 1.2.2 investigates the application of Artificial Intelligence (AI) techniques on law, which has been an active research area long before the introduction of e-government. The careful reader might observe among the cited projects that the European research community is more active on the broader domain of legal informatics. As explained in an earlier example of cross-border data protection law, this could be attributed to the fact that the EU is composed of a number of countries with different regulatory requirements, thus the need for sharing and management.

1.2.1 Retrieval of Legal Documents

As technology advances, an increasing amount of information is digitized, among which we have government-related information such as regulations and laws. Researchers at the San Diego Supercomputer Center (SDSC) observed that governments are putting more information on the Internet, but information still remains difficult to locate and

² The Business Gateway project is formerly called the Business Compliance One-Stop project. The web address for this portal is <http://www.business.gov>.

access [8]. This is because information is distributed among several databases belonging to different government agencies. Researchers at SDSC propose to use a web-based mediation approach using XML as the data transfer protocol.

Apart from providing easier access to information distributed in many government databases, information clustering and classification remain an active research area in a legal domain. Most of the recent research focused 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 [71].” Examples of long-existing legal resource vendors based on this paradigm are LexisNexis³ and Westlaw⁴.

Data mining techniques, in particularly text mining algorithms are sought to perform a classification on legal documents [109]. Information retrieval techniques are used as well; for example, Schweighofer et al. attempted a content-based clustering and labeling of European law, taking into account the importance of different terms [94]. 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 [75]. A similar approach is used in the SALOMON project that identified and extracted relevant information from case laws, such as keywords and summaries [74]. Finally, a natural language search capability is provided by online legal research services such as Westlaw.

³ LexisNexis online legal research system can be accessed at <http://www.lexisnexis.com>.

⁴ Westlaw online legal research service can be accessed at <http://www.westlaw.com>.

1.2.2 Artificial Intelligence and Law

Berman and Hafner [12] observed that legal rights of individuals are “severely compromised by the cost of legal services,” and as a result suggested the potential of Artificial Intelligence (AI) to improve legal services. Rissland et al. also noted that “the law offers structure and constraints that may enable AI techniques to handle law’s complexity and diversity [89].” A lot of research work is focused on the application of AI, in particular, knowledge-based systems, on law [10, 84-86, 103]. For instance, Thomson et al. [98] suggested that IT can be an aid to legal practitioners in handling their cases with greater efficacy by providing expert systems such as case management tools. Some has taken a logic reasoning approach to model legal information, such as the use of deontic functions to describe legal knowledge in [102], and the well known tax law reasoning software, Taxman, described in [70].

To aid legal reasoning and interpretation, most knowledge bases develop upon a rule-based system or a network representation. A rule-based system can be developed as suggested in [64]: “the process of formulating the rulebase of the system, i.e., the collection of patterns, patternsets, and hypothesization and confirmation rules it uses, is an empirical one. It requires human rule developers to examine many stories, create rulebase components according to their intuition.” However, rule based system is criticized for its lack of flexibility, especially in logic programming, to accommodate the frequent ambiguity and vagueness in legal issues [109]. Another approach for knowledge base development is graph or network representation. It requires knowledge engineers and domain experts to create the representation structure themselves, which is often a difficult and subjective task [109]. Zeleznikow and Hunter concluded that it is “flawed [to believe] that law is straightforward and unambiguous... as to the limits of logic for modeling law [109].”

Although most agree that it is a difficult task to model legal knowledge using existing AI techniques, concentrating on a specific domain, such as compliance assistance, reduces

the problem to a solvable one. Design standards processing has been an active research area in Civil Engineering. For instance, Yabuki proposed in [108] a system to represent design standards, and to check for completeness and consistency. In addition, the need for an automated compliance checking system for hazardous waste regulation is realized in [104] and a logic-based prototype is proposed and implemented in [66].

Due to the complexity of legal language, Natural Language Processing (NLP) techniques have been considered inappropriate for legal case texts [22]. As Brüninghaus and Ashley has noted, it is because “the language used in legal documents is too complex. Sentences in the court’s opinions are exceptionally long and often have a very complex structure.” They acknowledged that many problems are still far from being solved, but also suggested that “recent progress in NLP has yielded tools that measure up to some of the complexities of legal texts [22].”

One of the complexities of legal language is its open texture property. Gardner addressed the open texture problem, or in other words, incomplete definition of many legal predicates, of the law in [47]. Examples are phrases such as “reasonably certain” and “a reasonable time” that are intentionally or unintentionally arguable in meaning. It is suggested that “framers of legal rules have often abandoned clear directives in favor of open textured rules [12].” Some goes further to attribute the difficulty in modeling law using AI techniques to the open texture problem: “Legal concepts, therefore, cannot be modeled by unassailable, universally quantified necessary and sufficient conditions. In a word: they are incurably open-textured [89].”

1.3 Thesis Outline

The objective of this research is to develop an integrated framework for regulation representation to support a comparative analysis that facilitates user understanding and retrieval of related provisions from different sources of regulations. Industry designers,

planners, policy-makers as well as interested individuals are potential users who can benefit from the exploration of relevant provisions provided by this regulatory framework.

The rest of this thesis is organized into the following four chapters. Multiple facets of research constitute this cross-disciplinary study of regulatory and related legal information, and we have briefly examined some related literature on the broader perspective of legal informatics in this chapter. Research work related to specific chapters will be introduced in the first section of that chapter.

- Chapter 2 presents the development of a regulatory repository. We select several regulations from different sources to be included in our corpus. To populate the repository, a shallow parser is implemented to consolidate different formats of regulations, for example, HTML or PDF, into a XML representation. The shallow parser performs feature extraction on the XML regulations to include available domain knowledge as well as generic features. Examples of features and their representation format in XML are presented.
- Chapter 3 discusses the theory of a relatedness analysis utilizing different computational properties specific to regulations. The semantics of *relatedness* and *similarity* are investigated, and a similarity score is defined as the metric of relatedness between two provisions from different regulation trees. The computation of the base score, defined as a linear combination of different feature matching, as well as subsequent score refinements, based on the structures of regulations, are explained. The mathematical model is defined using a compact matrix representation of the scores and the structures.
- Chapter 4 examines the system performance, results and applications. Performance evaluation is conducted through a user survey, where user-ranked relatedness between provisions is compared with the machine-predicted ranking. Our system shows a reduction in error compared to a traditional retrieval model.

Results of comparisons between different groups of regulations are analyzed, with examples of related provisions documented per group. Finally, application to the domain of electronic-rulemaking is experimented. Various results obtained from the comparison between a drafted rule and its associated public comments are shown.

- Chapter 5 summarizes the development of this integrated framework for regulation representation and analysis, which lays down important groundwork for a rich set of future research in this area. Results of comparisons on different groups of regulations using the computation characteristics are listed, along with observed limitations of the current implementation. Suggestions of potential future research directions are provided.

Chapter 2

Repository Development

The complexity and diversity of regulatory documents make understanding and retrieval of regulations a non-trivial task. In particular, the existence of multiple jurisdictions, such as the Federal and state governments, leads to differences in formatting, terminology and context among regulations. Affected as well as curious citizens are entitled to easy access, retrieval and comparisons of different regulations, but in reality, we lack the infrastructure as well as tools to support such kind of explorations. Therefore, there is a need for a consolidated repository for regulatory documents such that tools can be developed to better understand and analyze regulations across different sources.

This chapter describes the development of a repository for regulatory documents from a variety of sources. The goal is to provide a consolidated platform for regulatory analysis and comparisons, and to provide users with an environment to retrieve and browse through relevant provisions with ease. For example, a simple ontology is developed to aid exploration of related provisions; however, an in-depth similarity analysis is imperative for a more sophisticated comparison. This chapter presents the design and implementation of a consolidated repository, which prepares for the development of analysis tools to be described in Chapter 3. This chapter is organized as follows: First, research work related to repository development, in particular, feature extraction and

document representation formats, is introduced in Section 2.1. The selected sources of regulations, including disabled access and drinking water standards, are listed in Section 2.2. In Section 2.2.1, we give a brief overview of the current representation format for regulations such as Portable Document Format (PDF) [77] and HyperText Markup Language (HTML) [61], and discuss their advantages and disadvantages. Section 2.2.2 explains why a new data format is needed, where several important computational properties of regulations are observed. Section 2.3 then gives the basic structure of a simple eXtensible Markup Language (XML) [41] representation of regulations. In this section, we first justify our selection of XML as the basic representation format; namely XML is capable of encapsulating the computational properties, such as adding structure and domain knowledge to regulations. To minimize human effort in data format conversion, we describe the development of a shallow parser to transform regulations into XML format as shown in Section 2.3.2.

Section 2.4 introduces feature extraction, which is an important process to aid document analysis. Features, which are defined as the evidences that identify related provisions, are extracted and incorporated into the XML element representation of provisions. Two different types of features, namely generic features and domain-specific features, are extracted with the help of a software tool and parsers developed for this task. Sections 2.4.1 to 2.4.7 give examples of features extracted from the domains of accessibility and drinking water control. Results are shown in Section 2.5, where regulations can be viewed in their natural tree hierarchy with nodes representing provisions. Relevant provisions can also be retrieved from an ontology developed based on the extracted concepts.

The proposed regulatory repository is implemented mostly in Java, with some utilizations of an open source project Jakarta Lucene [2] that provides simple text indexing capability. Some parts of the shallow parser are developed in Perl, and a software tool, Semio Tagger [95], is used for concept extraction. Several XML rendering tools have

been experimentally used to show different display formats for regulations. Help with implementation and development is credited in footnotes in appropriate sections.

2.1 Related Work

In this section, we will review some groundwork for the development of a text repository. The definition of and the reasons for feature extraction in a large data set are discussed. Relevant work on document structure extraction as well as the use of eXtensible Markup Language (XML) is introduced. A fair amount of work on keyword search is cited, which explains our focus on provision matching instead of query matching.

Feature extraction is an important step in repository development when the data is voluminous. Feature extraction is a form of pre-processing, for example, combining input variables to form a new variable. Often features are constructed by hand based on some understanding of the particular problem being tackled [15]. Automation of this process is also possible. In particular, in the field of Information Retrieval (IR), software tools exist to fulfill “the task of feature extraction ... to recognize and classify significant vocabulary items [15].” As suggested, one potential feature is key phrases. IBM’s Intelligent Miner for Text [36] and Semio Tagger [95] are examples of fully automated key phrase extraction tools. Most commercial tools use a combination of linguistic heuristics, pattern matching and lexical analysis for this task.

As suggested above, an example of feature extracted from texts is key phrases that are important within a corpus. Key phrases capture the sequencing information of terms, and experiments have shown that phrases can convey more important information than the terms separated. For example, as pointed out by Jones and Willett [64], “joint venture is an important term in the Wall Street Journal database, while neither joint nor venture are important by themselves. In fact, in a 800+ Mbytes database, both joint and venture would often be dropped from the list of terms by the system because their idf (inverse

document frequency, which represents the uniqueness of a term in the entire corpus) weights were too low.”

With the understanding of *what* feature extraction represents, we move on to the question of *why* we need it. The curse of dimensionality [9] refers to the exponential growth of hypervolume as a function of dimensionality. When the data dimension grows, the curse of dimensionality leads to inconclusive comparisons between points in space, since all points seem as far as one another. One of the motivations for feature extraction is to avoid the curse of dimensionality. The goal is to reduce data dimensions by including only the important features.

In a textual domain, feature extraction denotes the inclusion of important phrases or other features instead of all of the terms. The *joint venture* example explains the importance of feature extraction in a large corpus. Indeed, feature extraction is particularly useful in a domain-center corpus than among general-purpose texts. For instance, laws are developed based on specific areas of application and jurisdiction, where a general index term extraction would fail to capture any domain knowledge that are available. Example of domain knowledge includes ontologies and field-specific handbooks, which will be introduced in Sections 2.4.3 and 2.4.6 respectively. As we shall see in later sections, domain knowledge can be gracefully incorporated into the corpus through feature extraction.

In addition to data cleaning and pre-processing such as feature extraction, there is a need to extract structure out of documents, such as chapters and subsections as well as hyperlinks and references. There are multiple studies on extracting document structure; for example, a road map approach to exploit different tree structures of documents is presented in [105]. Kerrigan illustrated in [66] the extraction of references using a tabular parsing system with a context-free grammar. In addition, searching in structured or semi-structured data is a major research focus recently, such as among data in the eXtensible Markup Language (XML) [41] – one that has become almost the de facto representation for semi-structured data. Database management systems for XML has

been developed [53], with some work focusing on searching and retrieval techniques based on structured text [23], allowing a mix of queries to search on content and structure [4]. Ganesan et al. examined the hierarchical structure to compute similarity, assuming that attributes are confined to the leaves in the hierarchy [46]. Query language for structured text search has been developed and implemented in [27].

Due to the proliferation of the Internet, an extensive amount of research focusing on retrieving relevant documents based on a keyword search has been done [13]. Well-established techniques such as query expansions [62, 90] have been deployed to increase retrieval accuracy, with a significant amount of subsequent developments [3, 32, 87, 107] to improve performance. Thus, most repositories are equipped with a search and browse capability for viewing and retrieval of documents. In this research, we assume that at least one relevant document will be located by the user either with a keyword search, a SQL⁵-like query or by browsing through an document classification hierarchy such as one supported by Semio Tagger [95]. From there, related documents are suggested to the user by our system. In essence, we focus on refining the back end comparison technique for documents rather than matching queries at the front end.

2.2 The Need for Structure and Feature Identification in a Regulatory Repository

In this section, we will explain the need for structure and feature identification in a regulatory repository based on the example domains of accessibility and drinking water regulations. Documented below is a list of regulations and codes of practice selected from these two areas for repository development.

⁵ SQL is a relational database query language.

In the domain of disabled access, two US Federal documents are incorporated into our corpus: the Americans with Disabilities Act Accessibility Guidelines (ADAAG)⁶ [1] and the Uniform Federal Accessibility Standards (UFAS)⁷ [101]. In addition, Chapter 11 of the International Building Code (IBC) [63], titled “Accessibility,” is included to reflect the similarity and dissimilarity between federal regulations and private non-profit organization mandated codes. To illustrate the differences between American and European regulations, we include in our corpus the British Standard BS 8300, titled “Design of Buildings and Their Approaches to Meet the Needs of Disabled People – Code of Practice” [21], as well as a selected part from the Scottish Technical Standards (Part S on “Access to and Movement within Buildings, and Protective Barriers”) [97].

In the domain of environmental protection, we focus on comparisons between Federal and State drinking water regulations. Parts 141 to 143 on national drinking water standards are selected from the US Code of Federal Regulations Title 40 (40 CFR titled “Protection of the Environment”) [28], along with drinking water provisions from the California Code of Regulations Title 22 (22 CCR titled “Social Security,” Division 4 on Environmental Health) [26]. A fire code from the IBC, Chapter 9 titled “Fire Protection Systems,” is included as well to demonstrate the dissimilarity across different domains.

Apart from comparisons between different sources of regulations, we intend to apply our system on other domains as well, such as electronic rulemaking (e-rulemaking). E-rulemaking defines the process in which the electronic media, such as the Internet, is used to provide a better environment for the public to comment on proposed rules and regulations. Therefore, to compare the drafted rules with their associated public comments, we include in our corpus a newly drafted chapter for the ADAAG prepared by the US Access Board [37], titled “Guidelines for Accessible Public Rights-of-way” [37].

⁶ The ADAAG is published as Appendix A to Part 36, entitled “Nondiscrimination on the Basis of Disability by Public Accommodations and in Commercial Facilities,” of Title 28, entitled “Judicial Administration,” of the Code of Federal Regulations.

⁷ The UFAS is adopted by the General Services Administration (GSA) in 41 CFR 101-19.6, and by the Department of Housing and Urban Development (HUD) in 24 CFR part 40.

This proposed chapter received a large amount of public comments which are also incorporated in our corpus for analysis purposes.

Based on the above list of regulatory documents, we give a brief overview of the current digital publication standards for regulations in Section 2.2.1. It explains why a different consolidated format is needed for regulation representation. Several important computational properties of regulations are also noted in Section 2.2.2, which explains why there is a need for structure and feature identification in this domain.

2.2.1 Overview of the Current Standard of Digital Publication of Regulations

A brief survey on the electronic publications of regulations and supplementary documents shows that there is currently no central format for such publications. Some of the regulations, e.g., the ADAAG [1] and the UFAS [101], are provided in HyperText Markup Language (HTML) [61] format. Some are stored in Portable Document Format (PDF) [77], such as the Scottish Standards [97]. Indeed, even within one formatting language, there exists no central publishing standard. For instance, Figure 2.1 shows an example of two HTML regulations, namely the UFAS and the UK Disability Discrimination Act (DDA) [35]. The first formatting difference between the UFAS and the DDA lies in the section numbering style where one uses full path, such as 4.2(1), while the other lists only a partial path, e.g., listing only the number (1) instead of 4.2(1). In addition, compared to a plain HTML format adopted by the ADAAG, the entire DDA is written as a HTML table with the first column being the section title, and the second column is the main text. Within PDF regulations, examples of formatting differences include the single-columned BS 8300 [21] and the double-columned IBC [63].

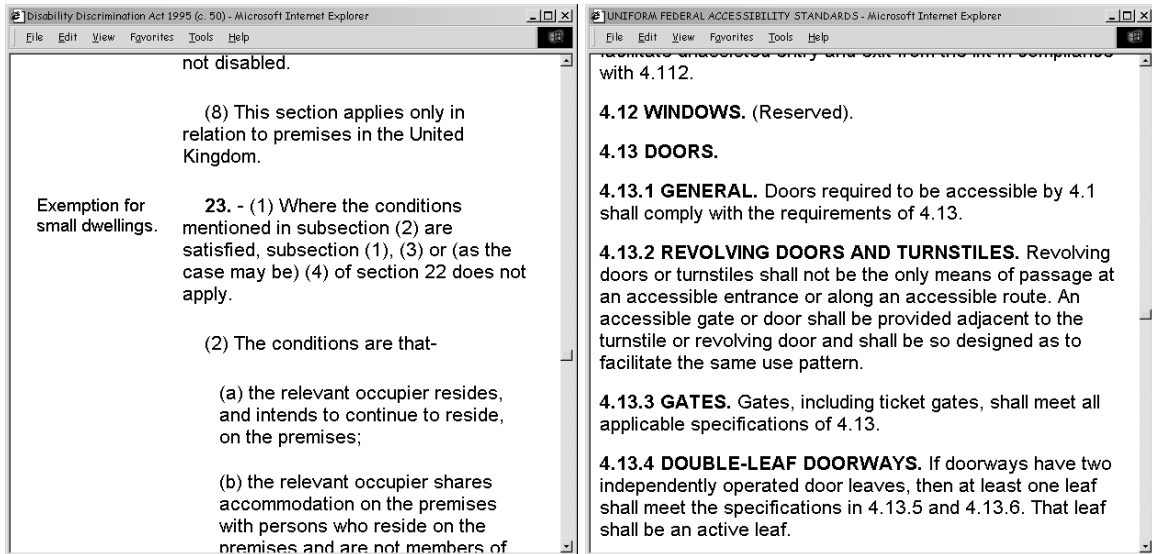


Figure 2.1: Formatting Differences between Regulations in HTML

As shown above, the current format and language for digital regulation allow for wide differences in formatting and styling, which cater to the needs of different regulatory agencies. For instance, figures and drawings are common in accessibility regulations, and they are directly embedded in PDF regulations just as regular text. Tables prevail in environmental regulations for chemical concentration requirements, and HTML is best in bundling data with specific rendering style, such as column width. Both HTML and PDF are convenient languages for wrapping data with style for purpose of public distribution. However, the same flexibility also leads to difficulty in reusing this digital information: as shown in Figure 2.2, the source code of a HTML table can be messy and the structure rendered on a browser is undecipherable on the source.

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 structural characteristics that are not found in generic text, which makes them interesting to analyze.

- Regulations assume a deep tree hierarchy. They are semi-structured documents that are organized into a tree structure; for example, Section 11.4.5(a) can be interpreted as a subpart or a child node of Section 11.4.5, which makes it a sibling of Section 11.4.5(b) as well. This regulatory structure is crucial in understanding contextual information between sections.
- Sections are heavily cross-referenced within one regulation. For instance, Section 11.4.5(a) can refer to Section 8.2 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 [101].” 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.

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 [28] is focused on environmental protection. Domain experts from the field of environmental protection might identify other computational properties, such as chemical properties from the periodic table, that they want to annotate for purpose of understanding as well as analysis in a regulatory infrastructure.

As a result, we need a consolidated format that is capable of incorporating all of the above computational properties of regulations, instead of a data and style bundle such as HTML or PDF. In particular, a comprehensive regulatory infrastructure should be able to include both structural and feature information. Finally, an ideal representation format for regulations should encapsulate provision information as well. This is due to the voluminous nature of regulations, where analysis only makes sense on a per provision basis. For example, a similarity comparison between the entire ADAAG and the entire UFAS, both over a hundred pages, would likely complicate the analysis instead of enhance understanding.

2.3 XML Representation of Regulations

XML [41] is chosen as the communication model because of its expressiveness to represent the organization of provisions, its ability to format semi-structured data and its flexibility in tagging compared to HTML. XML is a simple, yet flexible electronic publishing media for use over the Internet. It is similar to HTML, with the flexibility to define new tags and metadata, such as `<feature>` as a feature element, in addition to pure styling tags, such as `` as a bold font tag. It is more stringent than HTML which is relatively forgiving in dealing with unclosed tags and careless formatting. Instead, an XML document has to be well-formed in order to be rendered; mismatched tags cannot

be displayed. In essence, XML is a balance between a strictly structured data representation, such as a relational database, and a completely free-form format, such as PDF.

In consolidating regulations into XML format, provisions can be first encapsulated as an XML node. The tree hierarchy of regulations can be captured by properly structuring these XML nodes. Features, including domain-specific information, can be easily added as extra XML elements as well. Most of the HTML tags are still valid in XML, for example, tables in HTML can be directly embedded in XML. In the following sections, we briefly describe the basic XML structure and the development of a shallow parser to transform regulations into this XML format. Section 2.4 explains the process of feature extraction to refine XML regulations for analysis purposes. We shall refer to Figure 2.3 below for an illustration of the different components in the repository development.

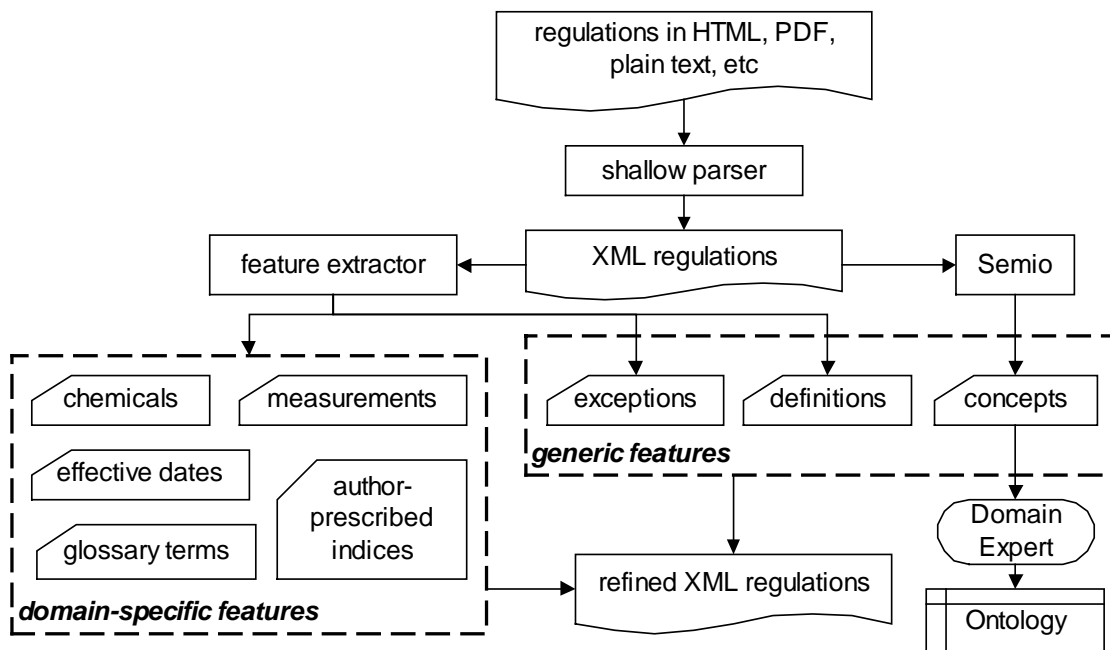


Figure 2.3: Repository Development Schematic

2.3.1 Basic Structure of XML Regulations

The basic XML structure mirrors the natural hierarchy of regulations as shown in Figure 2.4. The root of this XML tree is a `<regulation>` node, with the regulation name (e.g., “Americans with Disabilities Act Accessibility Guidelines”) and type (e.g., “Federal”) defined as attributes. Let’s take Section 4.7.4 from Figure 2.4 as an example. The unit of extraction is provision, and therefore Section 4.7.4 is represented as one single XML element. There is one terminological clarification - we use the terms “section” and “provision” interchangeably to represent the unit of extraction as well as the unit of comparison to be discussed in Chapter 3. 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.

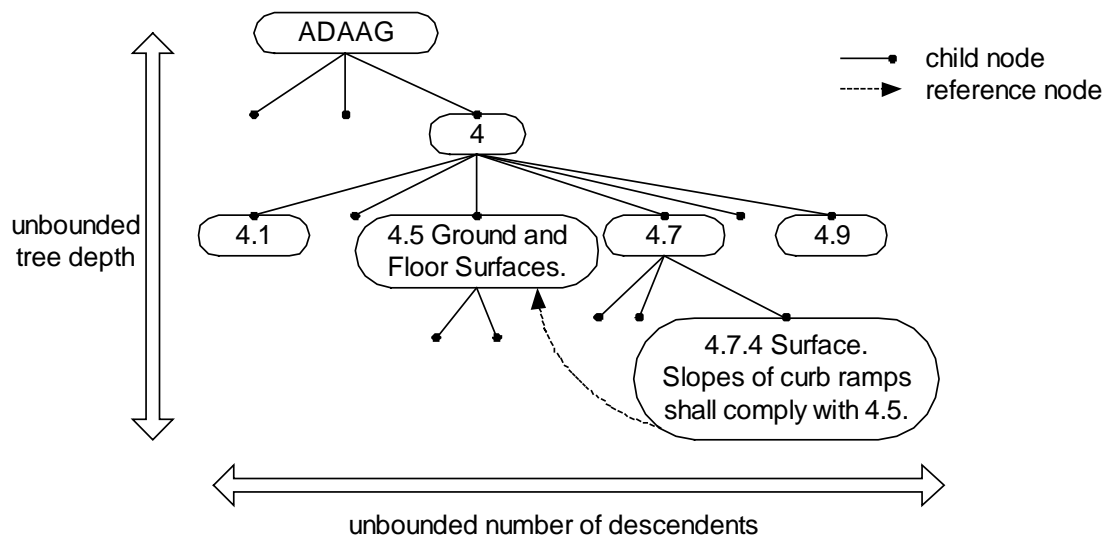


Figure 2.4: Regulation Structure Illustrated with Selected Sections from the ADAAG

Each provision is represented as an XML element called `<regElement>`. Envelope information such as section ID (i.e., 4.7.4) and section title (i.e., “Surface”) are extracted

as attributes in a `<regElement>`. The actual content of a provision (i.e., “Slopes of curb ramps shall comply with 4.5”) is placed in a subelement of the `<regElement>`, and is termed `<regText>`. Referential structure is also captured in a subelement called `<reference>`, with the ID of the referenced section (i.e., 4.5) and its reference frequency (i.e., 4.7.4 references 4.5 once) as attributes. Non-structural characteristics of regulations, or *feature* as we defined before, can be added as subelements as well. Finally, the tree structure of regulations is captured by properly structuring these `<regElement>`s. For instance, Section 4.7.4 is a subpart of Section 4.7, and therefore it is placed as a child node of the `<regElement>` representation of Section 4.7. Figure 2.5 illustrates the basic XML structure. As provisions tend to be lengthy, only excerpts of the regulation is shown here with ellipsis marks to indicate omitted parts.

```
<regulation id="adaag" name="Americans with Disabilities Act
Accessibility Guidelines" type="Federal">
  ...
  <regElement id="adaag.4" name="Accessible Elements and
  Spaces: Scope and Technical Requirements">
    ...
    <regElement id="adaag.4.5" name="Ground and Floor
    Surfaces">
      ...
    </regElement>
    ...
    <regElement id="adaag.4.7" name="Curb Ramps">
      ...
      <regElement id="adaag.4.7.4" name="Surface">
        <reference id="adaag.4.5" num="1" />
        <regText>
          Surfaces of curb ramps shall comply with 4.5.
        </regText>
      </regElement>
    </regElement>
    ...
  </regElement>
  ...
</regulation>
```

Figure 2.5: XML Representation of Regulation Structure

2.3.2 The Shallow Parser for Transformation into XML⁸

Data cleaning and consolidation can easily account for up to 90% of the total data mining time especially when there exist multiple data sources [39]. Therefore, our goal is to minimize human effort in data cleaning, which is unavoidable as will be explained below, and to automate the consolidation process as much as possible. To this end, a shallow parser is developed to extract and reconstruct the regulation's natural hierarchy from HTML or PDF to XML, since our corpus is composed of these two initial data formats. Deep parsing, which gives semantic structure to texts, is not necessary here as we focus on information retrieval rather than linguistic analysis of regulatory documents. A schematic of the shallow parser is shown in Figure 2.6.

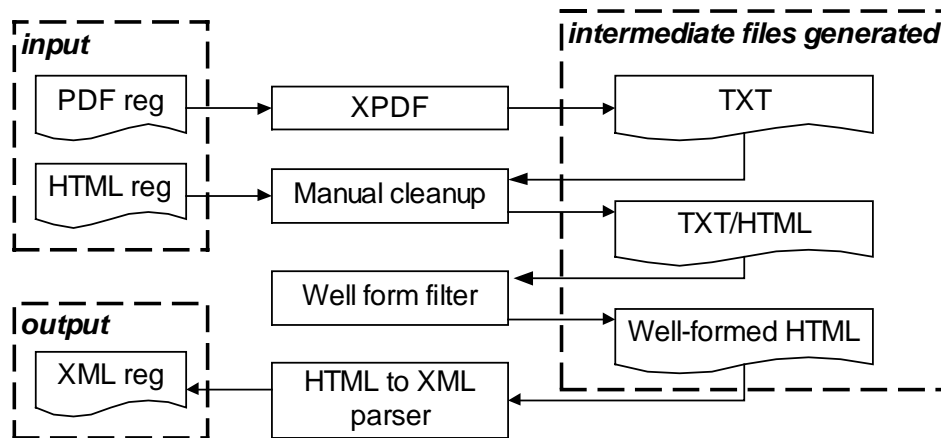


Figure 2.6: A Schematic of the Shallow Parser

As shown in Figure 2.6, the shallow parser takes a HTML or PDF regulation as input, and outputs the corresponding well-formed XML file, with added tags and removed formatting. For regulations in PDF, the process starts by transforming the encoded PDF

⁸ Ms. Pooja Trivedi and Mr. Haoyi Wang, both graduate students at Stanford University, helped in part of the shallow parser implementation.

document into machine understandable plain text format using the Xpdf text extractor [106]. It suffices to say that Xpdf is an open source software for PDF file conversion that handles both single-columned and double-columned PDF documents. However, these plain text regulations are still not ready for parsing and manual cleanup is unavoidable. For instance, the HTML tags can be erroneous such as `` instead of ``; plain text converted from PDF can include dangling figure captions. Our design of the shallow parser minimizes human effort, since most recognizable errors or inconsistencies are handled by the next automated step – a well form filter.

HTML is forgiving and does not have to be well-formed in order to be rendered in most browsers. As a result, a well form filter is developed to transform plain text and HTML documents into well-formed HTML documents. For instance, most PDF regulations are embedded with headers, footers and page numbers, which also remain in the plain text version. These decorative texts are automatically removed by the filter, in addition to illegal characters such as “<” and “>”. The filter also handles non-closing tags in HTML, such as a paragraph tag `<p>` without its ending tag `</p>`, which is not allowed in XML. There are inconsistent anchor tags to delineate provisions in most HTML regulations, for example, Section 2.1 is usually represented as `Section 2.1` so that other sections can link to it. Not surprisingly, sections are anchored in different styles while some are not anchored at all. Therefore, handcrafted rules are developed to locate section beginnings, and the filter identifies, unifies and adds anchor tags to sections. Our set of handcrafted rules uses pattern matching to delineate provisions by recognizing section IDs followed by section titles. Plain text is handled in a similar fashion, as it is simply a HTML document with no anchor tags.

After the filtering process, these well-formed HTML documents are fed into a HTML to XML parser. The parser extracts information such as section titles and references, as well as structures section elements to reflect regulation’s natural hierarchy. It starts by extracting section ID and section title through pattern matching, for example, delineating

section title by matching its end of line character that follows after a heading. For example, upon matching “Section 2.1 Curb Ramps,” the parser creates a new node with the syntax `<regElement id="adaag.2.1" name="curb ramps">`. The actual content of a section is placed in a `<regText>` subelement. The parser then structures elements properly by differentiating the relationship between consecutive nodes. The following steps explain how this is performed:

1. The parser first takes the current section ID and matches the successor node’s ID. If the successor’s ID is the current ID concatenated with a dot or a bracket, followed by a single number or letter, the successor is created as a child node of the current node. For example, Section 2.1(a) is a child of Section 2.1.
2. If the successor is not a child node, it is tested to see if it is a sibling node. We increment the last digit or letter of the current ID by one, and if this incremented ID is now equal to the successor’s, the successor is a sibling node. For example, Section 2.1(b) is a sibling of Section 2.1(a).
3. If the successor is neither a child nor a sibling node, it bears no immediate relationship to the current node. The current node is closed with a `</regElement>` tag. For example, Section 2.2 follows the closing of Section 2.1(b).

After the entire regulation is structured as an XML tree, references are extracted and tagged as subelements of `<regElement>` nodes. The parser first generates a list of all existing section IDs, and performs pattern matching on the contents of all `<regText>` elements to locate references based on the existing list. Since most regulations only reference its own provisions, for instance, the ADAAG does not reference the UFAS at all, only internal references are extracted. Indeed, it would be impossible to automate reference extraction for random references out to documents not in the corpus. Simple expansion of references is performed upon locating phrases such as “Sections 2.1 through 2.4,” where in this case all sections in between Section 2.1 and Section 2.4 are included.

A frequency count is kept for each reference per each section as well, since the more a section is linked to, the more important it is in the context of the current section. An example for a reference to Section 2.1 twice would be `<reference id="adaag.2.1" num="2" />`. A complete example of the outputted XML regulation is illustrated as shown earlier in Figure 2.5.

2.4 Feature Extraction for Comparative Analysis

After data cleaning and consolidation, features are extracted and added to the repository. As explained above, features represent non-structural characteristics from regulations, and they could be domain-specific information. As we are interested in analyzing regulations, features are defined as evidences that signal relatedness or similarity in this context. Once extracted and highlighted, features become a handy reference for one to follow through provisions. For example, as will be introduced in Section 2.5.3, an ontology can be developed on top of the extracted concept features to aid retrieval of relevant provisions.

We define two different types of features. As shown in Figure 2.3, we have 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 shown in Figure 1.2 [50], reproduced here as Figure 2.7, best illustrates the reason for including both types of features. Two directly conflicting provisions from the ADAAG [1] and the California Building Code (CBC) [25] are shown. This conflict is due to the fact that the ADAAG focuses on wheelchair traversal while the CBC focuses on the visually impaired when using a cane, and 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.

<p>ADA Accessibility Guidelines <u>4.7.2 Slope</u> 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.</p> <p>California Building Code <u>1127B.4.4 Beveled Lip</u> 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.</p>

Figure 2.7: Example of Two Conflicting Provisions

A software tool and parsers developed for this task are used to extract and add features as additional tags in sections where they appear. Some of the features can be applied generically on other sets of regulations, while some are specific to our domains; for instance, numeric measurement might only make sense in the domain of disabled access code but not in human rights law. In addition, what defines *evidence* in a certain domain of regulations is also subjected to the knowledge engineer's judgment. In this context, we strive to be as generic as possible, and all of the extracted features can be easily extended to other engineering domains as well. Each of these features is discussed in the following sections with examples to illustrate the XML representation of feature elements.

2.4.1 Concepts

Traditional Boolean model or Vector model in the field of Information Retrieval (IR) provides a mechanism for text analysis. Indexing the texts using all of the words, except stopwords, generates a huge multi-dimensional space with one axis representing one word [91]. Using singular value decomposition (SVD) as the dimensional reduction tool, which will be discussed in Section 3.1.2, similar terms are supposed to be reduced to a “concept” on a single axis. However, SVD is computationally intensive and the initial sparseness of the matrix is destroyed after dimension reduction. As an alternative to the bag-of-word Vector model and the SVD technique, we use concepts. Concepts are defined as key phrases formed by pulling together terms. The number of phrases identified is relatively small compared to that of traditional index terms, and they also allow us to capture sequencing information on words.

There are commercial software products available for key phrase extraction, and Semio Tagger [95] is one of them. Based on linguistic analysis and other techniques, the Tagger identifies a list of noun phrases, or *concepts*, that are central to the corpus. If we take the ADAAG and the UFAS as an example, they generate just over a thousand concepts together. Below is an example of a typical concept element identified in a corpus of accessibility regulations.

```
<concept name="maneu v clearanc" num="2" />
```

Each provision is tagged with its concepts, for example, “maneuvering clearances,” along with the corresponding count of appearances of that concept (num). A parser is developed to tag provisions with their associated concepts, and to keep a frequency count of concept appearances. To increase the number of matches and to consolidate the vocabulary, both the concepts and the texts in the provision are stemmed with Porter’s Algorithm [78] before matching; for example, the word “clearances” is stemmed to its root form “clearanc”.

2.4.2 Author-Prescribed Indices

Machine-generated phrases, such as those obtained from the method described in Section 2.4.1 above, represent a good measure of important concepts in the body text of provisions. Another source of potentially important phrases comes from author-prescribed indices in reference books or even the regulation itself. This type of human-written information sometimes can be more valuable than machine-generated phrases.

Index terms from Chapter 11 of the IBC [63], titled “Accessibility,” are tagged against the repository. The syntax is identical to a concept tag except that the element name is replaced with `index`. Below is an example of an `<index>` tag.

```
<index name="valet park" num="1" />
```

Here, the phrase “valet parking” comes from the list of index terms in the IBC. A shallow parser is used again to locate and tag index terms to appropriate provisions where they appear, and a frequency count is kept as well for each phrase per provision. Again, matching is performed on stemmed index terms and texts.

2.4.3 Definitions and Glossary Terms

In regulation documents, there is often a designated section in an early chapter that defines the important terminologies used in the code, such as Section 3.5 in the ADAAG, titled “Definitions.” These human-generated terms are more likely to convey key concepts than machine-extracted concepts. In addition, the definition of a term gives meaning to a term, which is useful for comparisons. Below is an example of a `<definition>` element, which shows the definition of the term “accessible” as given in Section 3.5 of the ADAAG:

```
<definition>
  <term> Accessible </term>
  <definedAs> Describes a site, building, facility, or
  portion thereof that complies with these guidelines.
  </definedAs>
</definition>
```

Similarly, engineering handbooks provide in the glossary important terms used in the field. For instance, the Kidder-Parker Architects' and Builders' Handbook provides an 80-page glossary that defines “technical terms, ancient and modern, used by architects, builders, and draughtsmen” [67]. Below is an example of a `<glossaryDef>` element that defines the term “return head”.

```
<glossaryDef>
  <term> Return Head </term>
  <definedAs> The continuation of a molding, projection,
  etc., in an opposite direction. </definedAs>
</glossaryDef>
```

The difference between a `<definition>` and a `<glossaryDef>` is that definition comes from the regulation itself, while `glossaryDef` comes from outside sources other than the regulation. The syntax of the element is exactly the same, and both are extracted by a shallow parser developed for this task.

2.4.4 Exceptions

Exceptions are a special property of regulations – they amend the body text of provisions. They can be regarded as part of the body text in `<regText>`; however, mixing regular content with exceptions does not help analysis of provisions, since exceptions are fundamentally negated provisions. Therefore, they are captured in an `<exception>` element as follows.

```
<exception>  
    This does not apply to parking provided for official  
    government vehicles owned or leased by the government  
    and used exclusively for government purposes.  
</exception>
```

The above example is an exception from a section in the UFAS. Extracting and highlighting this information can potentially help one to locate possible compliance leeway in exceptional cases.

2.4.5 Measurements

In accessibility regulations, measurements play a very important role; in particular, they define most of the conflicts. For instance, one provision might suggest a clear width of 13 to 15 inches, while another one might require 16 to 17 inches. It is therefore crucial to identify measurements and the associated quantifiers if there is any. In our context, measurements are defined to be length, height, angle, and such. They are numbers preceding units. Quantifiers are phrases that modify a measurement, such as “at most,” “less than,” “maximum” and so on. Quantifiers can be reduced to a root of either “max” or “min”; for example, the terms “at most” and “less than” are maximum requirements, thus both reduce to “max.”

Similar to concept tagging, our parser takes a list of units, quantifiers and their roots as input. This list can be easily generated by a knowledge engineer or a careful reader of the regulation. Handcrafted rules are developed to match synonymous measurements, such as parts per million (ppm) and milligrams per liter (mg/L). In the domain of disabled access and drinking water standards, a non-exhaustive list of units and quantifiers is selected below to illustrate measurement extraction:

- Units: inch, foot, degree, second, pound, parts per million (ppm), parts per billion (ppb), parts per trillion (ppt), parts per quadrillion (ppq), nephelometric turbidity unit (NTU).

- Quantifiers: minimum, maximum, at least, at most, higher than, greater than, more than, less than, steeper than, fewer than, faster than, or less, up to, below, over, exceeding.

We first identify numbers followed by units, for example, the number 2 followed by the unit lbf (pound-force) as in 2 lbf. The quantifier is an optional attribute in a measurement tag and is identified if it appears in the vicinity of the measurement. Negation, if appearing right in front of the quantifier, is extracted as well and the final quantifier is reduced to its root “max” or “min”; an example is shown below for a measurement of up to two pounds that appears once in the provision.

```
<measurement unit="lbf" size="2" quantifier="max"  
num="1" />
```

In addition, range measurement, for example, 2 to 3 inches appearing twice, is identified and is shown as follows:

```
<measurement unit="inch" size1="2" size2="3" num="2" />
```

Again, a shallow parser is developed specifically for this task. The measurement feature can be interpreted as a domain-specific knowledge, since it is developed primarily for accessibility and drinking water regulations. However, this feature can be easily extended to other domains by incorporating other types of measurements, such as Volt, Watt and so on for energy bills.

2.4.6 Chemicals – Drinking Water Contaminants

As we focus on drinking water standards in environmental regulations, certain chemicals play an important role in this domain. In particular, the US Environmental Protection Agency (EPA) publishes an index of national primary drinking water contaminants [79]. This list contains about a hundred potential drinking water contaminants; examples include “trans-1,2-dichloroethylene,” “vinyl chloride” and so on. An ontology is

developed based on the index of drinking water contaminants published by the EPA as well as supplementary materials, and an excerpt is shown in Figure 2.8⁹. A category name is preceded by an exclamation mark, while elements belonging to the category are signaled with a plus sign. For instance, a domain expert can easily codify synonymic/acronymic information such as “total trihalomethane” and “tthm” as shown in the ontology. This further illustrates the need to incorporate domain knowledge, where most intelligent mining tools are likely to fail to identify such type of information even with the help of a dictionary¹⁰.

```

!Disinfectants and Disinfection-byproducts
  !Disinfectants
    ...
    !Chlorine
      +chlorine
      +cl2
      +hypochlorite
      +hypochlorous acid
  !Disinfection Byproducts
    +d/dbp
    +d/dbps
    +dbp
    +dbps
    ...
    !Total Trihalomethanes
      +trihalomethane
      +tthm
      +tthms
    ...

```

Figure 2.8: Ontology Developed on Drinking Water Contaminants

To incorporate this piece of domain knowledge, the parser takes the ontology as a flat list and tags the drinking water contaminants as <dw> subelements in provisions where they

⁹ This is a modification of Bill Labiosa’s ontology work.

¹⁰ In this particular example, the term “tthm” cannot be found in either Webster or Oxford dictionary. Merriam-Webster Collegiate Dictionary is a product of Merriam-Webster, Inc.; Oxford English Dictionary is a product of Oxford University Press.

appear. As shown below, stemming and frequency counting are performed as in `<concept>` and `<index>`.

```
<dwc name="total coliform" num="1" />
```

Drinking water contaminants serve a similar purpose as author-prescribed indices, where human-generated knowledge should be included in provisions when available. The phrase “total coliform” might be extracted as a concept already, however its sheer presence in the dwc list adds to its importance in this particular domain.

2.4.7 Effective Dates

It is best to consult domain experts for feature identification, as they can provide insights on what is truly important in the field that are easily overlooked by model developers. Domain expert Labiosa¹¹ points out that effective dates are important in drinking water monitoring and control. Regulating agencies roll out new effective dates for provisions as they are updated continuously over the year. Effective dates could potentially reflect on a hidden triggering event for provision revisions, for instance, a newly passed bill or statute might require updates on relevant provisions. *Ideally*, related provisions are updated concurrently and should share similar effective dates. Of course, this is assuming that every group meets their deadline of publication, which might not be always true.

In our system, handcrafted rules are used to locate effective dates. Similar to measurements, dates are sometimes modified with quantifiers, and here is a non-exhaustive list of quantifiers obtained from drinking water regulations: after, effective on, beginning, starting, subsequent to, as of, since, in effect on, before, prior to, until, by,

¹¹ Mr. Bill Labiosa worked for the EPA for several years on national drinking water standards, and is currently a doctoral student in Environmental Engineering at Stanford University.

adopted, expire on, no later than, and through. As a result, we have the following four types of date elements:

- `<date date="January 24, 1978" num="1" />`
- `<date to="May 18, 1994" num="2" />`
- `<date from="October 13, 1978" num="2" />`
- `<date from="January 1, 1993" to="December 31, 2001" num="1" />`

The first one is a simple date entity without any quantifier. The second is an upper bound date entity (e.g., prior to May 1, 2003) while the third is a lower bound date entity (e.g., no later than May 1, 2003). The last one represents a range of dates, such as “from May 1, 2003 to June 1, 2003”. All date attributes include a frequency count similar to other features.

2.5 Results

The repository is complete with regulations properly transformed into XML format. Regulation hierarchies are reconstructed by structuring XML elements accordingly, while both generic and domain-specific features are extracted using handcrafted rules and a phrase extraction tool. The resulting documents are shown in the following sections. First, Section 2.5.1 displays provisions from accessibility and environmental regulations in XML format tagged with the complete set of structural and feature markups. Apart from displaying the documents in plain XML format, we can also view them as trees with nodes representing provisions, and this is shown in Section 2.5.2. Finally, concept ontologies can be developed to aid retrieval of provisions, and an example is shown in Section 2.5.3.

2.5.1 Examples with Complete Set of XML Markups

Presented below are three examples with the complete set of feature markups. The first example, shown in Figure 2.9, comes from the ADAAG definition section, and it shows the extracted definition, concept and index features. Since provisions tend to be lengthy, only excerpts are shown with ellipsis marks to represent omitted features and content texts. For instance, in Figure 2.9, there are indeed more concepts extracted from Section 3.5 of the ADAAG, aside from the single concept “access aisle” shown here.

```
Original section 3.5 from the ADAAG
3.5 DEFINITIONS.
...
ACCESSIBLE.
Describes a site, building, facility, or portion thereof that
complies with these guidelines.
...
CLEAR.
Unobstructed.
...

Refined section 3.5 in XML format
<regElement name="adaag.3.5" title="definitions">
  <concept name="access aisl" num="2" />
  <index name="facil" num="25" />
  <definition>
    <term> accessible </term>
    <definedAs> Describes a site, building, facility, or
      portion thereof that complies with these guidelines.
    </definedAs>
  </definition>
  <definition>
    <term> clear </term>
    <definedAs> Unobstructed. </definedAs>
  </definition>
  ...
</regElement>
```

Figure 2.9: Concept, Definition and Index Tags

As shown in Figure 2.10, the second example is a typical provision from the UFAS, which contains exception and measurement tags. The last example in Figure 2.11 is a drinking water provision from the 40 CFR, which illustrates two features specific in environmental regulation: `dwc` (drinking water contaminant) and `date`. A different type of measurement is shown as well, where 0.05 milligrams per liter is translated to 0.05 ppm (parts per million) for consistency with other measurements.

```
Original section 4.6.3 from the UFAS  
4.6.3 PARKING SPACES.  
Parking spaces for disabled people shall be at least 96 in  
(2440 mm) wide and shall have an adjacent access aisle 60 in  
(1525 mm) wide minimum (see Fig. 9). Parking access aisles ...  
EXCEPTION: If accessible parking spaces for vans designed for  
handicapped persons are provided, each should have an ...  
  
Refined section 4.6.3 in XML format  
<regElement name="ufas.4.6.3" title="parking spaces">  
  <concept name="access aisl" num="3" />  
  <measurement unit="inch" size="96" quantifier="min"  
    num="1" />  
  <reference name="ufas.4.5" num="1" />  
  ...  
  <regText> Parking spaces for disabled people shall ...  
</regText>  
  <exception> If accessible parking spaces for ... </exception>  
</regElement>
```

Figure 2.10: Measurement and Exception Tags

```

Original section 141.11.b from the 40 CFR
§ 141.11 Maximum contaminant levels for inorganic chemicals.
(a) The maximum contaminant level for arsenic applies only to
community water systems ...
(b) The maximum contaminant level for arsenic is 0.05
milligrams per liter for community water systems until January
23, 2006.

Refined section 141.11.b in XML format
<regElement id="40.cfr.141.11.b" name="">
  <dwc name="arsen" times="1" />
  <concept name="commun water system" times="1" />
  <measurement unit="ppm" size="0.05" quantifier="max" />
  <date to="January 23, 2006" num="1" />
  ...
  <regText>
    The maximum contaminant level for arsenic is 0.05
    milligrams per liter for community water systems until
    January 23, 2006.
  </regText>
</regElement>

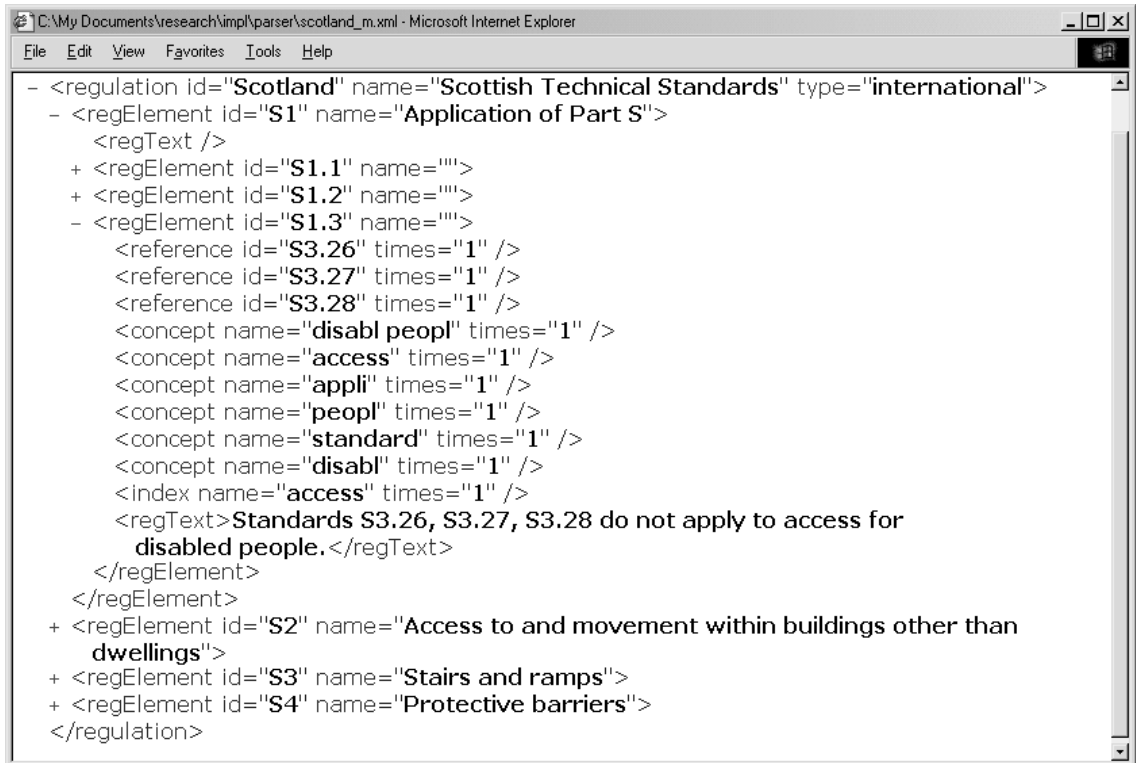
```

Figure 2.11: Drinking Water Contaminant and Effective Date Tags

2.5.2 Natural Tree View of Regulations

One of the many flexibilities of XML is its separation of data from style. Stylesheets can be written in order to render XML documents in the desired manner; without specifying a style, most browsers simply display XML documents as an expansible tree similar to a folder explorer. For instance, Figure 2.12 shows the default display of an XML regulation rendered in Internet Explorer without any stylesheet. Clearly, this exploration style is not very user friendly, which explains why a stylesheet is needed. As shown in Figure 2.13, a simple stylesheet written in eXtensible Stylesheet Language, XSL [42], is used to render an XML regulation in a different tree format. Each node represents a provision; clicking on any node results in a popup window displaying the provision content as shown in Figure 2.14. This tree view is obtained by modifying a publicly

available stylesheet called Tree Chart [100] that utilizes XSL Transformations (XSLT) to convert an XML document into an HTML chart that resembles a tree.



```

- <regulation id="Scotland" name="Scottish Technical Standards" type="international">
- <regElement id="S1" name="Application of Part S">
  <regText />
+ <regElement id="S1.1" name="">
+ <regElement id="S1.2" name="">
- <regElement id="S1.3" name="">
  <reference id="S3.26" times="1" />
  <reference id="S3.27" times="1" />
  <reference id="S3.28" times="1" />
  <concept name="disabl peopl" times="1" />
  <concept name="access" times="1" />
  <concept name="appli" times="1" />
  <concept name="peopl" times="1" />
  <concept name="standard" times="1" />
  <concept name="disabl" times="1" />
  <index name="access" times="1" />
  <regText>Standards S3.26, S3.27, S3.28 do not apply to access for
    disabled people.</regText>
  </regElement>
</regElement>
+ <regElement id="S2" name="Access to and movement within buildings other than
  dwellings">
+ <regElement id="S3" name="Stairs and ramps">
+ <regElement id="S4" name="Protective barriers">
</regulation>

```

Figure 2.12: XML Regulation Rendered in Internet Explorer without a Stylesheet

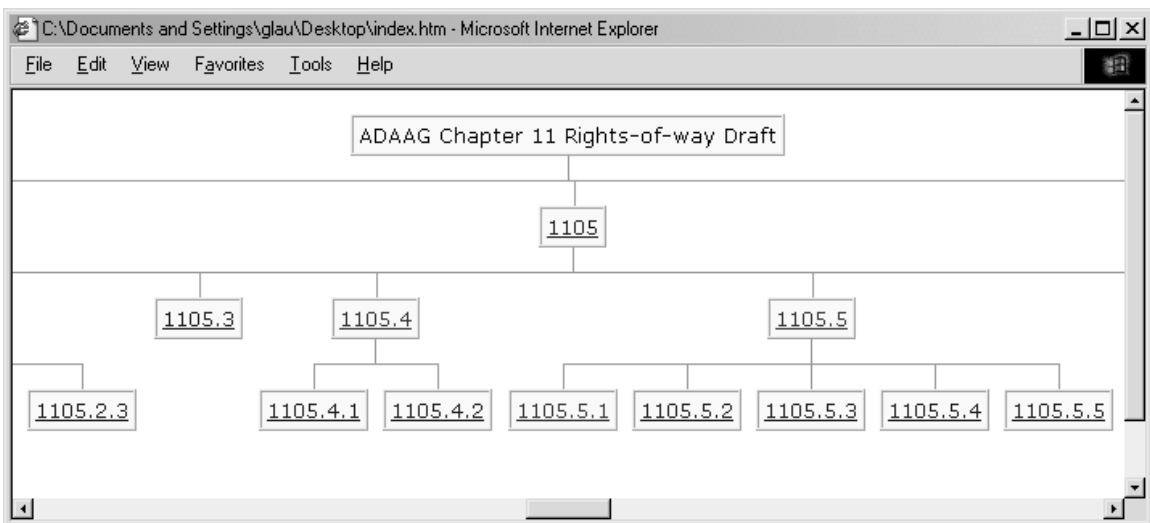


Figure 2.13: Tree View of XML Regulation Rendered with an XSL Stylesheet

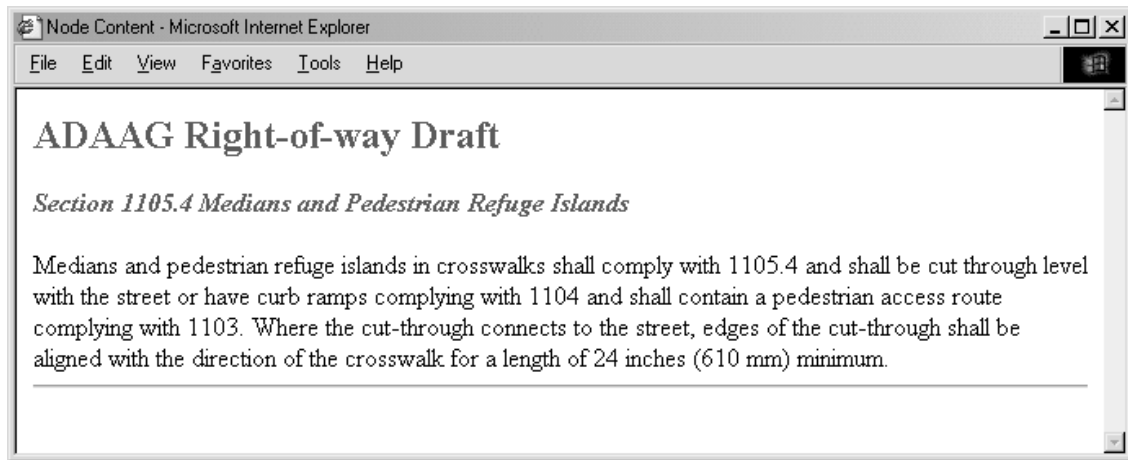


Figure 2.14: Content of a Provision Obtained by Clicking on a Node in Figure 2.13

If a natural folder explorer view or a stylesheet display format is not enough, standalone software applications can be used to render XML documents in specific styles. SpaceTree [56], a dynamically rescaling tree browser developed at the Human-Computer Interaction Lab at University of Maryland, is an example of such applications. As shown in Figure 2.15, the same XML regulation “Rights-of-Way Draft” can be displayed in a fancier tree format compared to a pure stylesheet rendering such as that shown in Figure 2.13. Just as the folder explorer style, tree nodes can be expanded and collapsed by clicking on nodes, while branches are dynamically rescaled to fit the screen space. We modify SpaceTree to display node contents by double clicking, so that details of a provision can be viewed separately in a popup window.

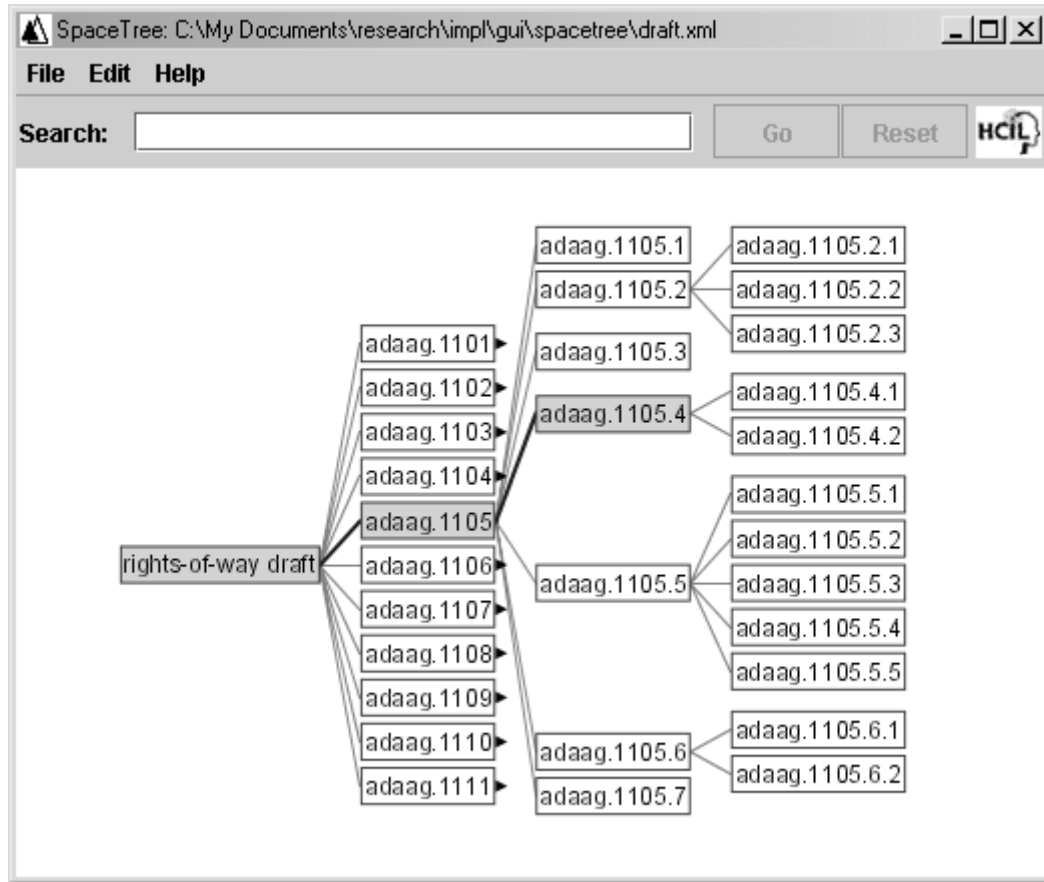


Figure 2.15: SpaceTree Display of an XML Regulation

2.5.3 Concept Ontology

Besides reading regulations based on its natural hierarchy, users might find it helpful to browse through an ontology [60] with documents categorized based on concepts as well. Semio Tagger [95] is one of several software products that provide such a capability. It identifies a list of concepts that are central to the corpus, as described in Section 2.4.1. It also provides a concept latching tool to help knowledge engineers to categorize the concepts and create an ontology. This is a semi-automated process where knowledge

engineers review and edit the list of concepts extracted by the Tagger, optionally add their own concepts, and arrange the concepts in a logical ontology.

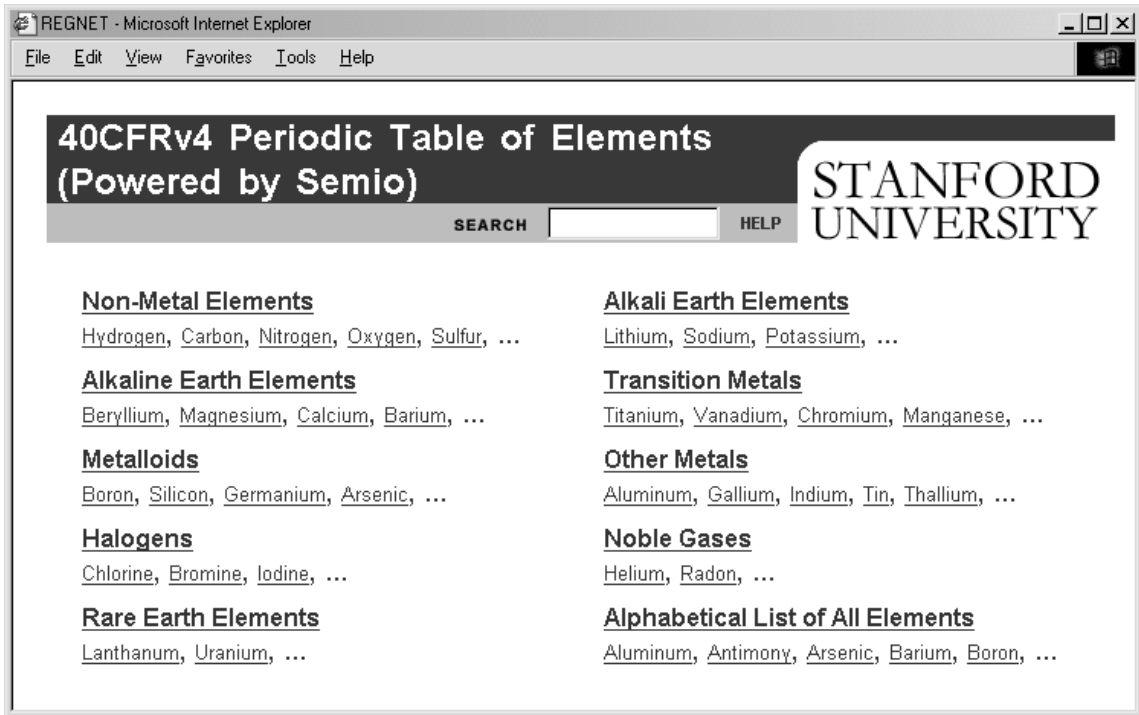


Figure 2.16: An Ontology Based on Environmental Regulations

As a result, provisions are clustered according to the ontology, and users can click through the structure to view relevant sections classified according to concepts. Figure 2.16 shows an ontology generated based on the periodic table of elements mapped onto environmental regulations.¹² Users can also perform a search on the list of concepts.

¹² The ontology shown is a joint effort between Bill Labiosa and Charles Heenan; both are graduate students at Stanford University.

2.6 Summary

The existence of multiple sources of regulations leads to a variety of formats and structures. In order to consolidate different regulations, we present the development of a regulatory repository as the platform for document retrieval and comparative analysis. This chapter first examines related work on the area of feature extraction and document structural extraction. The definition of feature extraction is quoted, and examples of automated feature extraction tools, in particular noun phrase extraction software, are given. Different techniques of feature extraction as well as the reason why feature extraction is desired are explained. Aside from feature identification, structure extraction is also examined; specifically, representation formats for semi-structured documents are introduced. Since our system focuses on provision comparisons instead of query matching, various work on keyword search is referenced as well.

A list of the interested regulations is given, where the focus is on regulations and codes of practice from the domains of disabled access and drinking water control. A brief survey of current digital publication of regulations reveals the need for a representation format that can gracefully encapsulate semi-structured data such as regulations. XML is selected as the system standard because of its capability of structure and feature inclusions. A semi-automated process is performed to transform different formats of regulations into XML format.

A shallow parser is developed to minimize human effort in data cleaning and consolidation, where each provision is encapsulated in a corresponding XML element. The hierarchy of provisions is reconstructed by properly structuring these XML elements. Finally, features, or evidences, are extracted from the corpus semi-automatically; this includes generic features, such as concepts, and domain-specific features, such as measurements. A text mining tool is used along with handcrafted rules to extract and tag features as XML elements in provisions.

Examples of provisions in XML format with the complete set of structural and feature markups are shown. Because of its intrinsic hierarchical nature, XML regulations can also be viewed as trees, with tree nodes representing regulation provisions. Therefore, we present several trees obtained using different rendering techniques. An ontology, based on elements from the periodic table, is developed on top of the identified concepts to allow for easy retrieval of provisions following the classification. As a result, users can browse through regulatory documents according to its natural hierarchy or based on concept clusters.

Chapter 3

Relatedness Analysis

With the regulatory repository, users can browse and search through the regulations easily. However, upon finding a relevant provision for a particular design scenario, it is still difficult to locate further desired materials with the volumes of regulations available. It becomes a more difficult task to search through multiple codes with multiple terms to locate more related provisions, if there is any. Nonetheless, there is a need to locate as much relevant information as possible, since as noted by Berman and Hafner [12], “[a] vast amount of information ... must be collected and integrated in order for the legal system to function properly.” In addition, they have pointed out that the chance of missing relevant information increases as the repository size grows unavoidably over the years [12]:

“The process of *finding the law* (including statutes, prior court cases, administrative rulings and procedural requirements) may involve searching a database of millions of potentially relevant documents. The proliferation of legal documents is a major cause of the growing cost of legal services noted by Harvard President (and former law professor) Derek Bok. Furthermore, as the chance of missing a relevant document increases, the legal status of a case becomes increasingly difficult to determine.”

Clearly, there is a need for an analysis tool to provide a reliable measure of relatedness of pairs of provisions, and to recommend similar sections of a selected provision based on a similarity measure. This chapter discusses the theory and implementation of a proposed relatedness analysis framework for regulations, where system evaluations and results obtained from the analysis follow in Chapter 4. The goal is to identify the most 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. This is obtained by first comparing regulations based on conceptual information as well as domain knowledge through a combination of feature matching. In addition, regulations possess specific structures, such as the tree hierarchy and the referential structure as shown in Figure 2.4. These structures also represent useful information in locating related provisions, and are therefore incorporated into our analysis for a more accurate comparison.

This chapter is organized as follows. Section 3.1 reviews the literature of relatedness analysis, which consists of three parts: definitions of different related areas of study, document comparisons and hyperlink topology. Section 3.2 starts with the discussion on the meaning of similarity and relatedness. Section 3.2.1 defines the basis of comparisons, such as the operators and units used in the computation. The definition of the similarity score and its matrix representation follows in Section 3.2.2. The analysis starts with an initial similarity score computation introduced in Section 3.3. The base score represents a linear combination of feature matching, which are discussed in Sections 3.3.1 to 3.3.2. Specifically, we introduce a traditional Boolean matching model in Section 3.3.1, while a new non-Boolean matching model via a vector space transformation is proposed in Section 3.3.2. Score refinements based on the structure of regulations are presented in Section 3.4. Section 3.4.1 addresses the natural hierarchical structure of regulations through a process termed neighbor inclusion. Section 3.4.2 introduces reference distribution, which incorporates the referential structure of regulations into the analysis. 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. A stable ranking of the most related sections is produced as a result, and similar sections from different regulations can be retrieved and recommended to users based on the comparison. Section 3.5 gives a summary on the analysis.

3.1 Related Work

This chapter 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 3.1.1 defines the fields of information retrieval, information extraction and text mining. Section 3.1.2 examines different techniques for textual comparisons, such as the Vector Model and Latent Semantic Indexing. Academic citation analysis and different researches based on hyperlink structure of the Web are reviewed in Section 3.1.3.

3.1.1 Information Extraction, Retrieval and Mining

Data mining [44] emerges from the fields of machine learning, statistics, artificial intelligence, pattern recognition and psychology. It defines the process in which patterns in data are discovered by generating hypotheses and predictions. Some individuals distinguishes data mining from Knowledge Discovery in Databases (KDD), where KDD refers to the overall process of useful knowledge discovery which includes data cleaning, preparation, and mining. Others distinguish the role of statistics from data mining, where statistical tools are used to validate hypothesis [52] generated from data mining. Most of the techniques involved in data mining, such as neural network, regression analysis and decision trees, are existing techniques. Data mining gains new momentum from the growing amount of data available coupled with the increasing power of computer

processors, which leads to the accrual demand of business intelligence building upon the data [18].

Text mining [58] is the application of data mining techniques, such as clustering and nearest neighbor analysis, on a large textual database. A close relative of text mining is Information Extraction (IE), which defines the fact-finding process from documents [55]. Just like data mining and KDD, a vague distinction exists between text mining and IE: text mining generates new findings that are unknown of prior to mining, while IE extracts existing facts from text.

A confusingly similar field termed Information Retrieval (IR) deals with “the representation, storage, organization of, and access to information items [5].” In general, IR is perceived as more associated with information indexing and ranking of relevance [31], while IE is concerned with factual extraction. It seems unclear as to how the distinction is drawn. Thus, we shall use the terms Information Extraction, Information Retrieval and text mining interchangeably in this context to represent the broader aspect of knowledge discovery among information available.

3.1.2 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 model¹³ [91, 93]. Most of these techniques are bag-of-word type of analysis, which means that they are

¹³ The Vector model is also called the Vector space model.

word order insensitive [5]. As our technique is based on the Vector model, we will briefly go over the basic mathematical formulation here.

In the Vector space 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 $\vec{d}_M = (w_{1,M}, w_{2,M}, \dots, w_{n,M})$, 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 [5]:

$$f_v = \frac{\vec{d}_M \cdot \vec{d}_N}{|\vec{d}_M| \times |\vec{d}_N|} \quad (3.1)$$

$$= \frac{\sum_{i=1}^n w_{i,M} \times w_{i,N}}{\sqrt{\sum_{i=1}^n w_{i,M}^2} \times \sqrt{\sum_{i=1}^n w_{i,N}^2}}$$

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 [92]. A simple approach is to use the count of term appearance as the term weight. One of the more popular algorithms is the $tf \times idf$ approach [38, 92], 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 is equal to the frequency count of term appearance in documents. The formula

to compute the *idf* component in a *tf×idf* model is $\log(k/k_i)$, where k is the total number of documents, and k_i is the number of documents in which the particular index term i appears. This means that words appearing in all documents in the corpus will have an *idf* factor of 0. The *log* formula implements the intuition that a frequently-used term is not useful in distinguishing similarities between documents. For example, a stopword, defined as a word that occurs frequently in the text of a document such as articles and prepositions, will most likely result in a zero *idf* score. Essentially, *tf* represents the intra-cluster similarity, while *idf* accounts for the inter-cluster dissimilarity. Based on a *tf×idf* model, the index term weight $w_{i,M}$ is equal to the frequency of term i in document M multiplied by $\log(k/k_i)$.

Without the help of thesauri, this type of models cannot capture synonyms which can potentially convey important information. We introduce the Latent Semantic Indexing (LSI) model [34], which aims to fill the gap between terms and concepts. LSI uses an algorithm called Singular Value Decomposition (SVD) [54] 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. In this thesis, LSI will be used as the benchmark to compare with our experimented results. Its mathematical formulation is briefly introduced here.

A term-document matrix K is populated with the weights of the index terms in the documents, with rows representing terms and columns representing documents. As suggested above, queries can be formulated as pseudo-documents. The matrix K^TK is the document similarity matrix by assuming the cosine between normalized document vectors as the similarity measure. Latent Semantic Indexing proposes to decompose the K matrix using Singular Value Decomposition as follows:

$$K = PQR^T \tag{3.2}$$

where the matrices P and R^T represent the eigenvalues derived from KK^T and K^TK , and the diagonal matrix Q stores the singular values. For some $s \ll \text{rank}(Q)$, we take only the largest s singular values from Q and zero out the rest to form Q_s . The number of singular values s should be large enough to include all of the important concepts, but small enough to reduce noise. Equation (3.2) reduces to

$$K_s = P_s Q_s R_s^T \quad (3.3)$$

$$\begin{aligned} K_s^T K_s &= (P_s Q_s R_s^T)^T (P_s Q_s R_s^T) \\ &= R_s Q_s^T P_s^T P_s Q_s R_s^T \\ &= R_s Q_s^2 R_s^T \quad \because P_s^T P_s = I, Q_s^T = Q_s \end{aligned} \quad (3.4)$$

with P_s and R_s being the corresponding reduced matrices. Equation (3.3) shows the new term-document matrix K_s computed in a reduced space, where Equation (3.4) represents the corresponding formulation of the similarity between documents. The (i, j) element in $K_s^T K_s$ denotes the similarity score between documents i and j , computed in a new concept space with reduced dimensions. If a query is modeled as a pseudo-document i , the (i, j) element represents the similarity score between the query and document j .

There are some investigations into improving the LSI, such as the Probabilistic Latent Semantic Analysis (PLSA) [59]. Experiments based on the PLSA on small sets of documents are performed [19]. 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 [33]. As an alternative, work has been done in the area of linguistic analysis and ambiguity resolutions [33, 40] to detect redundant documents, on a very focused document set.

3.1.3 Hyperlink Topology

Due to the evolution of the World Wide Web, there has been a lot of research work related to academic citation analysis [48]. For instance, CiteSeer is a scientific literature

digital library that provides academic publications indexed with their citations [17]. Different types of hyperlink topology and fitting models are examined extensively for different purposes [24, 57, 96]. One of the examples is Google's PageRank algorithm [20, 76]. This model ranks the importance of web pages by simulating the navigation pattern of Web users. It assumes that users follow hyperlinks from a starting page with an assigned probability p and jump to a random page with probability $(1 - p)$. The weight of cited pages are normalized according to the number of links the start page contains. In essence, 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 [68]. 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 [51]. In our work, the heavily referenced nature of regulations provides extra information about provisions just like 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.

3.2 Relatedness Analysis Measure

Before we discuss the theory and implementation of our proposed relatedness analysis, we will first lay out some groundwork and definitions in this section. Section 3.2.1 introduces the basis of analysis, such as the chosen unit of comparison and definitions of

operators used later in the chapter. Section 3.2.2 defines the similarity score along with its matrix representation, and briefly explains the process of obtaining and refining the score with a schematic. Here, we will begin with a discussion on the subject of *similarity* and *relatedness*.

The term “similar” is defined in the Merriam-Webster Dictionary¹⁴ to be “having characteristics in common; strictly comparable; alike in substance or essentials; not differing in shape but only in size or position,” whereas the term “related” is defined to be “connected by reason of an established or discoverable relation.” One could argue that the terms *similarity* and *relatedness* represent fundamentally different concepts, since objects that are “connected by reason” are not necessarily “alike in substance.” Whether *similarity* or *relatedness* is a better description depends on the situation; in our domain of legal informatics, materials that are strictly comparable or alike in essentials deserve attentions naturally. However, regulations that are connected by reason of a discoverable relation are probably more interesting. For instance, the controversial example of conflicting provisions shown in Figure 2.7 is indeed, according to dictionary definitions, *related* but not *similar*. A flush slope of curb ramp is not “strictly comparable” to a curb ramp with a beveled lip, but both elements are certainly “connected by reason.” It is intrinsic that *relatedness* includes *similarity* according to this interpretation, while the reverse condition does not necessarily hold.

The relationship between *similarity* and *analogy* has been studied by psychologist Gentner and Markman [49], based on the concept of *analogy* defined by Johannes Kepler [65]. They suggested that “similarity is like analogy”, and a structure mapping algorithm for similarity alignment has been developed [43]. To further complicate the situation, we introduce the notion of deducing similarity or relatedness between two entities. How similarity or relatedness should be determined is never a precise science, and in most

¹⁴ The Merriam-Webster Dictionary is a product of Merriam-Webster, Inc.

cases, this could be subjective. Even in a confined and rule-driven domain such as the law, it is still unclear on how this judging of similarity is performed as quoted in [89]:

“In Anglo-American law *stare decisis* – the doctrine of precedent – governs much legal reasoning. *Stare decisis* requires that similar cases be decided similarly. While this doctrine puts the focus squarely on reasoning from case to case, it is silent on how “similarity” should be determined. In fact, similarity is not static; it can depend on one’s viewpoint and desired outcome.”

As such, it is difficult to precisely define what is truly meant by a relatedness or similarity analysis in a legal corpus, let alone the decision on whether similarity or relatedness better represents our interest. As shown in the quotation above, similarity is not static but depends on one’s desired outcome. It suffices to say that, in the domain of legal informatics, a comparative analysis among regulations and supplementary documents should desirably 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.

3.2.1 Basis of Comparison

Due to the recent proliferation of the Internet, an extensive amount of research focusing on retrieval of relevant documents based on keyword search has been performed. Well-established techniques such as query expansion have been deployed to increase retrieval accuracy as discussed in Section 2.1. As a result, 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 a concept ontology such as shown in Figure 2.8. 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 regulations rather than matching queries at the front end.

Apart from defining our goal of comparison, the unit of comparison needs to be specified as well. Here, since a typical regulation can easily exceed thousands of pages, a comparison between a full set of regulation and another is meaningless. 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 [1] and Section 1127B.4.4 in CBC [25] as in the example shown in Figure 2.7. As suggested in Section 2.3.1, we use the terms “section” and “provision” interchangeably to represent the unit of extraction as well as the unit of comparison. The actual and official terminology differs from regulation to regulation.

To help define the terminologies for the basis in our comparison, we show below an illustration of two partial regulation trees. As shown in Figure 3.1, 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 3.1. 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 between an entity and itself, can also be performed on a single tree using the same analysis.

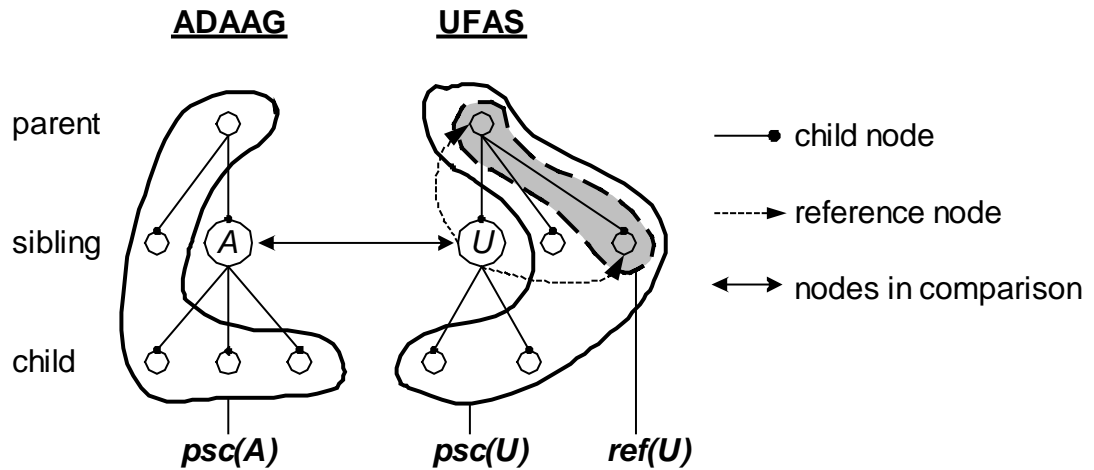


Figure 3.1: Immediate Neighboring Nodes and Referenced Nodes in Regulation Trees

3.2.2 Similarity Score

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. For instance, a self-comparison of a provision should result in a similarity score of 1, while a partial match between two provisions should result in a score that is greater than 0 and less than 1. The similarity score is denoted by $f(A, U) \in [0, 1]$ 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 $f(A, U) = f(U, A)$.

As we will be considering batches of similarity scores later in score refinements, matrix representations are employed to simplify the notations. A similarity score matrix Φ is defined to represent similarity scores between regulation A and regulation U . If regulations A and U consist of p and q number of sections respectively, the dimension of Φ is p by q where rows and columns denote sections from regulations A and U correspondingly. $\Phi(i, j)$ is defined to be the similarity score between Section i from regulation A and Section j from regulation U , i.e., $f(i, j)$. Subscripts are used to differentiate scores obtained from different analyses, such as $f_{\phi}(\bullet, \bullet)$ represents a single base score and Φ_{rd} represents a matrix of scores from reference distribution.

A schematic is shown below in Figure 3.2 for the similarity analysis core. It takes as an input the parsed regulations tagged with the associated features as well as any user-provided domain knowledge, and 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 well-prepared repository such as one described in Chapter 3, 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.

Referring to Figure 3.2, the analysis process starts with a base score computation which is a linear combination of scores from different features identified in the corpus. User-provided domain knowledge, such as ontologies or specific matching algorithms, is incorporated in feature matching. The base score is refined by incorporating the structure of regulations in the analysis, which includes neighbor inclusion and reference distribution. First, the immediate neighboring nodes, in particular the parent, the siblings and the children, are compared to modify the base score. The influence of the not-so-immediate neighbors is taken into account by reference distribution. Based on the assumption that similar sections reference similar sections, referenced nodes are also compared to refine the score. As a result, the final similarity scores between provisions

are obtained, and a stable ranking of the most related provisions is produced. Sections 3.3 and 3.4 below describe the reasoning, theory and implementation of our relatedness analysis.

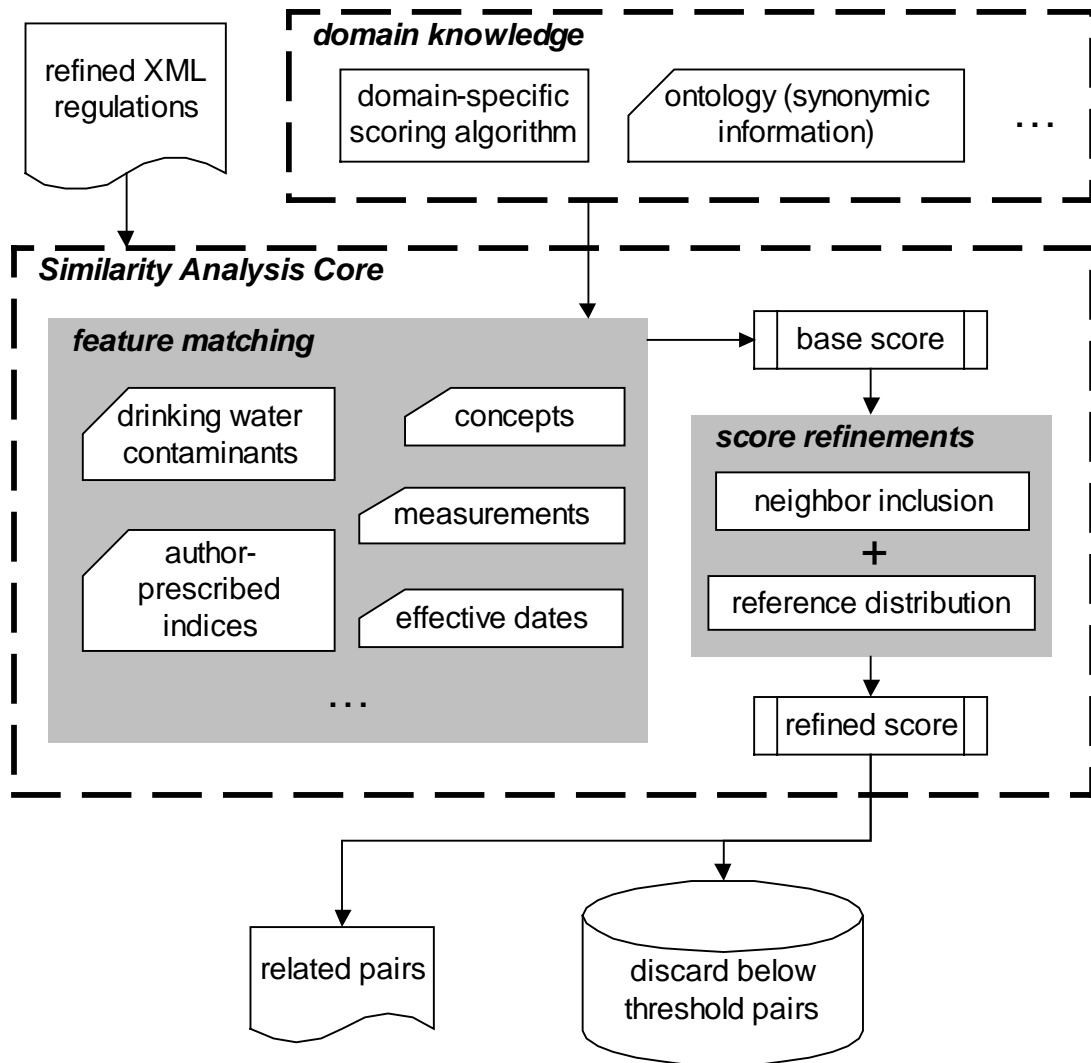


Figure 3.2: Similarity Analysis Core Schematic

3.3 Base Score Computation

The base score f_0 is a linear combination of the scores from each of the features i as shown in Equation (3.5) below. $F(A,U,i) \in [0, 1]$ represents the similarity score based on the comparison of feature i between Sections A and U , whereas β_i is the weight of feature i . Different features can be weighted differently, and the weights should sum up to 1 to assure a similarity score between 0 and 1. For instance, one might trust domain knowledge more than machine-generated phrases such as concepts, and thus assigns heavier weight on domain knowledge. In Chapter 4, we will discuss different combinations of weights and the effects on the corresponding results. A concise representation for the base scores between two regulations is the base score matrix Φ_0 , with its (i, j) element given by $\Phi_0(i, j) = f_0(i, j)$.

$$f_0(A,U) = \sum_{i=1}^N \beta_i \times F(A,U,i) \quad (3.5)$$

$$\sum_{i=1}^N \beta_i = 1$$

$F(A,U,i)$ = similarity score between Sections (A,U) based on feature i

N = total number of features

The scoring scheme for each of the features is discussed in the following sections; they essentially reflect how much resemblance can be inferred between the pair of sections based on that particular feature. We use the Vector model [91] as the basis of different feature comparisons, and its limitation is observed. Because of the limitation of the Vector model, namely the assumption of a Boolean matching between axes, we propose a new model to allow for a non-Boolean match which will be introduced in Section 3.4.

3.3.1 Boolean Feature Matching

In our relatedness analysis model, features are compared in the same way as the index terms in the Vector model [91]. The Vector Model is introduced in Section 3.1.2, where the degree of similarity of documents is evaluated as the correlation between their index term vectors. A document M is represented as a n -entry vector $\vec{d}_M = (w_{1,M}, w_{2,M}, \dots, w_{n,M})$, with n being the total number of index terms in the corpus. Equation (3.1) gives the similarity computation between two documents based on a cosine vector distance measure. Different term weight schemes [92], such as a *tf×idf* approach, are introduced in Section 3.1.2 as well.

In this section, we will introduce the Boolean feature matching model, which is fundamentally the Vector model with different *index term vectors* for different features. Two examples of features are given. Section 3.3.1.1 investigates the vector representation for concepts and author-prescribed indices that share the same approach. Section 3.3.1.2 applies the Vector model to drinking water contaminants, with a slightly more complicated representation due to the existence of synonyms introduced by a user-defined ontology.

3.3.1.1 Comparisons of Concepts and Author-Prescribed Indices

To compare provisions based on the extracted concepts, we employ techniques similar to the Vector model. As detailed in Section 2.4.1, the regulations are indexed with concepts and frequency count is kept in a `<concept>` XML element. With this information tagged in each provision, an n -entry provision-concept vector \vec{c}_M , where n represents the total number of unique concepts in the regulatory corpus, can be easily constructed per provision M . Compared to the Vector model, we have provisions instead of documents since our unit of comparison is provisions in regulations. With respect to Equation (3.1), the provision-concept vector \vec{c}_M replaces the document vector \vec{d}_M . We take the frequency count of concept i as the concept weight $w_{i,M}$ in $\vec{c}_M = (w_{1,M}, w_{2,M}, \dots, w_{n,M})$.

This is because concepts represent an already selected set of important noun phrases which we assume are not common terms such as stopwords, and therefore no *idf* factor is included in our model. The similarity score $F(A,U,i=concept)$, based on a concept comparison between Sections A and U , is obtained from the cosine similarity between the two provision-concept vectors \vec{c}_A and \vec{c}_U :

$$F(A,U,i=concept) = \frac{\vec{c}_A \bullet \vec{c}_U}{|\vec{c}_A| \times |\vec{c}_U|} \quad (3.6)$$

An alternative source of concepts can be the author-prescribed indices from reference books or from the regulation itself. An identical comparison procedure as concepts is adopted, since comparisons between indices, which are term-based features without domain knowledge, are analogous to comparisons between concepts. The frequency count of index i in provision M represents the index weight $w_{i,M}$ in a provision-index vector \vec{i}_M . The similarity score $F(A,U,i=index)$, based on author-prescribed index comparison between Section A and Section U , is obtained as follows:

$$F(A,U,i=index) = \frac{\vec{i}_A \bullet \vec{i}_U}{|\vec{i}_A| \times |\vec{i}_U|} \quad (3.7)$$

Indeed, it is interesting to study the difference in results produced by human-written indices versus machine-generated concepts, as well as results obtained using a combination of them. We shall defer the discussion to Chapter 4, where different weighting schemes are experimented and the effect on results are observed.

3.3.1.2 Comparisons of Drinking Water Contaminants

In Section 2.4.6, the extraction of drinking water contaminants (represented as *dwc*) is discussed, with an ontology developed by domain experts¹⁵ using a combination of the

¹⁵ The ontology is developed by Bill Labiosa and further modified here.

EPA index of drinking water contaminants with supplementary materials. For instance, according to the ontology in Figure 2.8, synonymic information can be readily identified, such as the terms “total trihalomethane” and “tthm”. The provided synonyms and acronyms represent Boolean domain knowledge which could be modeled using the Vector model computation shown in Equation (3.1).

Conceptually, we need a manual space reduction to accommodate synonyms and acronyms. Synonymic and acronymic term axes, for example, acronyms “total trihalomethane” and “tthm,” can be collapsed onto one axis to represent the combined term, “total trihalomethane = tthm”. The definition of an n -entry provision-contaminant vector \vec{t}_M is slightly different from \vec{c}_M or \vec{i}_M , where the n axes here represent the *consolidated* contaminants counting synonymic or acronymic contaminants as one. Due to the introduction of an ontology, the vector space is manually collapsed from the use of *unique* features to using *consolidated* features as axes. Based on this definition of the provision-contaminant vector \vec{t}_M , we have

$$F(A,U,i=dwc) = \frac{\vec{t}_A \bullet \vec{t}_U}{|\vec{t}_A| \times |\vec{t}_U|} \quad (3.8)$$

3.3.2 Non-Boolean Feature Matching

From the comparisons of drinking water contaminants shown in Section 3.3.1.2, it appears that we are stretching to the limits of the Vector model using a manual space reduction. This is due to the introduction of domain knowledge, specifically, an ontology that defines synonyms in the case of drinking water contaminants. To incorporate potential non-Boolean domain knowledge, we need to modify the Vector model for feature matching. The development of a non-Boolean feature matching model is best illustrated using an example – Section 3.3.2.1 will introduce the modified Vector model using a measurement matching with non-Boolean domain knowledge, where Section

3.3.2.2 gives an example of a different non-Boolean feature matching with effective dates.

3.3.2.1 Comparisons of Measurements

Before we discuss the comparison technique for measurements, we first review what a measurement feature contains. An example XML measurement tag is defined with the following attributes as shown in Section 2.4.5:

```
<measurement unit="ft" size="2" quantifier="max" num="1" />
```

where “unit,” “size” and “num” are required attributes; “quantifier” is optional. In range measurements, “size” is replaced by “size1” and “size2” which are both required. One approach to comparing two measurements is to employ a technique equivalent to concept comparisons. Namely, an n -entry provision-measurement vector \vec{m}_M can be constructed per provision M , where n here represents the total number of *unique* measurements identified in the corpus. The provision-measurement vector \vec{m}_M would contain the frequency count of measurement i as the weight $w_{i,M}$. *Unique* measurements are defined to be measurement tags that differ from each other in any required or optional attributes except the “num” field. In this model, a measurement of “2 ppm” will be regarded as a non-match to “2 ppm max” and “2 ppm min” as they map to different *unique* axes.

The above approach illustrates a crucial limitation in the Vector model – each term is a Boolean match with other terms. For instance, a measurement of “2 ppm” can only be defined to be either a 0 or 100 percent match of “2 ppm max,” where a 0% match follows from the above *uniqueness* definition and a 100% match results from treating the two measurements as identical. The intrinsic Boolean match property can be attributed to the fact that the Vector model assumes independence between term axes, which, in most cases, is a simple and elegant approach. Indeed, Baeza-Yates and Ribeiro-Neto claimed that this independence assumption could be advantageous [5]. They suggested that a indiscriminate application of term dependencies to all documents in the collection might

in fact hurt the overall performance due to the locality of many term dependencies. Thus, consideration of term dependencies in practice could be a disadvantage.

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 such as user-defined ontologies, axis independence no longer holds. It is unrealistic to assume that the real world can be modeled as a black and white match, and as a result, domain knowledge is potentially non-Boolean. For instance, a domain expert might interpret “2 ppm” as 70% relevant to “2 ppm max”. In essence, the degree of match between two features is no longer 0 or 100%.

To allow for such flexibility in modeling domain knowledge and user-defined comparison algorithms, we could potentially ask users or domain experts to define their own vector matching algorithm to replace the Vector model. However, this approach puts the burden of mathematical modeling on domain experts, who are not necessarily comfortable with formulating vector comparison models. One can only realistically assume that a domain expert can reasonably answer questions such as “how similar are the measurements ‘2 ppm’ and ‘2 ppm max,’” but not questions in the form of “what is the degree of match between the measurement vectors [‘2 ppm’, ‘0 ntu’, ‘4 ppb’] and [‘3 ppm max’, ‘2 ntu’, ‘0 ppb’].” Our system lacks flexibility if only a *vector* matching algorithm, which answers the latter question alike, are accepted instead of a less sophisticated *feature* matching algorithm, which can be a blackbox that answers questions similar to the former. Therefore, we will base our analysis on the Vector model for consistency with feature comparisons that do not involve user-defined ontologies and domain knowledge, such as concept comparisons in Section 3.3.1.1. The Vector model is slightly revised to accommodate user-specified *feature* matching algorithms, instead of *vector* matching algorithms, per our discussion above.

A user-specified *feature* matching algorithm is defined to return a degree of match between two features, such as a 75% match between measurements “2 ppm” and “2 ppm max”. Identical to the definition of a similarity score, the returned degree of match should range between 0 and 1. A self-comparison or comparison of identical features should result in a 100% match, and the algorithm should be commutative as well. To devise a new vector comparison technique based on the Vector model, we will use the following user-defined matching algorithm example between two measurements m_1 and m_2 .

The pseudo-code shown in Figure 3.3 demonstrates a comparison algorithm that assigns 0, 50, 75 and 100 percent relevancy between two measurements. It first compares the “unit” and “size” attributes of two measurements; if they do not match, a zero score is returned. For instance, “2 ppm” and “2 ft” are completely independent according to this algorithm. The “quantifier” attributes are compared, and it returns a 100 percent match if the quantifiers are the same. A 75 percent match is assigned between a measurement without a quantifier and one with a quantifier. Finally, a measurement with a “max” quantifier is 50 percent related to one with a “min” quantifier. Figure 3.4 illustrates the example of this user-defined matching algorithm.


```

// unit match, e.g., "ppt" vs. "ntu"
BOOLEAN unit_match = Is_Identical(unit(m1), unit(m2));
IF (unit_match == false)
    RETURN 0;

// size match: single measurement or range measurement,
// e.g., "2" vs. "2-3"
BOOLEAN size_match;
IF (!range_measurement)
    size_match = Is_Identical(size(m1), size(m2));
ELSE
    size_match = Is_Identical(size1(m1), size1(m2)) &&
                Is_Identical(size2(m1), size2(m2));
IF (size_match == false)
    RETURN 0;

// quantifier match, e.g., "max" vs. null
IF (Is_Identical(quantifier(m1), quantifier(m2)))
    RETURN 1;
IF (quantifier(m1) == max && quantifier(m2) == min)
    RETURN 0.5;
IF (quantifier(m1) == min && quantifier(m2) == max)
    RETURN 0.5;
RETURN 0.75;

```

Figure 3.3: Pseudo-Code for a User-Defined Measurement Matching Algorithm

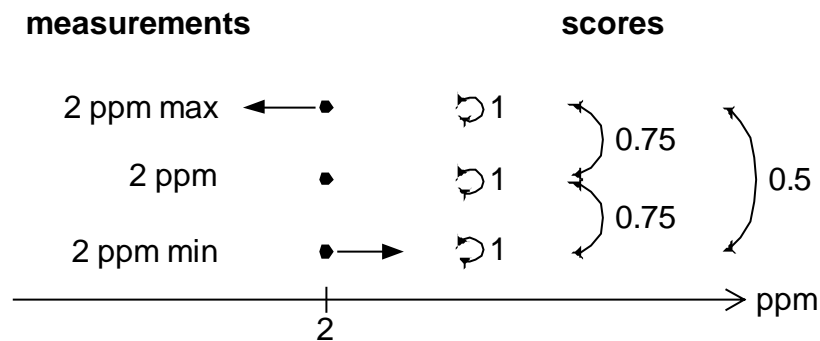


Figure 3.4: Illustration of an Example of a User-defined Measurement Comparison Algorithm

The information provided by a user-defined measurement comparison algorithm can be easily encoded into a matrix form. We let E be an $n \times n$ measurement matching matrix, where n is the number of *unique* measurements identified in the corpus. As suggested in the beginning of this Section, *uniqueness* is defined without prior domain knowledge such as a user-defined matching algorithm. Each *unique* measurement i corresponds to row i and column i of E . Each entry E_{ij} represents the degree of match between measurements i and j , such as $E_{ij} = 1$ between synonyms and $E_{ij} = 0.75$ between measurements “2 ppm” and “2 ppm max” to reflect a user-defined 75% match according to the pseudo-code in Figure 3.3. The diagonals E_{ii} are 1’s as self-comparisons are defined to result in a 100% match; in addition, E is symmetric (i.e., $E_{ij} = E_{ji}$) since user-defined matching algorithms are assumed to be commutative. Based on the algorithm shown as a pseudo-code in Figure 3.3, an example E matrix is given below, which represents a 3-dimensional vector space using “2 ppm,” “2 ppm max” and “2 ft” as the first, second and third measurement axes.

$$E = \begin{bmatrix} 1 & 0.75 & 0 \\ 0.75 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3.9)$$

To accommodate a degree-of-match algorithm such as the pseudo-code in Figure 3.3, we project the provision-measurement vector \vec{m} onto an alternate space before comparison. We will continue to use the earlier definition of the provision-measurement vector, namely each n -entry provision-measurement vector \vec{m}_M represents each provision M , where n here represents the total number of *unique* measurements identified in the corpus. 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 projects the provision-measurement vector \vec{m} onto an alternate space, and the resultant vector $\vec{m}' = D\vec{m}$ represents the *consolidated* measurement frequencies. After the transformation, we apply techniques similar to the Vector model to compare the consolidated frequency vectors \vec{m}'_A and \vec{m}'_U .

for Sections A and U respectively. The similarity score between Sections A and U based on a measurement comparison is given by the cosine between the consolidated frequency vectors \vec{m}_A' and \vec{m}_U' . The following shows the computation:

$$\begin{aligned}
 F(A,U,i=measurement) &= \frac{\vec{m}_A' \bullet \vec{m}_U'}{|\vec{m}_A'| \times |\vec{m}_U'|} & (3.10) \\
 &= \frac{D\vec{m}_A \bullet D\vec{m}_U}{|D\vec{m}_A| \times |D\vec{m}_U|} \\
 &= \frac{(D\vec{m}_A)^T D\vec{m}_U}{\sqrt{(D\vec{m}_A)^T D\vec{m}_A} \times \sqrt{(D\vec{m}_U)^T D\vec{m}_U}} \\
 &= \frac{\vec{m}_A^T D^T D\vec{m}_U}{\sqrt{\vec{m}_A^T D^T D\vec{m}_A} \times \sqrt{\vec{m}_U^T D^T D\vec{m}_U}} & (3.11)
 \end{aligned}$$

To determine the elements of D , we further investigate the meaning of cosine similarity between vectors. We will use the follow vectors as an illustrating example, where \vec{m}_A and \vec{m}_U represent the provision-measurement vector for provisions A and U respectively:

$$\vec{m}_A = \begin{bmatrix} w_{1,A} \\ w_{2,A} \\ w_{3,A} \end{bmatrix}, \quad \vec{m}_U = \begin{bmatrix} w_{1,U} \\ w_{2,U} \\ w_{3,U} \end{bmatrix} \quad (3.12)$$

$w_{i,j}$ denotes the frequency count of measurement i in provision j ; in this example, we have a total of three measurements identified in the corpus, which results in a 3-dimensional vector space. The cosine similarity between \vec{m}_A and \vec{m}_U can be computed as follows:

$$\begin{aligned}
 F(A,U,i=measurement) &= \frac{\vec{m}_A \bullet \vec{m}_U}{|\vec{m}_A| \times |\vec{m}_U|} \\
 &= \frac{w_{1,A} \times w_{1,U} + w_{2,A} \times w_{2,U} + w_{3,A} \times w_{3,U}}{\sqrt{w_{1,A}^2 + w_{2,A}^2 + w_{3,A}^2} \times \sqrt{w_{1,U}^2 + w_{2,U}^2 + w_{3,U}^2}} & (3.13)
 \end{aligned}$$

where the numerator is the degree of correlation between the two vectors based on the frequency count of the three measurements, and the denominator is the normalization factor. We can interpret the numerator as $(100\% \times w_{1,A} \times w_{1,U} + 100\% \times w_{2,A} \times w_{2,U} + 100\% \times w_{3,A} \times w_{3,U})$, since measurements 1, 2 and 3 are 100% match onto themselves. Extrapolating from this interpretation, the numerator can be expanded as $(100\% \times w_{1,A} \times w_{1,U} + 0\% \times w_{1,A} \times w_{2,U} + 0\% \times w_{1,A} \times w_{3,U} + 0\% \times w_{2,A} \times w_{1,U} + 100\% \times w_{2,A} \times w_{2,U} + 0\% \times w_{2,A} \times w_{3,U} + 0\% \times w_{3,A} \times w_{1,U} + 0\% \times w_{3,A} \times w_{2,U} + 100\% \times w_{3,A} \times w_{3,U})$. The 0% factor can be attributed to axis independence; for example, measurement 2 matches 0% of measurement 3 and therefore axes 2 and 3 are mutually independent.

We will now consider the change in cosine computation when synonyms are introduced into our system. For instance, we can take measurements 1 and 2 as synonymic measurements, which means measurement 1 matches 100% onto measurement 2. An example is “2 ppm” and “2 mg/l”¹⁶. This results in a reduced vector space as the two measurement axes collapse into one as shown in Equation (3.14). The first axis represents the synonymic measurements 1 and 2, while the second axis represents measurement 3. Their cosine changes correspondingly from Equation (3.13) to Equation (3.15).

$$\vec{m}_{A, \text{reduced}} = \begin{bmatrix} w_{1,A} + w_{2,A} \\ w_{3,A} \end{bmatrix}, \quad \vec{m}_{U, \text{reduced}} = \begin{bmatrix} w_{1,U} + w_{2,U} \\ w_{3,U} \end{bmatrix} \quad (3.14)$$

$$\begin{aligned} F(A, U, i=\text{measurement}) &= \frac{\vec{m}_{A, \text{reduced}} \bullet \vec{m}_{U, \text{reduced}}}{|\vec{m}_{A, \text{reduced}}| \times |\vec{m}_{U, \text{reduced}}|} \\ &= \frac{(w_{1,A} + w_{2,A}) \times (w_{1,U} + w_{2,U}) + w_{3,A} \times w_{3,U}}{\sqrt{(w_{1,A} + w_{2,A})^2 + w_{3,A}^2} \times \sqrt{(w_{1,U} + w_{2,U})^2 + w_{3,U}^2}} \quad (3.15) \end{aligned}$$

¹⁶ 1 part per million (ppm) converts to 1 milligram per liter (mg/l).

Again, the numerator can be expanded into $(100\% \times w_{1,A} \times w_{1,U} + 100\% \times w_{1,A} \times w_{2,U} + 0\% \times w_{1,A} \times w_{3,U} + 100\% \times w_{2,A} \times w_{1,U} + 100\% \times w_{2,A} \times w_{2,U} + 0\% \times w_{2,A} \times w_{3,U} + 0\% \times w_{3,A} \times w_{1,U} + 0\% \times w_{3,A} \times w_{2,U} + 100\% \times w_{3,A} \times w_{3,U})$. In the original 3-dimensional vector space, synonymic information would project a 100% match between seemingly independent axes, such as measurements 1 and 2, in a cosine computation $(100\% \times w_{1,A} \times w_{2,U})$. In a user-defined matching algorithm such as one shown in the pseudo-code in Figure 3.3, different measurements with a partial match can be incorporated into the analysis using a similar approach. For instance, if we have “2 ppm” and “2 ppm max” as measurements 1 and 2 respectively, the algorithm previously defined in Figure 3.3 would assign a 75% match between them, which can be modeled as $75\% \times w_{1,A} \times w_{2,U}$ and $75\% \times w_{2,A} \times w_{1,U}$ in a cosine equation. If we add onto this space an independent measurement axis 3, such as “2 ft,” the numerator of the cosine between Sections A and U becomes $(100\% \times w_{1,A} \times w_{1,U} + 75\% \times w_{1,A} \times w_{2,U} + 0\% \times w_{1,A} \times w_{3,U} + 75\% \times w_{2,A} \times w_{1,U} + 100\% \times w_{2,A} \times w_{2,U} + 0\% \times w_{2,A} \times w_{3,U} + 0\% \times w_{3,A} \times w_{1,U} + 0\% \times w_{3,A} \times w_{2,U} + 100\% \times w_{3,A} \times w_{3,U})$.

To obtain an augmented cosine as shown above, we propose to map the provision-measurement vectors onto an alternate space via a transformation matrix D . The cosine is then computed based on the transformed vectors. As is defined earlier, the resultant measurement vector $\bar{m}' = D\bar{m}$ represents the consolidated measurement frequencies. To determine the elements of D , we looked at the semantics of cosine in both synonym and partial matches. By comparing the numerator of Equation (3.11) with the semantics of cosine drawn above, we define D as $E = D^T D$, where E represents the measurement matching matrix based on a user-specified comparison algorithm. An example of E is illustrated earlier in Equation (3.9). For any symmetric $n \times n$ matrix E , there exists a $k \times n$ matrix D , where k is greater than or equal to 1 and less than or equal to n , such that the equality $E = D^T D$ holds. D can be computed using any matrix decomposition method. Theoretically, D represents the transformation matrix that maps the measurement vectors onto a different vector space to account for axis dependences prior to a cosine

comparison. As long as such a D matrix exists theoretically, the measurement vectors \vec{m}_A and \vec{m}_U are transformed onto *some* space via D where the cosine computation takes place. Consequently, the cosine between the transformed measurement vectors \vec{m}'_A and \vec{m}'_U as shown in Equation (3.10) can be written as Equation (3.11). In other words, the equality in Equation (3.11) only holds if such a D matrix exists. However, computationally speaking, we do not need to decompose E for the exact value of D , since by substituting $E = D^T D$ into Equation (3.11), the cosine becomes:

$$F(A,U,i=measurement) = \frac{\vec{m}_A^T E \vec{m}_U}{\sqrt{\vec{m}_A^T E \vec{m}_A} \times \sqrt{\vec{m}_U^T E \vec{m}_U}} \quad (3.16)$$

Clearly, the cosine can be computed directly using an accurately defined measurement matching matrix E and a decomposition into the transformation space D is not necessary. For example, using the E matrix defined in Equation (3.9) and the measurement vectors in Equation (3.12), the numerator of Equation (3.16) becomes $(w_{1,A} \times w_{1,U} + w_{2,A} \times w_{2,U} + w_{3,A} \times w_{3,U} + 0.75 \times w_{1,A} \times w_{2,U} + 0.75 \times w_{2,A} \times w_{1,U})$, which correctly models the 75% partial match between measurements 1 and 2. The denominator is $\sqrt{(w_{1,A}^2 + w_{2,A}^2 + w_{3,A}^2 + 2 \times 0.75(w_{1,A} \times w_{2,A}))} \times \sqrt{(w_{1,U}^2 + w_{2,U}^2 + w_{3,U}^2 + 2 \times 0.75(w_{1,U} \times w_{2,U}))}$ which represents the normalization factor for \vec{m}'_A and \vec{m}'_U .

By observation, Equation (3.16) reduces to the original Vector model if E is an identity matrix I . This is consistent with the feature comparison technique using a Boolean matching, such as a concept comparison where no dependency information is available. Indeed, a transformation matrix $D = I$ reflects the assumption of feature independence, where a feature either matches or does not match another feature; as a result, Equation (3.16) reduces to the Vector model. A more subtle observation lies in the case where synonyms are present. For any synonymic measurements, their corresponding axes can be collapsed onto one by summing the frequency counts of the synonyms. We will illustrate the dimension reduction using synonymic measurements i and j as an example. The dimension of the measurement vectors shown in Equation (3.17) can be reduced by

one as shown in Equation (3.18) by consolidating frequency counts of measurements i and j .

$$\vec{m}_A = \begin{bmatrix} \vdots \\ w_{i,A} \\ \vdots \\ w_{j-1,A} \\ w_{j,A} \\ w_{j+1,A} \\ \vdots \end{bmatrix}, \quad \vec{m}_U = \begin{bmatrix} \vdots \\ w_{i,U} \\ \vdots \\ w_{j-1,U} \\ w_{j,U} \\ w_{j+1,U} \\ \vdots \end{bmatrix} \quad (3.17)$$

$$\vec{m}_{A, \text{reduced}} = \begin{bmatrix} \vdots \\ w_{i,A} + w_{j,A} \\ \vdots \\ w_{j-1,A} \\ w_{j+1,A} \\ \vdots \end{bmatrix}, \quad \vec{m}_{U, \text{reduced}} = \begin{bmatrix} \vdots \\ w_{i,U} + w_{j,U} \\ \vdots \\ w_{j-1,U} \\ w_{j+1,U} \\ \vdots \end{bmatrix} \quad (3.18)$$

The measurement vectors are reduced from n -dimensions in Equation (3.17) to $(n-1)$ -dimensions in Equation (3.18), where n is the total number of unique measurements in the corpus without prior knowledge of synonyms. The measurement matching matrix E is also reduced from $n \times n$ to $(n-1) \times (n-1)$ by removing row j and column j as shown below, where $E_{ik} = E_{jk} \forall k$, and $E_{ik} = E_{\text{reduced},ik} \forall k \neq j$, with $1 \leq k \leq n$:

$$E = \begin{bmatrix} & i & & j-1 & j & j+1 & \\ 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & 1 & \cdots & \cdots & 1 & \cdots & \cdots \\ & & 1 & \cdots & \cdots & \cdots & \cdots \\ & & & 1 & \cdots & \cdots & \cdots \\ & & & & 1 & \cdots & \cdots \\ & & & & & 1 & \cdots \\ \text{sym.} & & & & & & 1 \end{bmatrix} \rightarrow E_{\text{reduced}} = \begin{bmatrix} & i & & j-1 & j+1 & \\ 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ & 1 & \cdots & \cdots & \cdots & \cdots \\ & & 1 & \cdots & \cdots & \cdots \\ & & & 1 & \cdots & \cdots \\ & & & & 1 & \cdots \\ & & & & & 1 \\ \text{sym.} & & & & & & 1 \end{bmatrix} \quad (3.19)$$

The cosine comparisons of the n -dimensional vectors and the $(n-1)$ -dimensional vectors should produce the same results as they represent the identical system. We will first show that the numerators of the cosine of the reduced and the original system are indeed the same:

$$\begin{aligned}
& \vec{m}_{A, \text{reduced}}^T E_{\text{reduced}} \vec{m}_{U, \text{reduced}} \\
&= \begin{bmatrix} \cdots & w_{i,A} + w_{j,A} & \cdots & w_{j-1,A} & w_{j+1,A} & \cdots \end{bmatrix} \begin{bmatrix} 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ & 1 & \cdots & E_{i,j-1} & E_{i,j+1} & \cdots \\ & & 1 & \cdots & \cdots & \cdots \\ & & & 1 & \cdots & \cdots \\ & & & & 1 & \cdots \\ \text{sym.} & & & & & 1 \end{bmatrix} \begin{bmatrix} \vdots \\ w_{i,U} + w_{j,U} \\ \vdots \\ w_{j-1,U} \\ w_{j+1,U} \\ \vdots \end{bmatrix} \\
&= \begin{bmatrix} \underbrace{\sum_{\substack{k=1 \\ k \neq i, j}}^n w_{k,A} E_{1k} + (w_{i,A} + w_{j,A}) E_{1i} \cdots \sum_{\substack{k=1 \\ k \neq i, j}}^n w_{k,A} E_{nk} + (w_{i,A} + w_{j,A}) E_{ni}}_{n-1 \text{ entities}} \end{bmatrix} \begin{bmatrix} \vdots \\ w_{i,U} + w_{j,U} \\ \vdots \\ w_{j-1,U} \\ w_{j+1,U} \\ \vdots \end{bmatrix} \\
&= \sum_{\substack{p=1 \\ p \neq i, j}}^n w_{p,U} \left(\sum_{\substack{k=1 \\ k \neq i, j}}^n w_{k,A} E_{pk} + (w_{i,A} + w_{j,A}) E_{pi} \right) + (w_{i,U} + w_{j,U}) \left(\sum_{\substack{k=1 \\ k \neq i, j}}^n w_{k,A} E_{ik} + (w_{i,A} + w_{j,A}) E_{ii} \right) \\
&= \sum_{p=1}^n w_{p,U} \left(\sum_{\substack{k=1 \\ k \neq i, j}}^n w_{k,A} E_{pk} + (w_{i,A} + w_{j,A}) E_{pi} \right) \quad \because E_{ik} = E_{jk} \quad \forall k, 1 \leq k \leq n, \text{ and } E_{ii} = E_{jj} \\
&= \sum_{p=1}^n w_{p,U} \sum_{k=1}^n w_{k,A} E_{pk} \quad \because E_{pi} = E_{pj} \quad \forall p, 1 \leq p \leq n \\
&= \vec{m}_A^T E \vec{m}_U \tag{3.20}
\end{aligned}$$

Substituting the above equation with $U = A$, we have

$$\vec{m}_A^T E \vec{m}_A = \vec{m}_{A, \text{reduced}}^T E_{\text{reduced}} \vec{m}_{A, \text{reduced}} \tag{3.21}$$

and the same equality holds for $A = U$. Therefore, the denominator of the cosine, $\sqrt{\vec{m}_A^T E \vec{m}_A} \times \sqrt{\vec{m}_U^T E \vec{m}_U}$, is also the same as the reduced system. As a result, the transformation through matrix $E = D^T D$ correctly produces the same result when synonymic information are modeled using two different spaces, namely the original n -dimensional space and a reduced vector space with the synonymic axes collapsed into one.

The vector space transformation developed above is shown to produce the desired results given a user-defined matching algorithm to define non-Boolean partial matches. The same analysis is also shown to reduce to the Boolean Vector model, such as one in Equation (3.6), if the user-defined algorithm represents an identity matrix that assumes axis independency. Therefore, a cosine computation based on our proposed vector space transformation always correctly models both Boolean and non-Boolean matches, as long as a correctly-populated matching matrix E is defined. Concept, author-prescribed index and drinking water contaminant comparisons defined in Section 3.3.1 can be performed using this transformation mode. As an illustrative example, we will define the comparison of effective dates in the next section.

3.3.2.2 Comparisons of Effective Dates

Prior to the discussion on the comparison technique for effective dates, we first review what is encapsulated inside an effective date feature. There are four types of date tags as defined in Section 2.4.7:

- `<date date="January 24, 1978" num="1" />`
- `<date to="May 18, 1994" num="2" />`
- `<date from="October 13, 1978" num="2" />`
- `<date from="January 1, 1993" to="December 31, 2001" num="1" />`

The first XML element represents an effective date without a quantifier, whereas the other three tags are quantified as a start date, end date and a range of dates respectively. As suggested in the previous section, we can employ a Boolean matching algorithm to compare effective dates analogous to concept comparisons. In a Boolean matching model, an n -entry provision-date vector \vec{d}_M can be constructed per provision M , where n represent the number of *unique* date features identified in the corpus. A unique date feature is defined to be an XML date element that differs from all other date elements in any field except the “num” field. The drawback of a simple Boolean match as described is obvious in this case: with the wide range of possible effective dates in our regulatory corpus, a Boolean match will likely result in zero cosines among all provisions. Thus, we will adopt the transformation model in Section 3.3.2.1 using the following example of a user-defined effective date matching algorithm on an effective date pair d_1 and d_2 :

```

// main function: if the beginning of a date range is before
// the end of another and vice versa, they overlap.
IF (Is_Before_Or_Equal(Start(d1),End(d2)) &&
    Is_Before_Or_Equal(Start(d2),End(d1)))
    RETURN 1;
ELSE
    RETURN 0;

// Start subroutine: returns the start date of a date element
Start(d) {
    IF (date(d) != null)
        RETURN date(d) - 1/2 yr;
    IF (from(d) == null)
        RETURN to(d) - 1 yr;
    RETURN from(d);
}

// End subroutine: returns the end date of a date element
End(d) {
    IF (date(d) != null)
        RETURN date(d) + 1/2 yr;
    IF (to(d) == null)
        RETURN from(d) + 1 yr;
    RETURN to(d);
}

```

Figure 3.5: Pseudo-Code of a User-Defined Effective Date Matching Algorithm

The above pseudo-code implements the matching concept as illustrated in Figure 3.6. A one-year time frame is used to determine whether two date features are related or not, except in the case where exact “to” and “from” dates are specified in the tag which are then used to represent the time frame. The algorithm assigns a score of 1 to any pair of date features that overlap with each other in that time frame, and 0 for all other cases. Essentially, this matching algorithm predicts that two provisions with effective dates close to one another, for example, within a year, are related. A typical *point* match in a Boolean model, which would result in a 0% match between “1/1/03” and “1/2/03,” is transformed into a less restrictive *range* match using this algorithm. The one-year time frame can be easily adjusted to reflect different expert opinions.

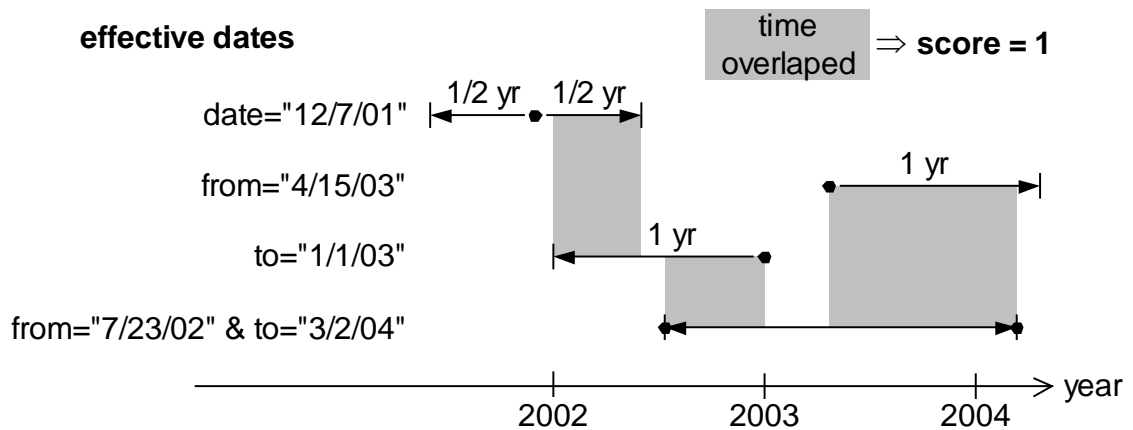


Figure 3.6: Illustration of an Example of a User-defined Effective Date Comparison Algorithm

To compute the start and end date of a time frame, the four types of date features are handled differently. A date feature with only a “date” field without a quantifier is regarded as the middle of a one-year time frame, where the start date is half year before the specified date in the tag, and the end date is half year after. A date feature with only a “from” field, which shows that the provision is effective from this date onwards, results in a start date as the specified “from” date and an end date a year after. Similarly, a date

feature with only a “to” field is the reverse case of a “from” date. For instance, as shown in Figure 3.6, a date tag `<date to="1/1/03" . . . >` is interpreted as a time frame from January 1, 2002 to January 1, 2003. Any other date feature with a time frame that overlaps with this period is regarded as a 100% match. Finally, we use the specified range as the time frame for a date feature with both “from” and “to” fields, instead of an imposed one-year period for comparison.

As defined at the beginning of this section, an n -entry provision-date vector \vec{d}_M is constructed per provision M , where n represents the total number of unique date features in the corpus. The frequency count of date feature i in provision M defines the weight $w_{i,M}$ in $\vec{d}_M = (w_{1,M}, w_{2,M}, \dots, w_{n,M})$. The similarity score $F(A, U, i=date)$, based on a user-defined date matching algorithm between Sections A and U, is given by Equation (3.22). We follow the vector space transformation model defined in Section 3.3.2, where E here represents the effective date matching matrix obtained using a user-defined effective date matching algorithm such as the pseudo-code in Figure 3.5.

$$F(A, U, i=date) = \frac{\vec{d}_A^T E \vec{d}_U}{\sqrt{\vec{d}_A^T E \vec{d}_A} \times \sqrt{\vec{d}_U^T E \vec{d}_U}} \quad (3.22)$$

3.3.3 Discussions of Other Feature Comparisons

In Sections 2.4.1 to 2.4.7, we listed a handful of features that are extracted from a regulatory corpus focusing on drinking water control and disabled access. The extraction list includes generic features that are applicable on all areas of regulations, as well as domain-specific features provided by knowledge experts. We discussed several feature comparison techniques in Sections 3.3.1 to 3.3.2, which represent a mixture of Boolean matches and non-Boolean matches. There remain several extracted features included in Chapter 2 that are not discussed here, such as the comparisons of glossary terms, exceptions and definitions. In addition, knowledge experts can always define additional

features and their matching algorithms that are important in relatedness identification in a particular domain.

As shown in Equation (3.20), our proposed vector space transformation is capable of modeling both traditional Boolean matches and non-Boolean matches introduced by user-defined algorithms. Therefore, if a domain expert desires to define matching algorithms for the remaining extracted features such as glossary terms, our transformation model can incorporate such information with ease and consistency with other feature comparisons. In addition, the base score f_0 is composed of the above feature comparisons as shown in Equation (3.5), with the flexibility to add on the remaining features that are not compared.

Since a non-Boolean model can also accommodate Boolean matches, it seems unavailing to implement a Boolean matching for any feature at all. For instance, a non-Boolean model can be easily adopted for dwc comparisons, shown in Section 3.3.1.2 using a Boolean model. Using our vector space transformation, an ontology matching matrix E can be defined with only 1's and 0's but no fractional matches to represent synonyms. We show below a simple pseudo-code to generate the E matrix using synonymic information from an ontology as shown in Figure 2.8. If one desires, the E matrix can be altered to reflect general-purpose dictionary information instead of domain-specific ontologies, or a combination of both.

```
// Populate the E matrix using synonymic information from an
// dwc ontology
FOR (i = 0; i < n; i++)
    FOR (j = 0; j < n; j++)
        IF (i == j)
            E[i, j] = 1;
        ELSE IF (Is_Synonym_From_Ontology(dwc(i), dwc(j)))
            E[i, j] = 1;
        ELSE
            E[i, j] = 0;
```

Figure 3.7: Pseudo-Code to Populate the E Matrix Using Ontology Information

With the drinking water contaminant matching matrix E correctly defined to represent ontology information, provision comparison can be performed using a vector space transformation approach. We define an n -entry provision-contaminant vector \vec{t}_M per provision M , where n is the total number of unique contaminants identified in the corpus. Again, uniqueness is defined without ontology knowledge of synonyms or acronyms. Using a dwc matching matrix E generated as shown in Figure 3.7, the computation of the similarity score $F(A, U, i=dwc)$, based on a dwc comparison between Sections A and U , is shown in Equation (3.23).

$$F(A, U, i=dwc) = \frac{\vec{t}_A^T E \vec{t}_U}{\sqrt{\vec{t}_A^T E \vec{t}_A} \times \sqrt{\vec{t}_U^T E \vec{t}_U}} \quad (3.23)$$

Comparing Equation (3.23) to Equation (3.8), our proposed vector space transformation model effectively collapses *unique* contaminants via matrix E to *consolidated* contaminants as manually performed in Equation (3.8). Although the framework for a non-Boolean model is readily available as shown, the value of using a Boolean model cannot be overlooked. It is faster computationally to perform a Boolean feature matching, as it is a simpler analysis after all. The non-Boolean model is recommended when domain knowledge is available.

3.4 Score Refinement Based on Regulation Structure

Score refinement utilizes the tree structure of regulations to refine the base score f_0 between provisions in order to obtain a better and more complete analysis. As shown in Figure 3.1, we take Sections A and U as our point of comparison, where these two interested nodes belong to two different regulation trees. The base score f_0 represents a computation of the similarity between two nodes based solely on the node contents, such

as the number of concepts and measurements Sections A and U share. We can interpret f_0 as a basis of relatedness analysis formed on the shared clusters of similar features between these two interested nodes A and U . Other nodes in these two regulation trees as well as the structures of the two trees are ignored in a base score analysis.

To utilize the tree structure of regulations and thus potentially include the influence of other nodes on the similarity between nodes A and U , we propose several score refinements in this section. In Section 3.2.1, we define $p_{sc}(A)$ to represent collectively the immediate neighbors of node A , while $ref(A)$ is defined to symbolize collectively the references to other nodes from node A . Both $p_{sc}(A)$ and $ref(A)$ represent a set of nodes in the same tree that are related¹⁷ to node A through the structure of the regulation tree that A belongs to. Our score refinements utilize this set of related nodes to reveal additional similarity evidences. Each set of nodes represents a different type of score refinements. Section 3.4.1 introduces neighbor inclusion which makes use of the p_{sc} set, while reference distribution, which is discussed in Section 3.4.2, accounts for evidences from the ref set. Our analysis is complete with score refinements, which incorporate the structure of nodes as well as node contents into comparisons.

As clusters of similarity scores are considered in the following sections, we will use the similarity score matrix Φ defined in Section 3.2.2 to formulate the analyses. The base scores developed in Section 3.3 are represented as Φ_0 . The refined similarity score matrices, such as Φ_{rd} for reference distribution, will be defined in subsequent sections.

3.4.1 Neighbor Inclusion

Neighbor inclusion defines the refinement process where the immediate neighbors of the interested node pair are included in the comparison to reveal potential hidden similarity between the interested pair. A direct comparison between two nodes might not identify

¹⁷ Related as defined in Section 3.2 - “connected by reason of an established or discoverable relation.”

similarities that are embedded in the neighboring nodes; for example, nodes A and U that are related without sharing the same features will have a low f_0 , while node A and $p_{sc}(U)$ can be related through similar features which results in a high f_0 .

Neighbor inclusion is composed of two types of analysis: a *self* versus *p_{sc}* comparison and a *p_{sc}* versus *p_{sc}* comparison. A *self* versus *p_{sc}* analysis, abbreviated as *s-p_{sc}*, compares the self-content of a node with the *p_{sc}* of the other interested node. For instance, using Section A and U as our running example, Section A itself is compared with $p_{sc}(U)$, and vice versa, to produce the score $f_{s-p_{sc}}(A, U)$. A *p_{sc}* versus *p_{sc}* comparison, abbreviated as *p_{sc}-p_{sc}*, is similar to an *s-p_{sc}* analysis, which takes into account the comparisons between $p_{sc}(A)$ and $p_{sc}(U)$ to produce the score $f_{p_{sc}-p_{sc}}(A, U)$.

Section 3.4.1.1 introduces a *p_{sc}-p_{sc}* comparison, with the *s-p_{sc}* comparison follows in Section 3.4.1.2. By separating the comparisons of *s-p_{sc}* and *p_{sc}-p_{sc}*, flexibility is maintained in our analysis where different weights can be assigned to these two comparisons. Section 3.4.1.3 discusses the refined score f_{ni} obtained from neighbor inclusion by weighting the *s-p_{sc}* and *p_{sc}-p_{sc}* comparisons differently. Indeed, the weight of an *s-p_{sc}* comparison should be higher than that of a *p_{sc}-p_{sc}* comparison, which will be explained in Section 3.4.1.3 as well.

3.4.1.1 *P_{sc}* Vs. *P_{sc}*

The set of nodes in $p_{sc}(A)$ is related to node A through a parent, sibling or child relationship. As defined in Section 3.2, similarity analysis aims to reveal entities that are “connected by reason of an established or discoverable relation”; therefore, we utilize the *p_{sc}* relationships between nodes to refine the comparison in an attempt to discover more similarity relationships. Continuing with our running example of comparisons between Sections A and U , their immediate neighbors are compared and the average similarity score obtained from neighbor comparisons defines $f_{p_{sc}-p_{sc}}(A, U)$. Essentially, we have *diffusion* of similarity between clusters of nodes in the tree; Figure 3.8 best illustrates the

idea. The similarity between $p_{sc}(A_1)$ and $p_{sc}(U_1)$, represented by clusters shaded in dark gray, diffuses to nodes A_1 and U_1 . Likewise, the dissimilarity between $p_{sc}(A_2)$ and $p_{sc}(U_2)$, shown using lightly-shaded clusters, spreads to nodes A_2 and U_2 . In other words, a p_{sc} - p_{sc} analysis 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. In Chapter 4, we will discuss several evaluation models, which include evaluation of results obtained from neighbor inclusion to assess the validity of the above statement.

To formulate a p_{sc} - p_{sc} comparison, we first show a pseudo-code in Figure 3.9 which will be later transformed into matrix notations. Each node in $p_{sc}(A)$ is compared with every node in $p_{sc}(U)$, so a $p_{sc}(A)$ set with x nodes and a $p_{sc}(U)$ set with y nodes will result in $x \times y$ comparisons. Each comparison is an f_0 base score computation. As shown in Figure 3.9, $f_{p_{sc}-p_{sc}}$ denotes the average base scores between the neighbors to reflect the *diffusion* of similarity among clusters of nodes explained above. The pseudo-code shows a new computation of the base score between nodes A and U whenever their neighbors are involved in a refinement; however, a table of pre-generated f_0 scores can be used at implementation time and $x \times y$ becomes the number of table lookups per comparison. In fact, we will now introduce a matrix denotation, which is the simplest representation and computation.

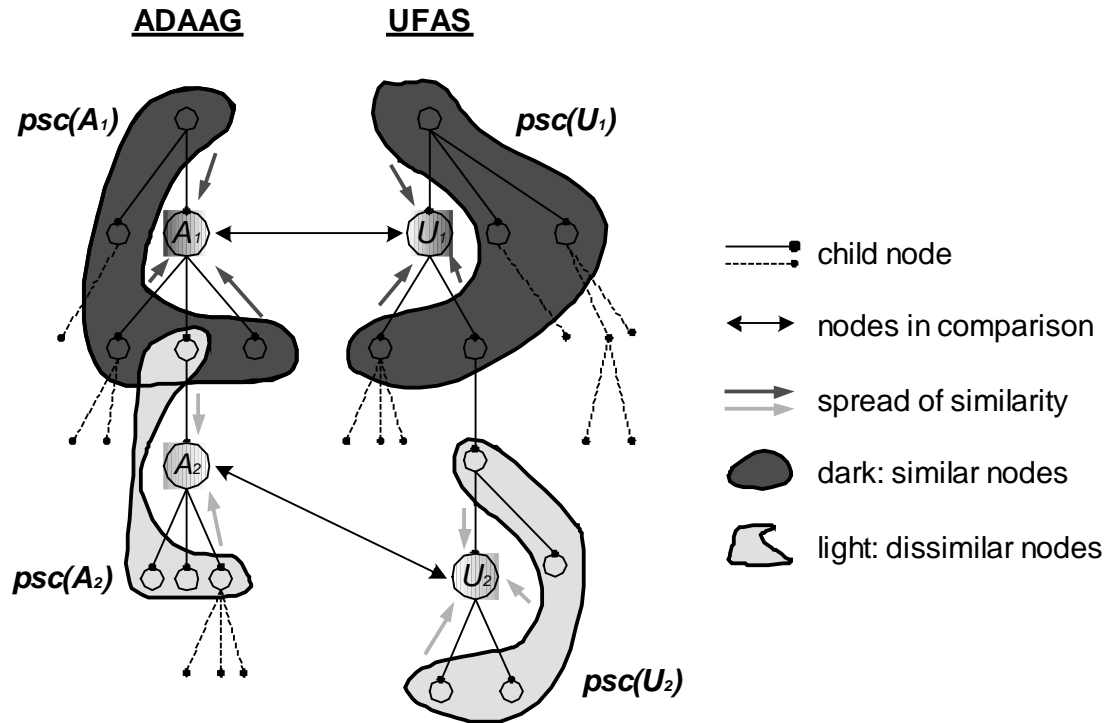


Figure 3.8: Diffusion of Similarity among Clusters of Nodes Introduced by a *psc-psc* Comparison

```

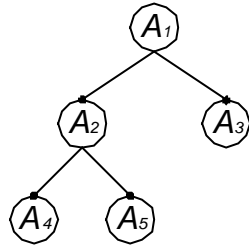
// compute f_psc-psc(A, U) between Sections A and U
DOUBLE score = 0;
FOREACH node a in psc(A)
    FOREACH node u in psc(U)
        score = score + f_0(a, u);
RETURN score / (Sizeof(psc(A)) × Sizeof(psc(U)));

```

Figure 3.9: Pseudo-Code of an $f_{psc-psc}$ Computation

In Section 3.2.2, Φ is defined to represent the similarity scores between two regulations. We will use the base score Φ_0 together with the neighbor structure of regulations to obtain a refined score $\Phi_{psc-psc}$ based on a *psc-psc* comparison as shown in Figure 3.9. A neighbor structure matrix N is defined for each regulation to be compared. N is a square

matrix where the dimension represents the number of sections in the regulation. Each Section i corresponds to row i and column i of N . Entry $N(i, j)$ is 0 if $i \notin psc(j)$; in other words, if Section i is not a psc of j , the entry is 0. Since psc is a reciprocal relationship, the condition is the same as $j \notin psc(i)$. For $j \in psc(i)$, entry $N(i, j)$ is $1/k$ where k is the total number of neighbors of i . As in the computation shown in the pseudo-code in Figure 3.9, k is equal to the size of set $psc(i)$. The diagonals are zero since a section cannot be a neighbor of itself by definition, i.e., $i \notin psc(i)$. Assuming that the tree is well defined without singleton nodes, there exists at least one neighbor for each node. As a result, there exists at least one non-zero entry for all rows and columns. It follows that the rows of N add up to one and all of the elements in N are nonnegative, thus N is a stochastic matrix [14]. An example N matrix is shown in Figure 3.10(b) with the tree structure in Figure 3.10(a).



$$N_A = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

(a) Example Tree of Regulation A (b) A Neighbor Structure Matrix N_A

Figure 3.10: A Neighbor Structure Matrix to Represent Tree Structure

We will now show that the similarity score matrix $\Phi_{psc-psc}$, which is composed of scores $f_{psc-psc}$ between sections in regulations A and U , can be represented as $N_A \Phi_0 N_U^T$. N_A and N_U are the neighbor structure matrices of regulations A and U respectively, and Φ_0 denotes their base scores with rows and columns representing sections from A and U correspondingly. Equation (3.24) shows the proof.

Let $a_i =$ Section i in regulation A (3.24)

$u_i =$ Section i in regulation U

$$Z = N_A \Phi_0 N_U^T$$

$$\lambda_{ij} = \begin{cases} 0 & \text{if } a_j \notin psc(a_i) \\ \frac{1}{sizeof(psc(a_i))} & \text{otherwise} \end{cases}$$

$$\mu_{ij} = \begin{cases} 0 & \text{if } u_j \notin psc(u_i) \\ \frac{1}{sizeof(psc(u_i))} & \text{otherwise} \end{cases}$$

$$\begin{aligned} Z(i, j) &= \sum_l \sum_k N_{A,ik} \Phi_{0,kl} N_{U,jl} \\ &= \sum_l N_{U,jl} \sum_k \lambda_{ik} f_0(a_k, u_l) \\ &= \sum_l \mu_{jl} \sum_k \lambda_{ik} f_0(a_k, u_l) \\ &= \frac{1}{sizeof(psc(a_i)) \times sizeof(psc(u_j))} \sum_{u_p \in psc(u_j)} \sum_{a_p \in psc(a_i)} f_0(a_p, u_p) \\ &= f_{psc-psc}(a_i, u_j) \\ &= \Phi_{psc-psc}(i, j) \end{aligned}$$

Therefore, a more compact representation of the pseudo-code computation shown in Figure 3.9 is $\Phi_{psc-psc} = N_A \Phi_0 N_U^T$. As is apparent in the above equations, the stochastic property of N_A and N_U translates to the normalization in the denominator. In developing a *self* versus *psc* analysis that follows in the next section, we will use the same approach with the same matrices N_A , Φ_0 and N_U as well as λ and μ defined in Equation (3.24).

3.4.1.2 Self Vs. Psc

If the immediate neighbors of a pair of interested nodes are compared to refine the similarity score between the pair as explained above, there exists an even more *direct*

comparison that should not be neglected in the analysis. Continuing with Sections A and U as the point of comparison, a *self* versus *psc* analysis compares the self content of node A with $psc(U)$, as well as the self content of node U with $psc(A)$, to produce $f_{s-psc}(A, U)$. The *diffusion* of similarity between clusters of nodes is more direct in this case; if Section A shares similarity with the immediate neighbors of Section U , and vice versa, an implied similarity exists between Sections A and U . Figure 3.11 illustrates the idea. The similarity between A_1 and $psc(U_1)$, represented by clusters shaded in dark gray, spreads to nodes A_1 and U_1 . Analogously, the dissimilarity between $psc(A_2)$ and U_2 , shown using lightly-shaded clusters, diffuses to nodes A_2 and U_2 .

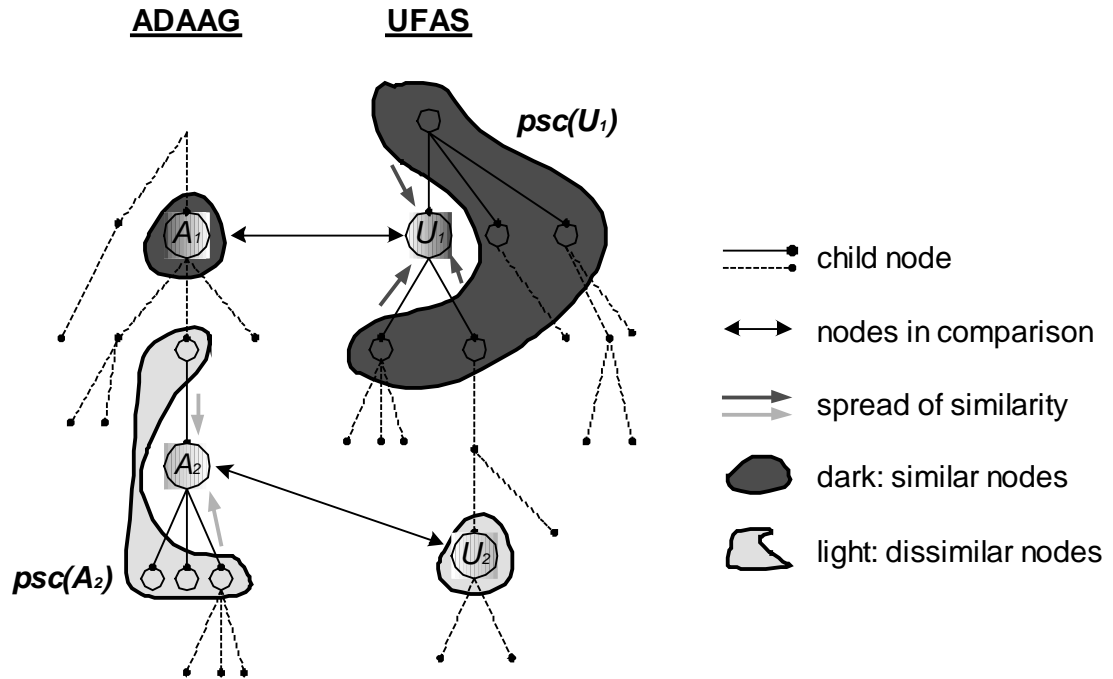


Figure 3.11: Diffusion of Similarity among Cluster of Nodes Introduced by an *s-psc* Comparison

In Figure 3.12, a pseudo-code that follows the same approach in the previous section is shown. Each node in $psc(A)$ is compared with node U , and vice versa, so a $psc(A)$ set

with x nodes and a $psc(U)$ set with y nodes will result in $x+y$ comparisons. Each comparison is an f_0 base score computation. f_{s-psc} represents the average base scores between node A and the neighbors of U as well as between node U and the neighbors of A . Again, $x+y$ does not reflect the number of node pair comparisons per $s-psc$ computation, since a table of pre-generated f_0 scores can be used at implementation time.

```

// compute f_s-psc(A, U) between Sections A and U
DOUBLE score1 = 0, score2 = 0;
FOREACH node a in psc(A)
    score1 = score1 + f_0(a, U);
FOREACH node u in psc(U)
    score2 = score2 + f_0(A, u);
RETURN (score1/Sizeof(psc(A)) + score2/Sizeof(psc(U))) / 2;

```

Figure 3.12: Pseudo-Code of an f_{s-psc} Computation

To formulate a similar matrix representation for an $s-psc$ comparison, we continue to use the defined neighbor structure matrices N_A and N_U for regulations A and U . Comparing f_{s-psc} with $f_{psc-psc}$, we observe that the computations are very similar. We showed in Equation (3.24) that $\Phi_{psc-psc} = N_A \Phi_0 N_U^T$; substituting N_A with the identity matrix I , the ij^{th} entry of the expression becomes a comparison between node i in regulation A and the psc of node j in regulation U . Node i in regulation A is regarded as the only psc of itself in this substitution, which translates to a direct comparison between i and $psc(j)$. Similarly, substituting N_U^T with I gives the expression for comparisons between the psc in regulation A and nodes in regulation U . Combining the two expressions, we have a matrix representation of an $s-psc$ analysis shown in Equation (3.25).

$$\begin{aligned}
 \Phi_{s-psc} &= \frac{1}{2} (I \Phi_0 N_U^T + N_A \Phi_0 I) \\
 &= \frac{1}{2} (\Phi_0 N_U^T + N_A \Phi_0)
 \end{aligned} \tag{3.25}$$

3.4.1.3 Combination of Both Analyses

Φ_{s-psc} and $\Phi_{psc-psc}$ represent correspondingly the average base scores obtained from a self-neighbor and neighbor-neighbor comparison. As suggested in Section 3.4, Φ_o lacks appreciation of the natural structure of regulations, and score refinements are supposed to fill the gap. However, Φ_{s-psc} and $\Phi_{psc-psc}$ alone also overlook a direct content comparison between nodes. Evidently, neighbor inclusion should not exclude a *self* versus *self* comparison, or in our notations, a Φ_o component. We introduce a weighted combination of Φ_o , Φ_{s-psc} and $\Phi_{psc-psc}$ as the similarity score Φ_{ni} from neighbor inclusion:

$$\Phi_{ni} = \alpha_0 \Phi_o + \alpha_{s-psc} \Phi_{s-psc} + \alpha_{psc-psc} \Phi_{psc-psc} \quad (3.26)$$

where α_0 , α_{s-psc} and $\alpha_{psc-psc}$ represent the weighting factor of the base, *s-psc* and *psc-psc* comparisons respectively. Clearly, the α 's should sum up to 1 to maintain a similarity score that ranges from 0 to 1 as defined earlier. α_0 should be greater than α_{s-psc} and $\alpha_{psc-psc}$, since f_o represents the most direct comparison between two nodes, while f_{s-psc} and $f_{psc-psc}$ represent a less direct diffusion of similarity from neighbors to nodes. Therefore, the base score should still be the most important component among the three scores. From an *s-psc* comparison to a *psc-psc* comparison, another layer of indirection is inferred, as a *psc-psc* comparison involves nothing but the neighbors, whereas an *s-psc* comparison still includes the interested nodes in the analysis. As a result, α_{s-psc} should be greater than $\alpha_{psc-psc}$. In other words, the weighting coefficients have the properties that:

$$\alpha_0 > \alpha_{s-psc} > \alpha_{psc-psc} > 0, \text{ and} \quad (3.27)$$

$$\alpha_0 + \alpha_{s-psc} + \alpha_{psc-psc} = 1$$

The intuition to prioritize between *s-psc* and *psc-psc* analyses explains why the two comparisons are separated as two different refinement processes. Chapter 4 discusses several experimented values for α 's and the corresponding changes in results. Other

usages of Φ_o , Φ_{s-psc} and $\Phi_{psc-psc}$, such as taking the set intersection of the most related provisions produced by different Φ 's as the final ranking, can also be experimented.

3.4.2 Reference Distribution

While neighbor inclusion accounts for the diffusion of similarity from the immediate neighboring nodes, reference distribution incorporates the influence of the not-so-immediate neighbors into the analysis. It utilizes the unique referential structure of regulations to further refine the similarity score. To understand the intuition behind reference distribution, we should note that regulations are heavily referenced documents, which contributes to the difficulty in reading and understanding them. Most regulatory documents are heavily self-referenced but not cross-referenced: they do not point to other regulations or outside materials as much as they cite provisions within the same regulation. For instance, the entire UFAS [101] only referenced the Code of Federal Regulations (CFR) four times, and it made no citation to the ADAAG [1] and vice versa.

The assumption behind reference distribution is that similar sections reference similar sections. In essence, the heavily referenced nature of regulatory documents provides extra information about provisions and the relationship between them, which can be useful in revealing potential hidden similarity between provisions. In Section 3.1.3, academic citation analysis [17] and link analysis [20] are briefly reviewed, where citations are used to help document comparisons. The algorithms cannot be directly imported to a regulatory domain – regulations are much more complicated and they exist as *separate* trees of provisions. One can visualize the problem as separate islands of information. Each island represents a regulation. Within an island, information is highly bridged with self-references among provisions. Across islands, there are very few or no connecting bridges, that is, cross-references between regulations.

To formulate a score refinement based on reference distribution, we will take Sections A and U as the interested pair of nodes. Analogous to neighbor inclusion, reference

distribution implies that similarity among referenced nodes of nodes A and U , as well as similarity between node A and $ref(U)$, and vice versa, diffuse to A and U themselves. Following the terminology of neighbor inclusion, reference distribution thus includes a *self* versus *ref* analysis and a *ref* versus *ref* analysis, abbreviated as *s-ref* and *ref-ref* respectively as shown in Figure 3.13.

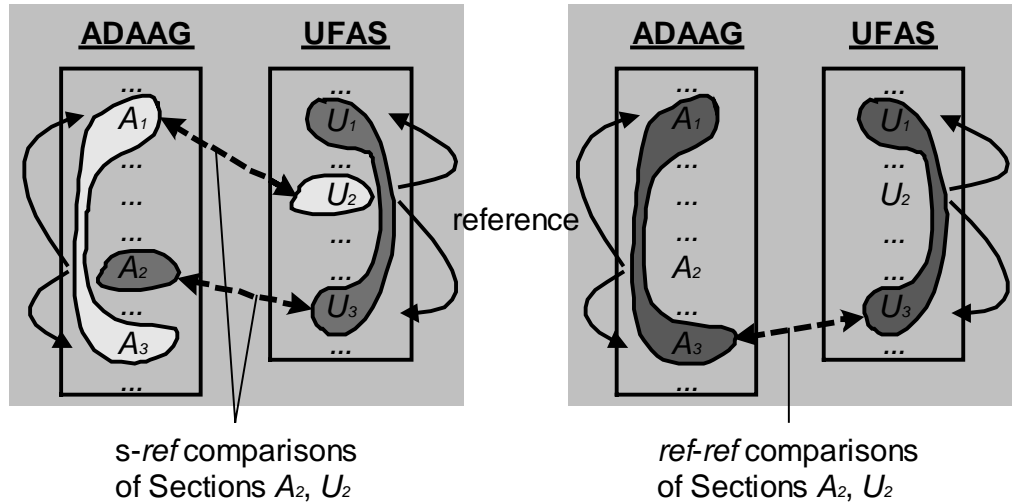


Figure 3.13: Illustrations of an *s-ref* and a *ref-ref* Comparison

We follow an approach similar to the *s-psc* and *psc-psc* analyses in neighbor inclusion. Pseudo-codes for an f_{s-ref} and $f_{ref-ref}$ computation are shown in Figure 3.14 and Figure 3.15, respectively. The computations of f_{s-ref} and $f_{ref-ref}$ are identical to that of neighbor inclusion, except that each score, based on node j referenced by node i , is multiplied with the number of times i references j . As a result, the normalization factor `sizeof(psc(node i))` is replaced with the total number of references including duplicates from node i . The reasoning behind the difference in formulation is that node i can potentially link to node j multiple times, which is indeed an observed fact in regulatory documents. A node j that is cited multiple times is naturally more important than other references, and thus the change in the computation. In fact, the neighbor inclusion computation, as shown in Figure 3.9 and Figure 3.12, can be interpreted as a

multiplication of one with each score between neighbors, as each neighbor receives one unit of importance.

```

// compute f_s-ref(A, U) between Sections A and U
// num_ref(o, j) returns the number of times i references j.
DOUBLE score1 = 0, score2 = 0, size1 = 0, size2 = 0;
FOREACH node a in ref(A)
    score1 = score1 + num_ref(A, a) × f_0(a, U);
    size1 = size1 + num_ref(A, a);
FOREACH node u in ref(U)
    score2 = score2 + num_ref(U, u) × f_0(A, u);
    size2 = size2 + num_ref(U, u);
RETURN (score1/size1 + score2/size2) / 2;

```

Figure 3.14: Pseudo-Code of an f_{s-ref} Computation

```

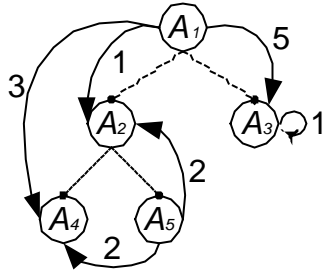
// compute f_ref-ref(A, U) between Sections A and U
DOUBLE score = 0, size = 0;
FOREACH node a in ref(A)
    FOREACH node u in ref(U)
        score = score + num_ref(A, a) × num_ref(U, u) ×
            f_0(a, u);
        size = size + num_ref(A, a) × num_ref(U, u);
RETURN score / size;

```

Figure 3.15: Pseudo-Code of an $f_{ref-ref}$ Computation

In neighbor inclusion, we showed that $\Phi_{psc-psc} = N_A \Phi_0 N_U^T$ and $\Phi_{s-psc} = \frac{1}{2} (\Phi_0 N_U^T + N_A \Phi_0)$. If we define a reference structure matrix R to replace the neighbor structure matrix N , we have the expressions $R_A \Phi_0 R_U^T$ and $\frac{1}{2} (\Phi_0 R_U^T + R_A \Phi_0)$, which will be shown below to represent respectively $\Phi_{ref-ref}$ and Φ_{s-ref} for a properly defined R . Similar to N , row i and column i of R denote Section i . Entry R_{ij} is 0 if $j \notin ref(i)$; note that ref is not a symmetric relationship as psc . For $j \in ref(i)$, entry R_{ij} is the number of citations from i to j divided by the total number of citations from i including duplicates. It follows that the rows of R add up to one for rows representing sections that make at least one reference; therefore, R

is not necessarily stochastic. An example R matrix is shown in Figure 3.16(b) with the referential structure in Figure 3.16(a).



$$R_A = \begin{bmatrix} 0 & 1/9 & 5/9 & 3/9 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2/4 & 0 & 2/4 & 0 \end{bmatrix}$$

(a) Example Tree of Regulation A

(b) A Reference Structure Matrix R_A

Figure 3.16: A Reference Structure Matrix to Represent References among Nodes

The proof shown in Equation (3.24) can be easily modified to show $\Phi_{ref-ref} = R_A \Phi_0 R_U^T$. λ and μ are redefined as

$$\lambda_{ij} = \begin{cases} 0 & \text{if } a_j \notin ref(a_i) \\ \frac{num_ref(a_i, a_j)}{\sum_{a_p \in ref(a_i)} num_ref(a_i, a_p)} & \text{otherwise} \end{cases}$$

$$\mu_{ij} = \begin{cases} 0 & \text{if } u_j \notin ref(u_i) \\ \frac{num_ref(u_i, u_j)}{\sum_{u_p \in ref(u_i)} num_ref(u_i, u_p)} & \text{otherwise} \end{cases}$$

which represent the proportional importance of reference j among all references from node i . $num_ref(i, j)$ returns the number of references from node i to node j in the above definition. Substituting matrix N with R and set psc with ref along with the redefined λ and μ , Equation (3.24) becomes

Let $Y = R_A \Phi_0 R_U^T$

$$\begin{aligned}
 Y(i, j) &= \frac{\sum_{u_p \in \text{ref}(u_j)} \sum_{a_p \in \text{ref}(a_i)} \text{num_ref}(u_j, u_p) \times \text{num_ref}(a_i, a_p) \times f_0(a_p, u_p)}{\sum_{u_p \in \text{ref}(u_j)} \sum_{a_p \in \text{ref}(a_i)} \text{num_ref}(u_j, u_p) \times \text{num_ref}(a_i, a_p)} \\
 &= f_{\text{ref-ref}}(a_i, u_j) \\
 &= \Phi_{\text{ref-ref}}(i, j)
 \end{aligned}$$

Thus, we have $\Phi_{\text{ref-ref}} = R_A \Phi_0 R_U^T$. Similarly, $\Phi_{s\text{-ref}} = 1/2 (I \Phi_0 R_U^T + R_A \Phi_0 I) = 1/2 (\Phi_0 R_U^T + R_A \Phi_0)$ follows by substituting R_A and R_U with the identity matrix I separately. The same linear combination approach shown in Section 3.4.1.3 for neighbor inclusion is used to combine $\Phi_{\text{ref-ref}}$ and $\Phi_{s\text{-ref}}$ to form the final similarity score Φ_{rd} from reference distribution:

$$\Phi_{rd} = \alpha_0 \Phi_0 + \alpha_{s\text{-ref}} \Phi_{s\text{-ref}} + \alpha_{\text{ref-ref}} \Phi_{\text{ref-ref}} \quad (3.28)$$

where $\alpha_0 > \alpha_{s\text{-ref}} > \alpha_{\text{ref-ref}} > 0$ and $\alpha_0 + \alpha_{s\text{-ref}} + \alpha_{\text{ref-ref}} = 1$. As a result, reference distribution accounts for the unique referential structure of regulations by combining comparisons of the referenced nodes $\text{ref}(A)$ vs. $\text{ref}(U)$ with comparisons of node A vs. $\text{ref}(U)$ and node U vs. $\text{ref}(A)$.

We note that referencing is directional unlike an immediate neighboring relationship, which leads us to further investigate the semantics of a reciprocal referential relationship. For node i , we define its ‘‘in reference’’ as the reference from other nodes to node i , while its ‘‘out reference’’ remains the regular reference to other nodes as defined in previous context. Reference distribution infers that similarity diffuses from the ‘‘out references’’ to the nodes that point to them. It seems that the reciprocal argument should hold as well, that is, similarity also diffuses from nodes to their citations. In other words, the ‘‘in

references” of a node i should be incorporated into the score refinement of node i to account for their potential influence on node i .

In fact, it is simple to include the “in references” using our developed analysis. Redefining the terminology, an *out-out* analysis is a “out reference” versus “out reference” comparison, which is a *ref-ref* comparison in previous context. Thus, we have $\Phi_{out-out} = R_A \Phi_0 R_U^T$ and $\Phi_{s-out} = \frac{1}{2} (\Phi_0 R_U^T + R_A \Phi_0)$ as shown earlier. R^T would correctly model the “in reference” structure if columns of R were normalized according to the number of “in references,” instead of rows of R currently normalized based on the number of “out references”. Therefore, we define matrix \bar{R} similar to R , where non-empty columns of \bar{R} are normalized to sum up to one, compared to a row normalization of R . Thus, the ij^{th} entry of \bar{R}^T represents the proportional weight of the “in reference” of node i from node j among all “in references” of node i . As a result, we have $\Phi_{in-in} = \bar{R}_A^T \Phi_0 \bar{R}_U$ and $\Phi_{s-in} = \frac{1}{2} (\Phi_0 \bar{R}_U + \bar{R}_A^T \Phi_0)$. One can go further to compare the “in references” to node i with the “out references” from node j in refining the similarity score between i and j , which can be achieved with $\Phi_{in-out} = \frac{1}{2} (\bar{R}_A^T \Phi_0 R_U^T + R_A \Phi_0 \bar{R}_U)$. Figure 3.17 illustrates the concept of *in-in* and *in-out* reference comparisons.

As much as it seems appropriate and concise to include the “in references” in a reference distribution analysis, one could also argue oppositely. In particular, an “in reference” is not an explicit reference; a provision passively *receives* an “in reference” from another provision, whereas a provision *explicitly* cites another provision as an “out reference”. While an *out-out* analysis assumes that similar sections reference similar sections, an *in-in* analysis would mandate a reverse flow of similarity such that similar sections are referenced by similar sections. An even more indirect *in-out* analysis infers that, tangentially, similar sections reference to and are referenced by similar sections. Comparisons of “in references” and “out references” raise some interesting issues on the true semantics of explicit and implicit referencing, which is beyond the scope of this dissertation. To demonstrate the concept of reference distribution without losing focus

on details, we will stay with the definition of Φ_{rd} in (3.28) where only explicit references are considered.

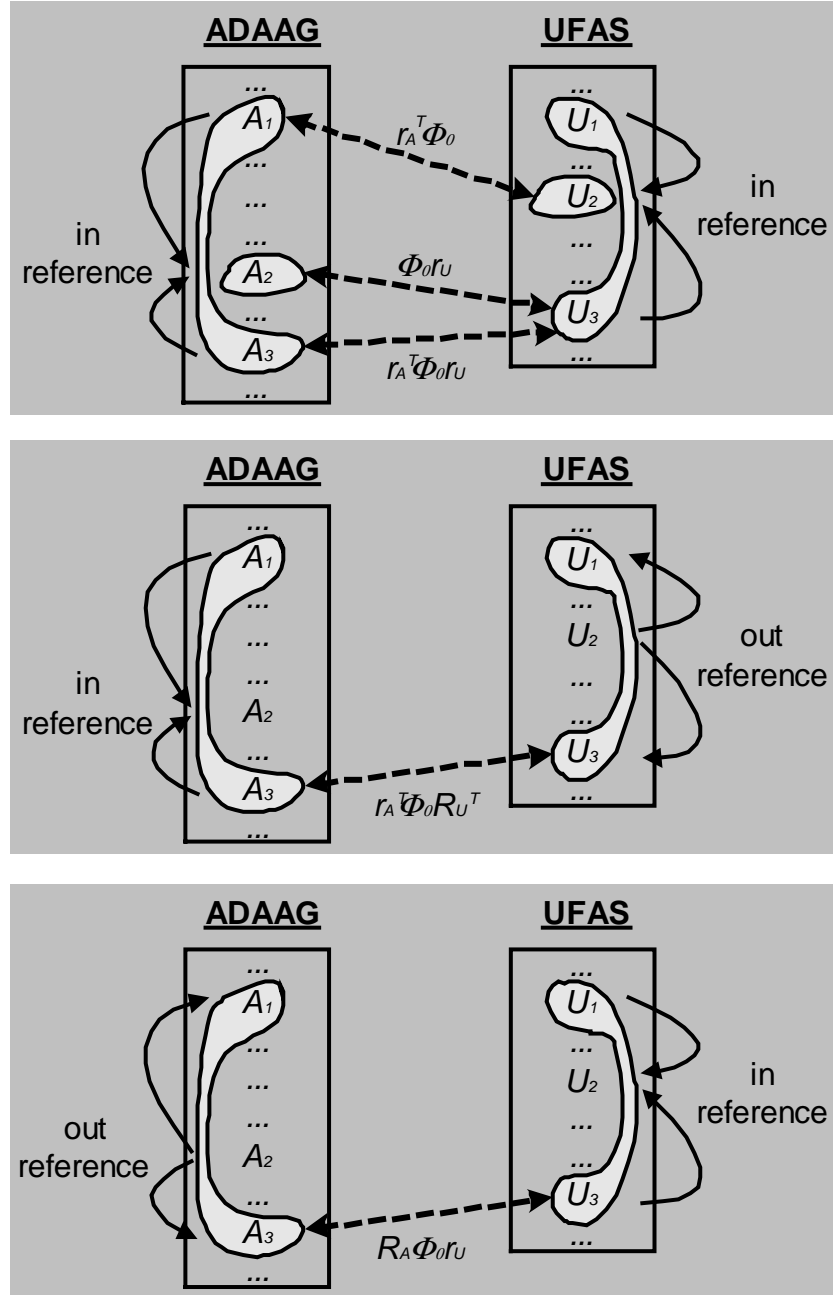


Figure 3.17: Illustrations of *In-In* and *In-Out* Reference Comparisons

Finally, with score refinements properly defined, our analysis is complete by combining Φ_0 , Φ_{s-psc} , $\Phi_{psc-psc}$, Φ_{s-ref} and $\Phi_{ref-ref}$ to form a final similarity score Φ_{final} . We define Φ_{final} as a linear combination using the α weights previously defined, and Equation (3.29) shows the result.

$$\Phi_{final} = \alpha_0 \Phi_0 + \alpha_{s-psc} \Phi_{s-psc} + \alpha_{psc-psc} \Phi_{psc-psc} + \alpha_{s-ref} \Phi_{s-ref} + \alpha_{ref-ref} \Phi_{ref-ref} \quad (3.29)$$

$$\text{where } \Phi_{psc-psc} = N_A \Phi_0 N_U^T$$

$$\Phi_{s-psc} = 1/2 (\Phi_0 N_U^T + N_A \Phi_0)$$

$$\Phi_{ref-ref} = R_A \Phi_0 R_U^T$$

$$\Phi_{s-ref} = 1/2 (\Phi_0 R_U^T + R_A \Phi_0)$$

$$\alpha_0 > \alpha_{s-psc} > \alpha_{psc-psc} > 0$$

$$\alpha_0 > \alpha_{s-ref} > \alpha_{ref-ref} > 0$$

$$\alpha_0 + \alpha_{s-psc} + \alpha_{psc-psc} + \alpha_{s-ref} + \alpha_{ref-ref} = 1$$

Chapter 4 will discuss the evaluations performed on results obtained using Φ_0 , Φ_{ni} , Φ_{rd} and Φ_{final} separately, as well as results of different tunings of α parameters.

3.5 Summary

A relatedness analysis, defined to identify materials that are alike in substance or connected by reason of a discoverable relation, is introduced in this chapter. A brief overview of related work starts the discussion. We distinguish techniques from different fields, such as data mining, information retrieval and knowledge discovery in databases. Popular techniques for document comparisons are reviewed, such as the Vector model, different term weighting approaches and the use of SVD for dimension reduction in LSI, and PLSA. Link analysis has gained new momentum due to the proliferation of the

Internet, with which legal documents share some of the referencing property. Thus, we explore recent studies in hyperlink topology that are based on academic citation analysis. Techniques such as Google's PageRank algorithm, HITS and the authority and hub interpenetration of hyperlinked documents are briefly examined.

Prior to defining the techniques of a relatedness analysis for regulations, an examination on the semantics of similarity and relatedness is first given. The basis of comparisons and the similarity measure are discussed. We define the goal, the unit and the operators of comparisons for a relatedness analysis based on a similarity score between 0 and 1. The computation of a similarity score, which includes a base score computation and several score refinements, are then introduced as shown in Figure 3.2.

The base score is a linear combination of scores from each feature matching. This 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 [91], 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.

Some features, such as the list of drinking water contaminants in environmental regulations, 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. Some domain-specific features are supplemented with feature dependency information defined by knowledge experts, who do not necessarily agree with a Boolean definition. Therefore, we propose a new vector space transformation based on the Vector model to accommodate non-Boolean matching. A matrix

representation is presented for the transformation. The formulation is shown to give accurate results on boundary cases, such as a complete axis independence and a dimension reduction introduced by collapsed axes.

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. In other words, similarities between the immediate neighbors imply similarity between the interested pair, which defines the basis of neighbor inclusion. A matrix representation is developed, where a neighbor structure matrix is defined to codify the neighbor relationship in a regulation tree.

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. 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.

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 . 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 chapter gives an overview of common evaluation models, and results obtained using different combinations of parameters are analyzed. Potential application of the developed analysis will be demonstrated as well.

Chapter 4

Performance Evaluation Models, Results and Applications

It is a challenging task to evaluate the *true* similarity between two documents. Human judgment is unavoidable; be it a match or non-match assessment between documents, a similarity score assignment, or a document ranking survey, different levels of subjectivity are introduced in the evaluation process. The seemingly simple question “how related are the two documents” indeed involves a lot of thinking. For instance, contextual information needs to be investigated since documents might not be self-contained. Terminological differences need to be anatomized with technical and domain-specific terms defined. Arguments on semantic interpretations need to be resolved for unintended or intended ambiguities in the documents. Obviously, similarity evaluation can be as complicated as one desires.

In this chapter, a performance evaluation model, results of comparisons and an example of system application are introduced. This chapter is divided into four parts: an overview of related work, a performance evaluation model using human input as the *true* similarity, a classification of results according to comparisons of regulations from different sources, and a system application on the domain of electronic-rulemaking. First, in Section 4.1, a

brief overview of existing performance evaluation models, such as precision and recall, is given. Section 4.2 explains the use of a user survey to gather human input as the *true* similarity between two regulations to evaluate our system performance. Section 4.2.1 gives a brief discussion on the format of the survey, and explains why rankings are selected as the appropriate user input instead of a direct similarity score assignment. Results obtained from our model is compared with that of Latent Semantic Indexing, which is introduced in Section 3.1.2, and the root mean square error between the *true* ranking and machine-generated ranking is tabulated in Section 4.2.2. To demonstrate the strength and weaknesses of different features as well as structural comparisons incorporated in our system, we include the results and observations based on different combinations of system parameters.

Different regulations are then paired up according to their sources, and the results of comparisons among different pairs of regulations are given in Section 4.3. General observations, such as the effect of different features on the results, are discussed in Section 4.3.1. Sections 4.3.2 to 4.3.6 document the comparisons in five different groups of regulations organized according to different sources. We will discuss, compare and contrast the average similarity of each group, as well as give illustrative examples of the usage of different features and structural analyses.

To demonstrate potential application of our system on domains other than regulation comparisons, we exploit the electronic-rulemaking process in Section 4.4. A short discussion is given on the observed impact of e-rulemaking on the efficiency of government agencies as well as rule makers. In essence, e-rulemaking creates a huge amount of data, i.e., public comments, effortlessly due to the “e” element in this process. This translates into a significant increase of workload for agencies, as the drafted rules need to be analyzed, compared and revised based on the generated public comments. We perform a relatedness analysis on a drafted regulation and its associated public comments, and several interesting examples are shown in this section. A summary follows in Section 4.5 to conclude the chapter.

4.1 Related Work

For retrieval evaluation, precision and recall are the most common metric when a benchmark of *relevant documents* is available. With a desired group of documents identified, it is easy to define the recall and precision measures. Recall is the fraction of relevant documents out of the retrieved documents, while precision is the fraction of correctly retrieved documents out of all of the relevant documents. It is best illustrated with Figure 4.1, where different document sets are defined. We then have

$$\begin{aligned}\text{Recall} &= \frac{|RV|}{|R|} \\ \text{Precision} &= \frac{|RV|}{|V|}\end{aligned}\tag{4.1}$$

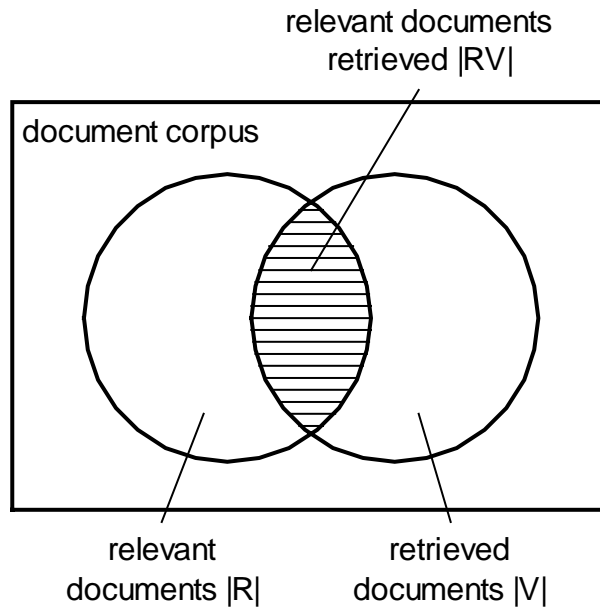


Figure 4.1: Precision and Recall

The problem with a precision and recall measure is that such benchmark does not always exist. A document corpus is not always fully examined with the relevant set of documents identified. For instance, in a legal corpus, it will be very difficult for any individual to thoroughly understand each document in the corpus and determine whether a document is relevant or not to a given query. In addition, what defines relevance could be subjective as well. Section 3.2 provides a further investigation into the meaning and definition of similarity and relatedness.

Without a benchmark, performance evaluation unavoidably involves human judgment in creating a benchmark for specific tasks. For instance, automated concept matching between different ontologies is compared with a human-generated matching [72]. Experts read through the list of automatically extracted matches; the addition and deletion of pairs form the basis of a precision and recall computation. Some work involves experts scoring matches, such as in [16] where 21 people are asked to score the relevance of automatically extracted synonyms on a scale of 0 to 10.

4.2 Comparisons to a Traditional Retrieval Model Using a User Survey

As discussed in related work in Section 4.1, there is no good metric to evaluate the performance of a textual similarity comparison system. Precision and recall, as suggested in Section 4.1, are only relevant when a targeted retrieval group is identified. In addition, subjective human judgment is inevitable in deciding what material is related or relevant to what. For instance, we briefly looked into what constitutes *similarity* and *relatedness* in Section 3.2, where it is unclear how this judging of similarity is performed even in a confined and rule-driven domain such as the law. The subjectivity of similarity judgment can be concluded in [89]: “similarity is not static; it can depend on one’s viewpoint and desired outcome.”

As unavoidable as described above, we seek human input to gauge the performance of our relatedness comparison system. We treat human input as the “correct” answer to comparisons between provisions from different sources. The “correct” answer could be subjective and incomplete, as different users certainly have different interpretations to different scenarios. Moreover, it is impossible for users to read through the entire corpus of regulations, thus creating an incomplete picture of understanding. For example, a simple task of understanding a single provision could result in an endless loop of contextual and referential lookups. Therefore, as will be introduced later in the design of a user survey, we try to provide a certain amount of contextual and referential information when available.

To assess the performance of our system with human input as the “correct answer,” we take a traditional retrieval model as the benchmark. Traditional techniques, such as the Vector model [91], simply compare the frequency counts of index terms between documents. A popular alternative is Latent Semantic Indexing (LSI) [34] which is introduced in Section 3.1.2. LSI uses Singular Value Decomposition (SVD) to reduce the dimension of term space into concept space by keeping only the s largest singular values; the claim is that synonyms that represent the same concept are mapped onto the same concept axis. In this work, we take LSI as the benchmark to compare with our system using a user survey as the “correct answer” of related provision retrievals. The value s is normally in the hundreds range, and we use $s = 300$ where the rest is zeroed out as noise.

4.2.1 User Survey and the Metric

Since it is impossible for our survey subjects to read through the entire corpus of regulations, ten sections from the ADAAG [1] and the UFAS [101] are randomly chosen as our point of comparison. To facilitate understanding, contexts are given to our subjects for provisions that are deep in their regulation tree. For example, upon reading Section 4.34.2 titled “clear floor space” of the ADAAG, it is unclear to the reader what context of “clear floor space” it is with respect to, and therefore the title of its parent,

namely “automated teller machines” from Section 4.34 of the ADAAG, is given as well. By providing only the title of contextual information, we aim to balance the volume of information given to the users with the focus of provisions in comparison. The task for users is to assign a ranking between each pair of provisions from ADAAG and UFAS based on their relatedness. This is because the similarity score is a relative measure, where a score of 0.5 is meaningless on its own without comparisons to other scores produced using the same scoring scheme. In addition, as we have discussed above the subjectivity of a user survey, different people would have different interpretation on a similarity score of 0.5 as well. Thus, only the rankings produced by different scoring schemes, such as our system, LSI or individual user, are compared.

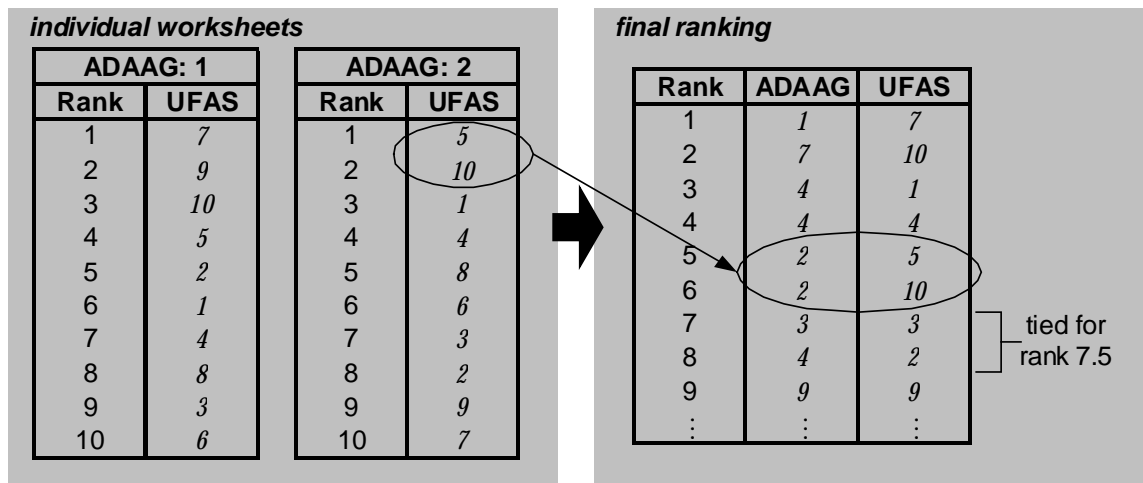


Figure 4.2: User Survey

A worksheet is included in the survey where users can first rank each section in the ADAAG from 1 to 10 in decreasing similarity with the ten selected sections from the UFAS. Users are then asked to compile the separate rankings into a final 1 to 100 pairwise rank, such as Section 3.4 from the ADAAG matching Section 4.2 from the UFAS as final rank 67. Ties are broken by taking the average rank as the true rank, for example, two sections tied for rank 67 is assigned a true rank of 67.5 (essentially, one section occupies rank 67 and one occupies 68, which results in $(67+68)/2$). Figure 4.2 shows an

example user survey. Ten surveys are collected and the results are analyzed, where the average ranking obtained from the ten surveys is regarded as the “correct” ranking.

4.2.2 Root Mean Square Error (RMSE)

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. A 100-entry ranking vector is formed per scoring scheme by representing each pair of sections as one entry. The i^{th} element of the ranking vector is the ranking of pair i . To obtain the difference between the “correct” ranking and the machine predicted ones, we compute the Root Mean Square Error (*RMSE*) between the ranking vectors as shown in Equation (4.2):

$$\begin{aligned} RMSE &= \sqrt{\frac{\text{Residual Sum of Squares (RSS)}}{\text{degrees of freedom}}} \\ &= \sqrt{\frac{|\vec{r}_h - \vec{r}_m|^2}{\text{length of } \vec{r}_h}} = \sqrt{\frac{(\vec{r}_{h1} - \vec{r}_{m1})^2 + \dots + (\vec{r}_{hN} - \vec{r}_{mN})^2}{N}} \end{aligned} \quad (4.2)$$

where \vec{r}_h and \vec{r}_m are the human-generated and machine-predicted ranking vectors respectively. In essence, the RMSE is the root of the Residual Sum of Squares (RSS) normalized according to the number of observations, in our case, $N = 100$. Several sets of parameters are experimented for our system, and the corresponding root mean square errors are computed and tabulated in Table 4.1 below.

Table 4.1: RMSE of Different Combinations of β and α Parameters

					$\beta_{concept}$						
					1	0	0	0.5	0	0.5	0.33
					β_{index}						
					0	1	0	0.5	0.5	0	0.33
					$\beta_{measurement}$						
					0	0	1	0	0.5	0.5	0.33
α_0	α_{s-psc}	$\alpha_{psc-psc}$	α_{s-ref}	$\alpha_{ref-ref}$							
1	0	0	0	0	24.9	16.0	12.0	24.5	15.6	24.9	24.4
0.8	0.15	0.05	0	0	25.7	29.1	16.9	25.7	28.1	25.6	25.7
0.8	0	0	0.15	0.05	25.6	18.0	12.0	25.2	17.5	25.6	25.1
0.8	0.075	0.025	0.075	0.025	25.8	28.4	13.7	25.7	27.1	25.6	25.5
Average					22.9						
LSI					27.4						

As shown in Table 4.1, the resulting errors are tabulated in the middle section where different combinations of β and α parameters are used. The average of the results from these different sets of parameters is shown last in the table, along with the error rate obtained using LSI. The parameters we experimented with in our system are the weighting coefficients of different features, namely $\beta_{concept}$, β_{index} and $\beta_{measurement}$, and the weights of different score computation techniques, such as α_0 , α_{s-psc} , $\alpha_{psc-psc}$, α_{s-ref} and $\alpha_{ref-ref}$.

As introduced in Chapter 2, three different kinds of features are extracted in the domain of accessibility, namely concepts, author-prescribed indices and measurements. The first three columns represent results obtained using only one feature comparison, such as concepts alone. The next three columns represent different mixtures of two feature comparisons, such as concepts and measurements each contributing to 50% of the base score. The last column shows an equally weighted linear combination of all three features that are found in accessibility regulations. The first row represent a base score computation without any score refinements, whereas the second and the third rows are a combination of the base score and either neighbor inclusion or reference distribution.

The last row is a mixture of the base score and both score refinements. As suggested in Equation (3.29), we have $\alpha_0 > \alpha_{s-psc} > \alpha_{psc-psc} > 0$ and $\alpha_0 > \alpha_{s-ref} > \alpha_{ref-ref} > 0$.

4.2.2.1 General Observations

Overall, our system outperforms the traditional bag-of-word model LSI, where the average root mean square error is 22.9 and 27.4 respectively. Majority of the combinations of parameters we experimented with produces better results ranging from slightly lowered to significantly reduced errors. The smallest error ($e = 12.0$) is obtained using the measurement feature, which further reinforces the importance of domain knowledge. Structural score refinements do not seem to affect the results in any noticeable trend, which could be attributed to the fact that the ten pairs of randomly selected sections are not particularly heavily referenced. In addition, survey subjects are not given with complete contextual and referential information, which could potentially result in a content-biased “correct” answer.

Comparing the runtime of both techniques, LSI uses singular value decomposition which is known to be computationally intensive. The comparisons between the ADAAG (701 sections) and the UFAS (549 sections) result in 1250 number of sections in total. The corpus for this analysis, consisting of the ADAAG and the UFAS only, is indeed quite small. With stopword elimination, there are only 1818 unique terms in this reduced corpus. Using a relatively small $s = 300$, the number of the largest singular values to remain, the runtime of LSI in MATLAB is on the order of minutes. As our relatedness analysis does not involve heavy computation except several sparse matrix multiplications, the runtime in MATLAB is on the order of seconds. The gain in performance is expected, since our pre-extracted and pre-stored features, such as concepts, essentially eliminate the need for a dynamic dimension reduction of term space into concept space.

4.2.2.2 Results With and Without Domain Knowledge

The first row of results tabulated in Table 4.1 is obtained based on the use of concept as the only feature in comparison, which represents a relatedness analysis without domain knowledge. The second and the third rows of results are based on the use of author-prescribed indices and measurements, which demonstrate the use of domain knowledge in a comparative analysis. Clearly, the use of domain knowledge results in smaller RMSE than analysis performed without domain knowledge. In particular, the use of measurements generates significantly smaller errors than any other combinations of features.

Since this survey is only conducted using accessibility regulations, we cannot 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, we do believe that domain knowledge has its values in locating related provisions, as is apparent in the domain of accessibility based on the survey. In the next section, drinking water standards are also compared using implemented domain knowledge such as effective dates and drinking water contaminants. Examples will be given to illustrate the importance of domain knowledge, such as ontologies, in the domain of drinking water regulations.

4.3 Comparisons Among Different Sources of Regulations

With a rich corpus of regulations from different sources, there are a number of interesting comparisons one can perform. As detailed in 2.2.1, our data comes from the Federal government, State government, private non-profit organization as well as European agencies in the domains of accessibility and drinking water control. A fire code is

included as well to show a cross-domain comparison. We divide the comparisons into five groups according to the data source as follows:

- Group 1: ADAAG vs. UFAS. The Americans with Disabilities Act Accessibility Guidelines (ADAAG) [1] and the Uniform Federal Accessibility Standards (UFAS) [101], both published by the US Access Board, are compared here to show the similarity between Federal regulations in the domain of accessibility. They are expected to be very similar as they are enacted by the same agency.
- Group 2: UFAS vs. IBC11. The UFAS and Chapter 11 from the International Building Code (IBC) [63] are compared; both of which are accessibility standards. We do not compare both the ADAAG and the UFAS with the IBC, since the ADAAG is expected to be similar to the UFAS and we do not anticipate new discoveries by adding the ADAAG into the analysis here. The UFAS serves as a representative of Federal accessibility regulations in this case. We will use IBC11 to represent Chapter 11 from the IBC for convenience.
- Group 3: UFAS vs. BS8300/STS. The UFAS is compared with both the British Standard BS 8300 [21] and Part S of the Scottish Technical Standards [97] to show the similarities and dissimilarities between US and European requirements on disabled access. BS8300 and STS will be used to represent the British and Scottish Standards.
- Group 4: 40CFRdw vs. 22CCRdw. Parts 141 to 143 from the US Code of Federal Regulations Title 40 (40 CFR) [28] are compared with Division 4 of the California Code of Regulations Title 22 (22 CCR) [26]. They represent drinking water standards enforced by the Environmental Protection Agency and California state government respectively. We will symbolize the selected parts of drinking water regulations using 40CFRdw and 22CCRdw.
- Group 5: 40CFRdw vs. IBC9. To contrast with same domain comparisons such as within accessibility or drinking water standards, we compare regulations from

two different domains. Parts 141 to 143 from the 40 CFR are compared with Chapter 9 of the IBC, titled “Fire Protection Systems,” to show the dissimilarity between the two domains. Analogous to Chapter 11 of IBC, we will represent the fire code as IBC9.

Table 4.2 shows the average similarity scores among the five groups of comparisons obtained using different combinations of features. The columns represent scores obtained from different features with the last column denoting a combination of all features. Each row represents one group of comparison as described above. A dash symbol shows that a specific feature type is not present among a specific group of comparison; for example, effective dates and drinking water contaminants (dwc) are only valid among environmental regulations. The average similarity scores tabulated here represent final scores combining Φ_0 , Φ_{ni} and Φ_{rd} using $\alpha_0 = 0.8$, $\alpha_{s-psc} = \alpha_{s-ref} = 0.075$ and $\alpha_{psc-psc} = \alpha_{ref-ref} = 0.025$ as shown in the last row of results with varying β parameters in Table 4.1.

Section 4.3.1 first gives a general overview of the results from Table 4.2. Observations based on the results from different groups are documented, as well as the effect of different features among different groups of comparisons. Sections 4.3.2 to 4.3.6 will look into details of each group of comparisons. Illustrative examples will be provided in each section to demonstrate the strength and weaknesses of different features, neighbor inclusions and reference distributions, along with interesting observations in specific groups of comparisons.

Table 4.2: Average Similarity Scores Among Comparisons Using Different Feature Sets

Groups of Comparisons	Concept	Index	Measurement	Effective date	Dwc	All
Group 1: ADAAG Vs. UFAS	0.0573	0.0805	0.0024	-	-	0.0467
Group 2: UFAS Vs. IBC11	0.0663	0.141	0.0002	-	-	0.0691
Group 3: UFAS Vs. BS8300/STS	0.0430/ 0.0337	0.0920/ 0.0606	0.0001/ 0.0001	-	-	0.0451/ 0.0314
Group 4: 40CFRdw Vs. 22CCRdw	0.0095	-	0.0001	0.0002	0.0066	0.0041
Group 5: 40CFRdw Vs. IBC9	1.7×10^{-6}	0	0	0	0	3.4×10^{-7}

4.3.1 General Observations

In the domain of accessibility, the average similarity scores are close to one another according to Table 4.2. Similarities within Group 2 seem to be the highest; however, it is possibly due to the bias in index term matching which will be discussed in Section 4.3.3. Similarities among Group 3 is relatively smaller than Groups 1 and 2. As will be explained in Section 4.3.4, terminological differences between American and European accessibility codes could be one of the reasons. It is clear that Group 5 is significantly less related than other groups, which is expected as the documents in comparisons are from two different domains.

The average similarity scores among drinking water regulations, as shown in Group 4 in Table 4.2, are relatively smaller than the average similarity scores among accessibility regulations. The natural question to ask is why would drinking water standards from the CFR and CCR be less related to one another than accessibility regulations. The answer is twofold; first of all, environmental regulations are much longer than accessibility codes. There are over 2600 provisions in each of the drinking water subparts in 40 CFR and 22 CCR. Accessibility regulations range from a maximum of about 700 provisions in the

ADAAG to 50 provisions in the STS. On average, the similarity between 2600×2600 pair-wise comparisons is much lower than that of 700×50 comparisons.

This observation leads to the second interpretation to the result. Part of the reasons why drinking water regulations are much more voluminous than accessibility regulations is its diversity of coverage. Drinking water standards cover a lot of topics such as the national primary drinking water, national secondary drinking water, consumer confidence reports and so on. In this regard, accessibility regulations are more focused and more similar to a topic within drinking water regulations. This explains why accessibility regulations are more similar to one another than drinking water standards. Higher similarity scores are expected for comparisons between topics in drinking water regulations.

Further investigations into the resulting ranking of scores reveal that the highest similarity scores among Group 1 are much higher than that of other groups. The final similarity scores of the top ranked provision pairs in Groups 2 to 4 are roughly 0.6, whereas the scores of the top ranked pairs in Group 1 range from 0.88 to 0.98 depending on different α and β values. Indeed, identical pairs of provisions are observed in Group 1 which will be discussed in Section 4.3.2. As will be introduced in Section 4.3.3, Group 2 shows much fewer pairs of *almost* identical provisions compared to Group 1, which is reasonable as IBC11 is not prepared by the same Federal agency as the UFAS and ADAAG. Several *almost* identical provisions are also identified in Group 4 as discussed in Section 4.3.5.

Besides examining the top ranked pairs of provisions, we also consider the mid-ranked provisions of different groups. For example, mid-ranked pairs from Group 2 are somewhat related to one another, and an example is given in Section 4.3.3. Relatively speaking, the mid-ranked results of Group 2 are not as similar to one another compared to the mid-ranked provisions found in Groups 1 and 4. Besides analyzing the average similarity scores from Table 4.2, investigations of the mid-ranked provisions provide a different perspective on judging similarities within a group.

Comparing different features, concepts always play a fairly important role across different groups of comparisons. This is understandable since terms form the basis of body text in regulations, and thus appear much more often than non term-based features such as measurements. Other term-based features, such as indices and drinking water contaminants, also result in average similarity scores bigger than those obtained using other features such as measurements.

Examining the column of scores obtained using a measurement comparison, Group 1 seem to share more measurement features than other groups, as the average similarity is order of magnitude bigger than the rest of the groups. In the domain of accessibility, Groups 2 and 3 both show a relatively low similarity score based on measurement comparisons. The result of Group 3 is easy to understand, as the measurements prescribed are indeed quite different between US and European codes. Potential future work in this area could be a unit conversion system that handles differences between the International System of Units (SI), such as meters, and the US Customary System (USCS) units, such as feet. Nevertheless, a simple unit conversion system will not improve the result of Group 2. The relatively low similarity score based on measurement comparisons between the UFAS and IBC9 is possibly due to the fact that the IBC9 is more performance based than the UFAS. There are only a few measurements prescribed in the IBC9, which is short and concise (120 sections) compared to the UFAS (554 sections). Measurements become a less useful indicator of similarity in this case. In fact, it will be similar to a Group 5 analysis, i.e., a cross-domain comparison, for comparisons between an entirely performance-based regulation and a strictly prescriptive code.

4.3.2 Group 1 Comparison: ADAAG Vs. UFAS

As shown in Table 4.2, the two Federal regulations are relatively similar to each other with respect to other groups of comparisons. Although the average similarity score of Group 2 appears to be higher than that of Group 1, there is a slight bias in the author-prescribed index comparisons for Group 2 which will be discussed in Section 4.3.3.

Term-based features, such as concepts and indices, result in average similarity scores bigger than those obtained using measurement features. As explained in Section 4.3.1, this is possibly due to the fact that term-based features form the majority of the body text of regulations.

The comparison results show that the first 50 top ranked pairs in Group 1 are almost identical provisions; some even share the same section ID. It is not uncommon that two regulations share provisions that are entirely the same, especially when the two regulations are published by the same agency as in this case. For instance, in the “Introduction” section of the UFAS, several adoptions are listed: “GSA¹⁸ adopted the UFAS in 41 CFR 101-19.6... HUD¹⁹ adopted the UFAS in 24 CFR part 40... USPS²⁰ adopted the UFAS in Handbook RE-4... DoD²¹ adopted the UFAS by revising Chapter 18 of DoD 4270.1-M.”

To justify for the proposed score refinements, we compare results obtained using the base score with results from neighbor inclusion and reference distribution. Two interesting examples are shown in this section, with more to come in subsequent sections. The first example shown in Figure 4.3 illustrates the use of neighbor inclusion, where we compare results of f_0 with f_{s-psc} and $f_{psc-psc}$, and some improvement is observed. For instance, Section 4.1.6(3)(d) in the ADAAG is concerned with doors, while Section 4.14.1 in the UFAS deals with entrances. As expected, a pure concept match could not identify the relatedness between door and entrance, thus $f_0 = 0$. With non-zero f_{s-psc} and $f_{psc-psc}$, 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, are identified indirectly through neighbor inclusions.

¹⁸ GSA stands for General Services Administration.

¹⁹ HUD stands for Department of Housing and Urban Development.

²⁰ USPS stands for United States Postal Service.

²¹ DoD stands for Department of Defense.

<p>ADA Accessibility Guidelines</p> <p><u>4.1.6(3)(d) Doors</u></p> <p>(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.</p> <p>Uniform Federal Accessibility Standards</p> <p><u>4.14.1 Minimum Number</u></p> <p>4.14 Entrances</p> <p>4.14.1 Minimum Number</p> <p>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.</p>

Figure 4.3: Related Provisions Identified Through Neighbor Inclusion

The second example shows the importance of an *s-ref* comparison in reference distribution. As shown in Figure 4.4, both Section 4.13.5 in the ADAAG and Section 4.3.3 in the UFAS discuss about the minimum clear width of a door, with different focuses. The base score is relatively low (0.20), while f_{s-ref} is considerably higher (0.88). In fact, Section 4.13.5 in the ADAAG references another section in the ADAAG that is identical to Section 4.3.3 in the UFAS, and vice versa. This explains why an *s-ref* comparison is needed in addition to the traditional out reference comparison (*ref-ref*). For instance, in this case, a *ref-ref* comparison does not identify much similarity between the out references from the two sections in comparison.

ADA Accessibility Guidelines**4.13.5 Clear Width**

Doorways shall have a minimum clear opening of 32 in (815 mm) with the door open 90 degrees, measured between the face of the door and the opposite stop (see Fig. 24(a), (b), (c), and (d)). Openings more than 24 in (610 mm) in depth shall comply with 4.2.1 and 4.3.3 (see Fig. 24(e)).

EXCEPTION: Doors not requiring full user passage, such as shallow closets, may have the clear opening reduced to 20 in (510 mm) minimum.

→ 4.2.1 Wheelchair Passage Width

The minimum clear width for single wheelchair passage shall be 32 in (815 mm) at a point and 36 in (915 mm) continuously (see Fig. 1 and 24(e)).

→ 4.3.3 Width

The minimum clear width of an accessible route shall be 36 in (915 mm) except at doors (see 4.13.5 and 4.13.6). If a person in a wheelchair must make a turn around an obstruction, the minimum clear width of the accessible route shall be as shown in Fig. 7(a) and (b).

Uniform Federal Accessibility Standards**4.3.3 Width**

The minimum clear width of an accessible route shall be 36 in (915 mm) except at doors (see 4.13.5). If a person in a wheelchair must make a turn around an obstruction, the minimum clear width of the accessible route shall be as shown in Fig. 7.

→ 4.13.5 Clear Width

Doorways shall have a minimum clear opening of 32 in (815 mm) with the door open 90 degrees, measured between the face of the door and the stop (see Fig. 24(a), (b), (c), and (d)). Openings more than 24 in (610 mm) in depth shall comply with 4.2.1 and 4.3.3 (see Fig. 24(e)).

EXCEPTION: Doors not requiring full user passage, such as shallow closets, may have the clear opening reduced to 20 in (510 mm) minimum.

Figure 4.4: Related Provisions Identified Through Reference Distribution

4.3.3 Group 2 Comparison: UFAS Vs. IBC11

Chapter 11 of the International Building Code, titled “Accessibility,” is a very focused and concise document with only 120 sections, which is relatively small compared to other regulations in this area. As discussed in the Section 4.3.1, smaller and more topic-focused regulations tend to be more related to other regulations in the same domain than a longer and more diversified code. However, it is still quite surprising to see that a Federal regulation is, on average, more similar to this private organization mandated regulation than another Federal code on the topic of accessibility. The result is slightly biased, since comparisons based on author-prescribed indices return a much higher similarity score than any other features. This is because the list of indices incorporated in our corpus indeed come from the back of the IBC. Consequently, there are more frequent usages of the index terms in the IBC than other regulations, which lead to the bias in scores.

Identical provisions are observed in Groups 1, and we have a similar but slightly different observation in Group 2 as well. *Almost* identical provisions are ranked on top in comparisons between the UFAS and IBC11, where one provision is the paraphrase of another provision as shown in Figure 4.5. A few pairs of almost identical provisions similar to this top the rankings, due to the shared concepts and author-prescribed indices.

So far, we have shown several examples of closely related provisions identified by our system across different groups of comparisons. In order to paint a complete picture of regulatory comparisons, we will now show an example of the non top-ranked results. Apart from the few almost identical pairs of provisions on top of the rankings, we observe that there are a lot of mid-ranked provisions in Group 2 that are somewhat related to one another. For instance, Figure 4.6 below shows a pair of provisions from the UFAS and IBC11 with final score ranked at around 200. They are somewhat related, as they both examine on the accessibility requirements for storage facilities. However, their focus is quite different and they are embedded in quite different contexts in their

own regulation tree, which explains why they are in the mid-rank section even with all feature matching, neighbor inclusion and reference distribution.

<p>Uniform Federal Accessibility Standards <u>4.3.2(1) [No Title; under Accessibility Route Location]</u> At least one accessible route within the boundary of the site shall be provided from public transportation stops, accessible parking, and accessible passenger loading zones, and public streets or sidewalks to the accessible building entrance they serve.</p> <p>International Building Code Chapter 11 <u>1104.1 Site Arrival Points</u> Accessible routes within the site shall be provided from public transportation stops, accessible parking and accessible passenger loading zones, and public streets or sidewalks to the accessible building entrance served.</p>

Figure 4.5: Almost Identical Provisions Prescribed by the UFAS and the IBC

<p>Uniform Federal Accessibility Standards <u>4.1.2(11) [No Title; under Accessible Buildings: New Construction]</u> If storage facilities such as cabinets, shelves, closets, and drawers are provided in accessible spaces, at least one of each type provided shall contain storage space complying with 4.25. Additional storage may be provided outside of the dimensions shown in Fig 38.</p> <p>International Building Code Chapter 11 <u>1107.6.1 Dispersion</u> Accessible individual self-service storage spaces shall be dispersed throughout the various classes of spaces provided. Where more classes of spaces are provided than the number of required accessible spaces, the number of accessible spaces shall not be required to exceed that required by Table 1107.6. Accessible spaces are permitted to be dispersed in a single building of a multibuilding facility.</p>

Figure 4.6: Mid-Ranked Related Provisions from the UFAS and the IBC

4.3.4 Group 3 Comparison: UFAS Vs. BS8300/STS

According to the average similarity scores in Table 4.2, the UFAS is slightly less related to European accessibility codes, such as the BS8300 and the STS, compared to the similarity between the UFAS and other American accessibility codes. This is partly because of spelling differences, such as the American spelling of “curb” versus the British version “kerb.” Terminological differences are observed as well, such as the chiefly British acronym “WC” appeared more than a hundred times in the BS 8300 without even defining its expanded form “water closet,” whereas the ADAAG never used the term “WC” at all. A potential future work in this area could be a dictionary check of synonyms and acronyms used across different continents. The following example will illustrate the difficulties in identifying non-trivial synonyms, where a dictionary can be of no use.

To illustrate the similarity between American and British accessibility standards, we compare the UFAS with the BS8300. Figure 4.7 and Figure 4.8 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 non-standard 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 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 4.7: Terminological Differences Between the UFAS and the BS8300

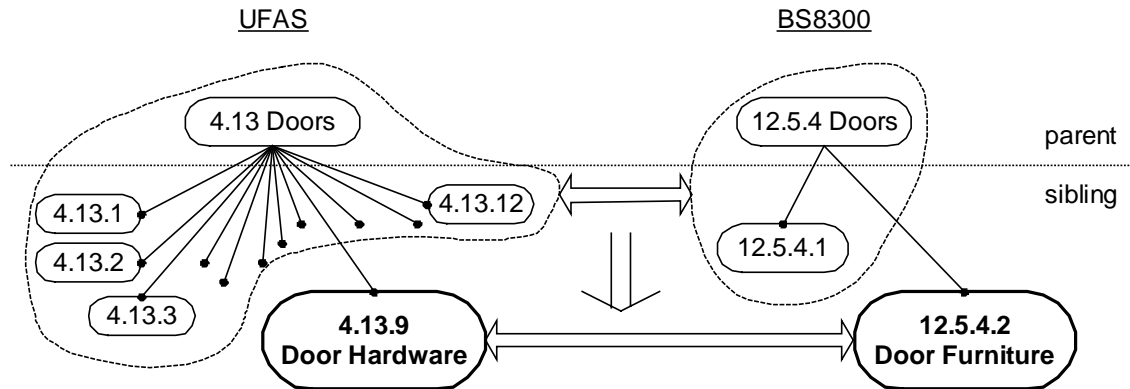


Figure 4.8: Similarities Between Neighbors Imply Similarities Between Section 4.13.9 from the UFAS and Section 12.5.4.2 from the BS8300

As shown in Table 4.2, the UFAS is also compared with the STS in addition to the BS8300. An observation based on the comparisons between the UFAS and the STS is given below in Figure 4.9, where reference distribution contributes to revealing hidden similarities between provisions. As shown in Figure 4.9, both sections from the UFAS and the STS are concerned about pedestrian ramps and stairs which are related accessible elements. However, even with neighbor inclusion, these two sections show a relatively low similarity score, which is possibly due to the fact that a pure term match does not recognize stairs and ramps as related elements. In this case, after considering reference distribution, these two provisions show a significant increase in similarity based on similar references. Again, this example shows how structural matching, such as reference distribution, is important in revealing hidden similarities which will be otherwise neglected in a traditional term match.

<p>Uniform Federal Accessibility Standards</p> <p><u>4.1.2 Accessible Buildings: New Construction</u></p> <p>(4) Stairs connecting levels that are not connected by an elevator shall comply with 4.9.</p> <p>Scottish Technical Standards</p> <p><u>3.17 Pedestrian Ramps</u></p> <p>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.</p>

Figure 4.9: Related Elements “Stairs” and “Ramp” Revealed Through Reference Distribution

4.3.5 Group 4 Comparison: 40CFRdw Vs. 22CCRdw

As shown in Table 4.2 and discussed in Section 4.3.1, Group 4 shows a smaller average similarity score than other accessibility groups, which is potentially because of the volume and diversity of coverage of drinking water regulations. Comparing different features among drinking water standards, relatedness appears to be captured by concepts and drinking water contaminants. In Section 4.3.1, we have already explained the importance of term-based features such as concepts. Drinking water contaminants are also term-based features, and the use of an ontology to help identify synonyms seems to boost the retrieval of similarity as well. Effective dates and measurements are comparatively less significant, possibly reflecting on the fact that they are non term-based features and the scoring schemes are more unsparing than that of drinking water contaminants or concepts.

Two examples are given to illustrate the similarity and dissimilarity between Federal and State drinking water regulations. The first example, shown in Figure 4.10, is a top ranked pair of related provisions on drinking water control of the chemical Barium required by the 40CFRdw and 22CCRdw. This pair of provisions is actually identical in text except the subject of governing agency changes between Environmental Protection Agency

(EPA) and California Department of Health Services (DHS). It is not uncommon that one agency directly adopts provisions issued by another agency. Indeed, in the domain of disabled access, our system identified a lot of identical provisions when comparing the ADAAG with the UFAS; however, as suggested in Section 4.3.2, this is more or less expected since both are Federal regulations issued by the Access Board.

In this example of Barium requirements, the text in the provision is actually somewhat unusual and does not seem to be written in standard regulatory language. The text appears to be a *notice* required by both the EPA and the California DHS, where the notice could potentially come from an outside source. The careful reader might also note that the EPA and the California DHS *do* have different Barium requirements – the EPA requires 2 parts per million while the California DHS sets the requirement at 1 part per million. It appears that the two agencies might have modified the notice according to their separate standards. This example also illustrates the importance of domain knowledge, where a measurement comparison would reveal that these two provisions are not identical, even though the wordings are almost the same.

Aside from adopting identical provisions between Federal and State agencies, differences are also observed between the two documents. For instance, the 40CFRdw makes use of many chemical acronyms, such as TTHM, whereas the full term “total trihalomethanes” is always spelled out in the 22CCRdw. Figure 4.11 shows a pair of provisions illustrating the case. Based on a pure concept match, the two provisions result in zero similarity. The similarity score based on a drinking water contaminant match is 0.49, due to the use of ontological information as shown in Figure 2.8 that identifies the acronym TTHM as a match to “total trihalomethanes,” as well as HAA with “haloacetic acids.” This example justifies for the incorporation of domain knowledge; without which, a user searching for TTHM or HAA will never find anything in 22CCRdw but only in 40CFRdw.

Code of Federal Regulations Title 40**141.32.e.16 Barium**

The **United States Environmental Protection Agency (EPA)** sets drinking water standards and has determined that barium is a health concern at certain levels of exposure. This inorganic chemical occurs naturally in some aquifers that serve as sources of ground water. It is also used in oil and gas drilling muds, automotive paints, bricks, tiles and jet fuels. It generally gets into drinking water after dissolving from naturally occurring minerals in the ground. This chemical may damage the heart and cardiovascular system, and is associated with high blood pressure in laboratory animals such as rats exposed to high levels during their lifetimes. In humans, **EPA** believes that effects from barium on blood pressure should not occur below 2 parts per million (ppm) in drinking water. **EPA** has set the drinking water standard for barium at 2 parts per million (ppm) to protect against the risk of these adverse health effects. Drinking water that meets the **EPA** standard is associated with little to none of this risk and is considered safe with respect to barium.

California Code of Regulations Title 22**64468.1(c) Barium**

The **California Department of Health Services (DHS)** sets drinking water standards and has determined that barium is a health concern at certain levels of exposure. This inorganic chemical occurs naturally in some aquifers that serve as sources of ground water. It is also used in oil and gas drilling muds, automotive paints, bricks, tiles and jet fuels. It generally gets into drinking water after dissolving from naturally occurring minerals in the ground. This chemical may damage the heart and cardiovascular system, and is associated with high blood pressure in laboratory animals such as rats exposed to high levels during their lifetimes. In humans, **DHS** believes that effects from barium on blood pressure should not occur below 2 parts per million (ppm) in drinking water. **DHS** has set the drinking water standard for barium at 1 part per million (ppm) to protect against the risk of these adverse health effects. Drinking water that meets the **DHS** standard is associated with little to none of this risk and is considered safe with respect to barium.

Figure 4.10: Direct Adoption of Provisions Across Federal and California State on the Topic of Drinking Water Standards

<p>Code of Federal Regulations Title 40 <u>141.132.a.2 [No Title; under Monitoring Requirements]</u> Systems may consider multiple wells drawing water from a single aquifer as one treatment plant for determining the minimum number of TTHM and HAA5 samples required, with State approval in accordance with criteria developed under §142.16(h)(5) of this chapter.</p> <p>California Code of Regulations Title 22 <u>64823(e) [No Title; under Field of Testing]</u> Field of Testing 5 consists of those methods whose purpose is to detect the presence of trace organics in the determination of drinking water quality and do not require the use of a gas chromatographic/mass spectrophotometric device and encompasses the following Subgroups: EPA method 501.1 for trihalomethanes; EPA method 501.2 for trihalomethanes; EPA method 510 for total trihalomethanes; EPA method 508 for chlorinated pesticides; EPA method 515.1 for chlorophenoxy herbicides; EPA method 502.1 for halogenated volatiles; EPA method 503.1 for aromatic volatiles; EPA method 502.2 for both halogenated and aromatic volatiles; EPA method 504 for EDB and DBCP; EPA method 505 for chlorinated pesticides and polychlorinated biphenyls; EPA method 507 for the haloacids; EPA method 531.1 for carbamates; EPA method 547 for glyphosate; EPA method 506 for adipates and phthalates; EPA method 508A for total polychlorinated biphenyls; EPA method 548 for endothall; EPA method 549 for diquat and paraquat; EPA method 550 for polycyclic aromatic hydrocarbons; EPA method 550.1 for polycyclic aromatic hydrocarbons; EPA method 551 for chlorination disinfection byproducts; EPA method 552 for haloacetic acids.</p>

Figure 4.11: Terminological Differences Between Federal and State Regulations on the Topic of Drinking Water Standards

4.3.6 Group 5 Comparison: 40CFRdw Vs. IBC9

A clear outlier in Table 4.2 is Group 5, where the results are several orders of magnitude smaller than the rest of the groups. This is expected since 40CFRdw and IBC9 are from two completely different domains, namely drinking water and fire protection standards. All of the features but concepts show a zero similarity score. Features such as drinking

water contaminants and effective dates only exist in environmental regulations, which explains why the fire code does not share any of them. Both domains contain measurements; however, they are very different kinds of measurements that are not shared between the two domains, such as “75 feet clearance” in the fire code and “2 parts per million” in drinking water standards. Concepts generate a close-to-zero similarity score, as there are still some common phrases that are shared, such as the phrase “common area” found in both domains.

<p>Code of Federal Regulations Title 40 141.85.a.1.iv.B.6 [No title; under Public Education and Supplemental Monitoring Requirements]</p> <p>Have an electrician check your wiring. If grounding wires from the electrical system are attached to your pipes, corrosion may be greater. Check with a licensed electrician or your local electrical code to determine if your wiring can be grounded elsewhere. DO NOT attempt to change the wiring yourself because improper grounding can cause electrical shock and fire hazards.</p> <p>International Building Code, Chapter 9 907.2.8.1 Fire Detection System</p> <p>System smoke detectors are not required in guestrooms provided that the single-station smoke alarms required by Section 907.2.10 are connected to the emergency electrical system and are annunciated by guestroom at a constantly attended location from which the fire alarm system is capable of being manually activated.</p>

Figure 4.12: Remotely Related Provisions Identified From a Drinking Water Regulation and a Fire Code

One example is shown below in Figure 4.12, where provisions from the two separate domains share some remote similarity. Section 141.85.a.1.iv.B.6 from the 40CFRdw is a small subsection under Section 141.85 on “public education and supplemental monitoring requirements.” This section happens to touch on the safety of *electrical systems* in public education. Section 907.2.8.1 from the IBC9 deals with fire detection systems that involves discussion of *electrical systems* as well. These two tangentially related

provisions that are top ranked among this group of cross-domain comparisons are one of the few related provisions found by our system with negligible similarity scores.

4.4 Electronic Rulemaking

Apart from the intended application on comparisons between regulatory documents and to demonstrate system scalability and extensibility, we have applied the prototype system to other domains as well, such as electronic-rulemaking (e-rulemaking). The process of e-rulemaking with participations from the public involves a non-trivial task of sorting through a massive volume of electronically submitted textual comments. Thus, our relatedness analysis system can potentially help to sort comments with respect to the drafted regulation.

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. For example, the Federal Register [45] documented a recent case where the Alcohol and Tobacco Tax and Trade

Bureau received over 14000 comments in 7 months, majority of which are emails, on a flavored malt beverages proposal. The call for public comments included the following statement:

“All comments posted on our Web site will show the name of the commenter but will not show street addresses, telephone numbers, or e-mail addresses.”

However, due to the “unusually large number of comments received,” the Bureau announced that it is difficult to remove all street addresses, telephone numbers and email addresses “in a timely manner.” Instead, concerned individuals are asked to submit a request for removal of address information as opposed to the original statement posted in the call for comments. As such, an effortless electronic comment submission process has turned into a huge data processing problem for government agencies.

In order to help screening and filtering of public comments, we applied our system on 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 [37] for the ADAAG [1], 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 results of a relatedness analysis are related pairs between the provision from the draft and individual comment. Figure 4.13 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 4.13, 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.

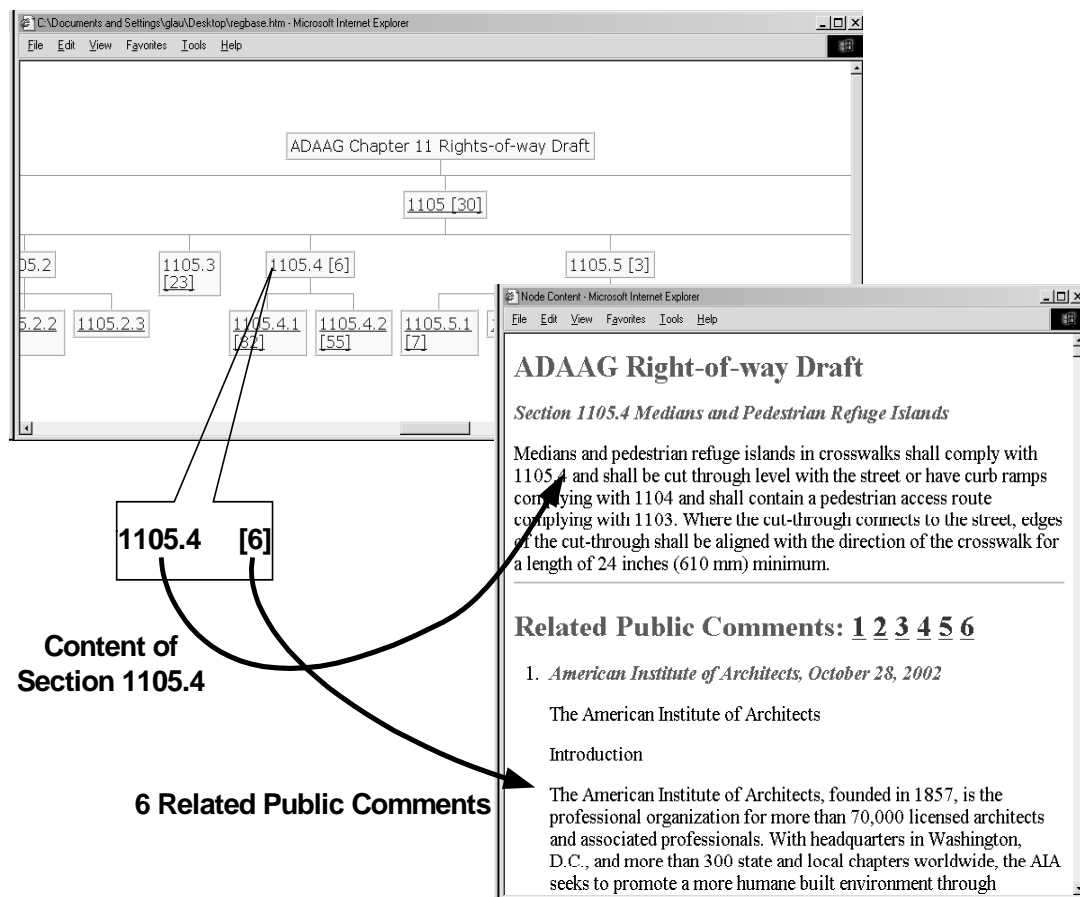


Figure 4.13: Comparisons of Drafted Rules with Public Comments in E-Rulemaking

Several interesting results are observed and presented in this section to illustrate the potential impacts as well as limitations of the use of a comparison framework on rulemaking. Figure 4.14 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.

A different type of comment screening is shown in Figure 4.15. 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
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 4.14: Related Drafted Rule and Public Comment

The pair of highly related provision and comment shown in Figure 4.16 suggests that a comparison between drafted provisions and comments is indeed the right approach. This commenter started by citing Section 1109.2 in the draft, followed by a list of suggestions and questions about Section 1109.2. Our system gathered the relatedness between Section 1109.2 and this comment through different features, such as the shared phrases. This piece of comment is a representative example of a lot of comments that are written

similarly: comments that are concerned about a single provision in the draft. Thus, a comparison between drafted provisions and comments is important to help users focus on the most related comments per provision.

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 4.15: A Piece of Public Comment Not Related to the Draft

Based on the observation made from the example shown in Figure 4.16, there seem to be room for improvement for an e-rulemaking portal. The public might find it helpful to submit comments on a per provision basis, in addition to a per draft basis. With technology available, it should be possible to develop an online submission system that allows for both types of comment submission. It saves participants time to paraphrase or cite their concerned provision. It also saves rule makers time to locate related comments either through human effort or an automated system. Comments submitted on a per draft

basis can still be analyzed and compared with the entire draft to identify any relevant provisions. On a side note, this commenter also suggested that it is important to forward the comments to the right person. An extension of this relatedness analysis framework could be developed to automatically inform any assigned personnel in charge of reviewing the provision within government agencies.

Apart from correctly identifying comments that are related to different provisions, limitations of our system are also observed. Section 1109.2 is related to another piece of comment as shown in Figure 4.17. The relatedness is revealed through the shared features between Section 1109.2 and the comment, which includes a direct quotation and revision of Section 1109.2. The identified relatedness is correct; however, suggested modifications and revisions of provisions cannot be automatically detected. In essence, our current system is able to uncover the relatedness but not the revised version of provisions embedded in the comments. To precisely locate revisions suggested in the comments, one can potentially 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.

Finally, Figure 4.18 shows a piece of public comment that is not identified as relevant to any provision in the draft. This reviewer commented on the general direction and intent of the draft, which explains why our system failed to sort this comment into any provision. Furthermore, this particular result suggests that a comparison between provisions and comments might not be enough. One could use the same analysis framework to compare comments with one another. For instance, this reviewer supported the positions of the American Council of the Blind (ACB) and the Washington Council of the Blind (WCB). While our system failed to associate this comment with any provision, comments submitted by ACB and WCB might give a clue to where this comment should belong. Essentially, clustering of comments alone could be as handy as the illustrated clustering of comments and provisions.

ADAAG Chapter 11 Rights-of-way Draft
1109.2 Parallel Parking Spaces

An access aisle at least 60 inches (1525 mm) wide shall be provided at street level the full length of the parking space. The access aisle shall connect to a pedestrian access route serving the space. The access aisle shall not encroach on the vehicular travel lane.

EXCEPTION: An access aisle is not required where the width of the sidewalk between the extension of the normal curb and boundary of the public right-of-way is less than 14 feet (4270 mm). When an access aisle is not provided, the parking space shall be located at the end of the block face.

Public Comment

Norman Baculinao, P.E., PTOE, August 26, 2002

1109.2 Parallel Parking Spaces. An access aisle at least 60 inches (1525 mm) wide shall be provided at street level the full length of the parking space. The access aisle shall connect to a pedestrian access route serving the space. The access aisle shall not encroach on the vehicular travel lane.

EXCEPTION: An access aisle is not required where the width of the sidewalk between the extension of the normal curb and boundary of the public right-of-way is less than 14 feet (4270 mm). When an access aisle is not provided, the parking space shall be located at the end of the block face.

1. This section needs to be clarified, i.e., where is the access isle located? that is, "will it be on the driver side or passenger side?"
2. The following is more of a question/concern about this requirement:

In downtown areas where parking is premium, this requirement will make it very difficult to install parallel accessible parking spaces. If I understood it correctly, access isles typically accommodate "lifts" which is usually located on the passenger side. If this is the case, then areas with adequate sidewalk width do not need access isles because "lifts" can be placed directly onto the sidewalk, EXCEPT, for left-curb side of ONE-WAY streets.

On the other hand, those that would use access isles using a wheel chair, (or for those that gets the wheel chair from a trunk or the back of the automobile), then access isle would be needed.

3. The requirement for the exception is install the parking stall at the end of the block. I am assuming the intent is to shorten the distance to the nearest access ramp. If this is the case, then can we allow a mid-block location so long as a "curb-cut" or access ramp is built either at the front or rear of the parking stall???

In the City of Pasadena, we have deferred all requests for accessible parallel parking until the guidelines is adopted so we are very anxious about final approval of this document.

I would really appreciate, if you **could forward this comments to the right individual and hopefully get a response back**. Please feel free to call me for any clarifications regarding this comments.

Sincerely,

Norman Baculinao, P.E., PTOE
 Traffic Engineering Manager
 Department of Transportation, City of Pasadena

Figure 4.16: Comment Intended for a Single Provision Only

ADAAG Chapter 11 Rights-of-way Draft
1109.2 Parallel Parking Spaces

An access aisle at least 60 inches (1525 mm) wide shall be provided at street level the full length of the parking space. The access aisle shall connect to a pedestrian access route serving the space. The access aisle shall not encroach on the vehicular travel lane.

EXCEPTION: An access aisle is not required where the width of the sidewalk between the extension of the normal curb and boundary of the public right-of-way is less than 14 feet... When an access aisle is not provided, the parking space shall be located at the end of the block face.

Public Comment

Bruce E. Taylor, P.E., October 25, 2002

Re: Request for Comments on the Draft Guidelines for Accessible Public Rights-of-Way.

The Oklahoma Department of Transportation has reviewed the proposed draft guidelines for accessible public rights of way, and subsequently reviewed the AASHTO recommendations relative to that guidance.

The Department concurs with responses conveyed in the "AASHTO Comments and Recommendations on the Draft Guidelines for Public Access", and fully supports the AASHTO efforts to address safety concerns and eliminate ambiguities within the proposed guideline language.

In addition, the Department would request that the Access Board adopt language that would allow the consideration of off-street parking as an alternative to ADA compliant on-street parallel parking. Proposed Section 1102.14, States;

Where on-street parking is provided, at least one accessible on-street parking space shall be located on each block face and shall comply with 1109.

Further, Section 1109.2, Parallel Parking Spaces, states;

An access aisle at least 60 inches (1525 mm) wide shall be provided at street level the full length of the parking space. The access aisle shall connect to a pedestrian access route serving the space. The access aisle shall not encroach on the vehicular travel lane. EXCEPTION: An access aisle is not required where the width of the sidewalk between the extension of the normal curb and boundary of the public right-of-way is less than 14 feet (4270 mm). When an access aisle is not provided, the parking space shall be located at the end of the block face.

Flexibility should be afforded the Engineer to allow off-street accessible parking, where available, in a reduced vehicular environment common to most minor streets adjoining heavily traveled thoroughfares. The Department would propose that the requirements of Section 1104.12 requiring one compliant parking space per block face, be removed, and **Section 1109.2 be revised to read;**

An access aisle at least 60 inches (1525 mm) wide shall be provided at street level the full length of the parking space. The access aisle shall connect to a pedestrian access route serving the space. The access aisle shall not encroach on the vehicular travel lane. EXCEPTION: An access aisle is not required where the width of the sidewalk between the extension of the normal curb and boundary of the public right-of-way is less than 14 feet (4270 mm). When an access aisle is not provided, the parking space shall be located at the end of the block face or on adjacent connecting streets.

The Department appreciates the opportunity to comment on the Draft Guidelines for Public Access. Should you have questions or comments, please advise.

Sincerely,
Bruce E. Taylor, P.E.
Chief Engineer
Oklahoma Department of Transportation

Figure 4.17: Suggested Revision of Provision in Comment

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

No relevant provision identified

Public Comment
Douglas L. Hildie, September 13, 2002

I am responding to a request from a fellow member of the blind community in this nation. She, and I, are members of the American Council of the Blind (ACB), its state affiliate the Washington Council of the Blind (WCB), and local chapters in our communities. **I support the positions of ACB, WCB**, and many people who are blind that, failure of national, regional, and local government to provide for the require and implement rational policies and practices resulting in the installation of tactile warnings and audible pedestrian signals at intersections would be unjustified and unjustifiable.

I am legally blind; I spent nearly twenty (20) years as a Vocational Rehabilitation Counselor serving blind clients, and it is obvious to me from my experience that the safety and welfare of blind people in general will be best served by an all inclusive approach that recognizes the needs of the many, not the needs of the few.

It is obvious, I believe, that blind people are not "all the same", any more than any group of individuals is "all the same". It is true for "sighted people", and for "blind people", that some will have varying degrees of functional ability. But, contrary to the ideological perspective being foisted upon the public at large by a foolish few in the broader community of blind persons, people who are blind cannot do everything others do with eyesight just by using a cane.

The blind community, like any community of people, is composed of some whose motives are not in the best interest of all. I hope you will "see through" the verbal shrapnel put out by a minority of blind people in this nation, and make the logical, rational, and right choice for the safety, health, and welfare of ALL blind people.

Thank you for considering my comments.

Douglas L. Hildie

Figure 4.18: Comment on the General Direction of Draft

4.5 Summary

This chapter evaluates the performance of our system compared to traditional techniques, the results of different regulation comparisons and the potential applications in detail. We first give a brief overview of performance evaluation models related to document retrieval. Precision and recall are defined, and their limitation is observed – it is difficult to develop benchmarks, i.e., a correct set of relevant documents per query, for document repositories. It is even less plausible in a legal domain, where individuals can hardly fully understand each provision and the complicated relationships between them. Several studies, which involve matching or scoring by human, are referenced here.

We then examine the development of a performance evaluation for our analysis as compared to traditional Information Retrieval techniques. We start the discussion with a reference to the difficulties in deciding whether two provisions are similar or related as cited in the previous chapter. We conclude that human judgment of similarity is inevitable in developing a metric for machine predictions, despite the fact that human input could be subjective. As a result, a user survey is devised for ranking the similarity of ten randomly chosen provisions from the ADAAG and ten from the UFAS. 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.

We choose to compare our system with Latent Semantic Indexing, as LSI claims to form concept axes instead of term axes based on a dimension reduction technique, which shares a similar goal as our concept extraction. The Root Mean Square Error (RMSE) is used to compute the ranking prediction error based on the survey results as the “correct” answer. We compute the RMSE for a LSI implementation using the 300 largest singular values, as well as different β (feature weight) and α (score refinement) parameters for our system. Overall, our system outperforms the LSI with RMSE of 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 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. Another explanation is that the “correct” answers do not make use of the structures either - the users are not given with much contextual and referential information in the survey for a complete understanding of the two regulations in comparison.

The second part of this chapter deals with the results obtained by comparing regulations from different sources. The comparisons are divided into five groups: 1) ADAAG vs. UFAS, 2) UFAS vs. IBC Chapter 11 on accessibility, 3) UFAS vs. UK and Scottish accessibility codes, 4) drinking water regulations from 40 CFR vs. 22 CCR, and 5) 40 CFR on drinking water control vs. IBC Chapter 9 on fire protection systems. The average similarity scores of each group based on different feature matching are tabulated and compared. One to two examples are drawn from each group to illustrate the use of different features and score refinements.

Identical or almost identical provisions are found in Groups 1, 2 and 4. Identical provisions are expected in Group 1, since both the ADAAG and the UFAS are Federal accessibility requirements prepared and published by the same agency, the Access Board. In Group 2, the almost identical provisions are paraphrase of one another. For Group 4, the subject of enforcing agency changes between the EPA and the California DHS among the “identical” provisions; indeed it is not uncommon for different agencies to directly adopt provisions from one another. Non-identical provisions are also retrieved in

different groups of comparisons, and some interesting results are shown. For instance, an example is given in Group 1 where the provisions are found to be related through reference distribution. Another example in Group 3 shows the hidden similarity between the American phrase “door hardware” and the chiefly British phrase “door furniture” identified through neighbor inclusion.

Comparing the similarity scores of different groups, Group 2 shows the highest score which is potentially biased. It is clear that index matching returns a much higher score among Group 2, which is possibly due to the fact that the list of indices are from the IBC. As a result, the index terms appear more often in the IBC which leads to a biased score in Group 2. Looking into the score rankings of each group, the degree of similarity is much higher among Group 1 than the rest of the groups, which makes sense as both are Federal disabled access regulations. The average similarity score of Group 4 is relatively small, possibly due to the significantly larger size of documents in comparison and the diversity of topics covered in drinking water regulations. Compared to accessibility codes, drinking water standards cover a lot of topics and are therefore not as focused, which results in a smaller average similarity score. Group 3 compares the UFAS with two European codes, and the results are not as similar as the UFAS compared to other American regulations. This is easy to understand, since there are obvious spelling differences, such as “curb” versus “kerb,” as well as terminological differences, such as “bathroom” versus “WC.” Finally, Group 5 is the outlier in the table of comparisons, with similarity scores that are mostly zero or orders of magnitude smaller than the rest of the groups. This is the anticipated result, since the regulations in comparison are indeed from two separate domains that are not related to one another.

Different features, such as concepts, measurements, author-prescribed indices, effective dates and drinking water contaminants, are compared as well. Primarily, we observe that term-based features, such as concepts and drinking water contaminants, show a relatively higher similarity score compared to non term-based features. This is understandable as terms form the basis of the body text of provisions, and as a result they occur more

frequently. An example of non term-based feature that does not identify much similarity is the measurement feature. Especially in Groups 2 and 3, measurements play a minor role – for Group 2, the IBC is relatively more performance-based than other regulations, which results in less prescriptive requirements such as measurements. For Group 3, the measurements are quite different between US and European standards, as the units are different. On the other hand, drinking water contaminants help retrieval of related sections among drinking water regulations, such as in the example given where TTHM is matched with “total trihalomethanes.” This further echoes the importance of domain knowledge.

Aside from comparisons among regulations, the last part of this chapter demonstrates a potential system application on e-rulemaking. The problem introduced by e-rulemaking, namely the vast amount of public comments received through the Internet, is briefly discussed using a recent e-rulemaking scenario as an example. We then apply our system on the comparisons between a drafted regulation and its associated public comments. Several interesting examples are noted, where individuals commented on different topics that are both related and not related to the draft. Limitations are also observed, where comments that deal with the general intent of the drafted rules are proved to be difficult to analyze. By screening through the public comments and sorting them according to their relatedness to provisions in the draft, it helps rule makers to review and revise the draft based on the public comments.

In this chapter, a performance evaluation model is developed and used to measure our system performance compared to that of traditional techniques. An assortment of results is obtained and analyzed based on the comparisons of different sources of regulations, such as the Federal government, the State government, private organization and European agencies. Potential system application is demonstrated on the e-rulemaking domain. Some future research directions and potential future tasks are mentioned in the next chapter.

Chapter 5

Conclusions and Future Works

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 statutes and back to past or related rules, facilitating improved understanding of legal requirements [30].” In this chapter, we will give a brief summary of the developed relatedness analysis system that links relevant provisions to one another. Based on the prototyped framework, some future research directions are described.

5.1 Summary

This thesis addresses some of the difficulties in dealing with government regulations such as national and regional codes. The existence of multiple jurisdictions often leads to multiple documents that need to be located and consulted for compliance requirements. In addition, regulations from different sources sometimes impose different or conflicting

requirements. Thus, there is a need for a regulatory infrastructure that promotes understanding, retrieval and analysis of regulatory documents. We proposed and developed a regulatory repository and a relatedness analysis framework, where the performance, results and potential application of the analysis are evaluated.

We developed a regulatory repository to consolidate different formats of regulations, such as HTML or PDF, into an XML format. Analysis tools can be built on top of this XML framework. A shallow parser is developed for this task, which uses a combination of handcrafted rules and text mining tools to structure regulatory documents into the designed XML format. Feature extraction is performed, where features are encapsulated as XML elements in provisions where they appear.

Based on the developed XML repository for regulations, the theory and implementation of a comparative analysis between regulatory provisions are 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 presented the mathematical formulation using a matrix notation.

Finally, performance evaluation is conducted through a user survey, where results obtained using our system are compared with results from traditional retrieval models. 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 on 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. Limitations are also observed, with some of the potential future work suggested in the following section.

5.2 Future Directions

The development of a relatedness analysis framework is only the beginning of many applications of IT on semi-structured documents, such as government regulations. In this section, three major research directions are described. First, several domains comprised of semi-structured documents are suggested as potential application areas of the developed tool. We will then outline some natural improvements to the current development, such as a deeper understanding of concept semantics through definitions. Finally, we will discuss the implications of our tool on the making of regulations and some suggested usage to streamline the rulemaking process.

5.2.1 Applications of A Semi-Structured Document Analysis Tool

The current system is designed for the domain of accessibility and further applied on the domain of drinking water standards. To allow for a more general application of regulatory comparisons on other domains, we can develop a more complete specification for semi-structured document representation in XML. For instance, the current implementation of the XML regulatory framework does not include tables, figures or equations. A complete XML standard should include all potential elements in regulations, where a formal representation format is needed for tables, figures and equations, especially in an engineering domain. For instance, we observe that a handful of figures are used to illustrate dimensions regarding wheelchairs and their access in accessibility regulations. Tables are used predominantly to specify chemical concentration requirements in drinking water standards.

Apart from assisting rule makers, interested and affected citizens to understand regulations, our tool can be used to aid legal research in law firms as well. In particular,

large corporations operating in multiple jurisdictions often need to conduct a so-called “50 state survey of the law” to identify and analyze different legal requirements on different topics. Assuming that a lawyer starts out with the jurisdiction that he or she is familiar with, our tool can be used to compare and bring together related materials from other jurisdictions for the task.

Another common task for lawyers is to do a historical research on legislation, which involves identifying how a particular provision evolved over time in past laws or bills. Currently, legislative historical research can be laborious. Using a relatedness analysis tool, one could automatically identify relevant pieces of legislative history given the corpus of regulatory documents.

Apart from applications on a legal domain, the proposed analysis technique is indeed general and can be applied to other semi-structured documents with a similar hierarchical and referential structure. Examples include user manuals or software specifications, which are often organized into chapters, sections and subsections with references within sections as well. A different set of domain-specific features will need to be identified aside from generic features such as concepts.

5.2.2 Improving the Analysis

Based on Tiebout’s theory of local expenditures [99], it is hypothesized that “if the consumers move to the community whose law happens to fit their preference pattern, they will be at their optimum,” and the relationship between market access and regulatory competition is studied [7]. Combined with the observation of cross-border data transfer laws [11, 88], tools to analyze regulatory competitions in different jurisdictions are needed. In order to analyze regulations in different jurisdictions, especially among the polyglot countries in the EU where regulatory competitions are at the peak, our relatedness analysis system needs to be improved. A translation module can be added to translate non-English regulations into English. Terminological differences need to be

resolved. In particular, one of the computational characteristics of regulations is the definition of terms included in most regulatory documents. In this research, the definitions are extracted and encapsulated in the XML regulatory framework. Future work on analyzing regulations can make use of the definitions of phrases to perform a better comparison. Techniques such as concept matching presented in [73] can help to resolve terminological differences using the provided definitions.

In the example shown in Figure 4.10, we observe that the developed system successfully identified the relatedness between the requirements of the chemical Barium in drinking water enforced by the Environmental Protection Agency and the California Department of Health Services. However, the requirements are indeed different – the California agency enforces a more stringent requirement (1 ppm) than the Federal government (2 ppm). Based on our framework, a potential improvement can be envisioned to capture differences between provisions. Assuming that the interested provisions are related, we first apply the relatedness analysis system to identify the most related pairs of provisions, such as the requirements on Barium by the California and Federal agencies. Different features, such as measurements, can be compared individually to capture differences between provisions. This will require a formal definition and formulation of a difference operator between provisions.

5.2.3 Impacts on the Making of Regulations

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 ... if ADAAG is to reference the A18 then the A18

should contain the requirements for this application [6].” 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.

After regulations are drafted, the public is invited to comment on the proposed rules. As suggested in Section 4.4, based on the developed framework, potential research direction in e-rulemaking includes automated forwarding of comments to corresponding personnel in agencies, as well as automated clustering of comments. Linguistic analysis could be investigated to help identify suggested provision revision embedded in comments. An online comment submission portal that allows for commenting per provision in addition to the existing per draft basis could also be valuable.

The focus of this research was to develop a comparative analysis framework for semi-structured documents, with applications to government regulations. In order to prototype an analysis framework, regulatory documents are first consolidated into a standardized XML format. Several computational properties are identified from regulations, and we proved the importance of the identified properties on extracting relatedness between regulations. In this work, we also demonstrated the use of IT on policy making, in particular, the communication between government agencies and the public via comments on proposed rules. The analysis of semi-structured documents, such as government regulations, is undoubtedly a very rich area for future research.

Bibliography

- [1] *Americans with Disabilities Act (ADA) Accessibility Guidelines for Buildings and Facilities*, US Architectural and Transportation Barriers Compliance Board (Access Board), Washington, DC, 1999.
- [2] *Jakarta Lucene*, The Jakarta Project, <http://jakarta.apache.org/lucene>, 2002.
- [3] R. Attar and A.S. Fraenkel. "Local Feedback in Full-Text Retrieval Systems," *Journal of the ACM*, 24 (3), pp. 397-417, 1977.
- [4] R. Baeza-Yates and G. Navarro. "Integrating Contents and Structure in Text Retrieval," *ACM Special Interest Group in Management of Data (SIGMOD) Record*, 25 (1), pp. 67-79, 1996.
- [5] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*, ACM Press, New York, NY, 1999.
- [6] D.C. Balmer. "Trends and Issues in Platform Lift," In *Proceedings of Space Requirements for Wheeled Mobility Workshop*, Buffalo, NY, October 9-11, 2003.
- [7] C. Barnard and S. Deakin. "Market Access and Regulatory Competition," In C. Barnard and J. Scott (Eds.), *The Law of the Single European Market: Unpacking the Premises*, Hart Publishing, Oxford, UK, pp. 197-224, 2002.

-
- [8] C. Baru, A. Gupta, Y. Papakonstantinou, R. Hollebeek and D. Featherman. "Putting Government Information at Citizens' Fingertips," *EnVision*, 16 (3), pp. 8-9, July-September, 2000.
- [9] R.E. Bellman. *Adaptive Control Processes*, Princeton University Press, Princeton, NJ, 1961.
- [10] T.J.M. Bench-Capon. *Knowledge Based Systems and Legal Applications*, Academic Press Professional, Inc., San Diego, CA, 1991.
- [11] D. Bender. "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, 2004.
- [12] D.H. Berman and C.D. Hafner. "The Potential of Artificial Intelligence to Help Solve the Crisis in Our Legal System," *Communications of the ACM*, 32 (8), pp. 928-938, 1989.
- [13] M.W. Berry and M. Browne. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.
- [14] D.P. Bertsekas. *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA, 1995.
- [15] C.M. Bishop. *Neural Networks for Pattern Recognition*, Oxford University Press; Clarendon Press, New York, NY, 1995.
- [16] V.D. Blondel and P.V. Dooren. *A Measure of Similarity Between Graph Vertices: With Applications To Synonym Extraction And Web Searching*, Technical Report, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2002.

- [17] K.D. Bollacker, S. Lawrence and C.L. Giles. "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, 1998.
- [18] R.J. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro and E. Simoudis. "Mining Business Databases," *Communications of the ACM*, 39 (11), pp. 42-48, November, 1996.
- [19] T. Brants and R. Stolle. "Finding Similar Documents in Document Collections," In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Workshop on Using Semantics for Information Retrieval and Filtering*, Las Palmas, Spain, 2002.
- [20] S. Brin and L. Page. "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, 1998.
- [21] *British Standard 8300*, British Standards Institution (BSI), London, UK, 2001.
- [22] S. Brüninghaus and K.D. Ashley. "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, 2001.
- [23] F.J. Burkowski. "Retrieval Activities in a Database Consisting of Heterogeneous Collections of Structured Text," In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, pp. 112-125, 1992.

- [24] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura and I. Silva. "Local versus Global Link Information in the Web," *ACM Transactions on Information Systems (TOIS)*, 21 (1), pp. 42 - 63, January, 2003.
- [25] *California Building Code (CBC)*, California Building Standards Commission, Sacramento, CA, 1998.
- [26] *California Code of Regulations (CCR)*, Title 22, California Office of Administrative Law, Sacramento, CA, 2003.
- [27] C.L.A. Clarke, G.V. Cormack and F.J. Burkowski. "An Algebra for Structured Text Search and a Framework for its Implementation," *The Computer Journal*, 38 (1), pp. 43-56, 1995.
- [28] *Code of Federal Regulations (CFR)*, Title 40, Parts 141 - 143, US Environmental Protection Agency, Washington, DC, 2002.
- [29] C. Coglianese. *E-Rulemaking: Information Technology and Regulatory Policy*, Technical Report, Regulatory Policy Program, Kennedy School of Government, Harvard University, Cambridge, MA, Report No. RPP-05, 2003.
- [30] C. Coglianese. "Information Technology and Regulatory Policy," *Social Science Computer Review*, 22 (1), pp. 85-91, 2004.
- [31] F. Crestani, M. Lalmas, C.J. Van Rijsbergen and I. Campbell. "Is This Document Relevant?... Probably': A Survey of Probabilistic Models in Information Retrieval," *ACM Computing Surveys*, 30 (4), pp. 528-552, 1998.
- [32] C.J. Crouch and B. Yang. "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, 1992.

- [33] D. Crouch, C. Condoravdi, R. Stolle, T. King, V. de Paiva, J. Everett and D. Bobrow. "Scalability of Redundancy Detection in Focused Document Collections," In *Proceedings of the 1st International Workshop on Scalable Natural Language Understanding (ScaNaLU-2002)*, Heidelberg, Germany, May 23-24, 2002.
- [34] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman. "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, 41 (6), pp. 391-407, 1990.
- [35] *Disability Discrimination Act 1995 (c. 50)*, Her Majesty's Stationery Office (HMSO), London, UK, 1995.
- [36] J. Dörre, P. Gerstl and R. Seiffert. "Text Mining: Finding Nuggets in Mountains of Textual Data," In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, pp. 398-401, 1999.
- [37] *Draft Guidelines for Accessible Public Rights-of-Way*, US Architectural and Transportation Barriers Compliance Board (Access Board), Washington, DC, <http://www.access-board.gov/news/prow-release.htm>, 2002.
- [38] S.T. Dumais. "Improving the Retrieval of Information from External Sources," *Behavior Research Methods, Instruments, and Computers*, 23 (2), pp. 229-236, 1991.
- [39] H.A. Edelstein. *Introduction to Data Mining and Knowledge Discovery*, Two Crows Corporation, Potomac, MD, 1999.
- [40] J.O. Everett, D.G. Bobrow, R. Stolle, R. Crouch, V. de Paiva, C. Condoravdi, M.v.d. Berg and L. Polanyi. "Making Ontologies Work for Resolving Redundancies Across Documents," *Communications of the ACM*, 45 (2), pp. 55 - 60, 2002.

- [41] *eXtensible Markup Language (XML)*, World Wide Web Consortium (W3C), <http://www.w3.org/XML>, 2004.
- [42] *eXtensible Stylesheet Language (XSL)*, World Wide Web Consortium (W3C), <http://www.w3.org/Style/XSL/>, 2004.
- [43] B. Falkenhainer, K.D. Forbus and D. Gentner. "The Structure-Mapping Engine: Algorithm and Examples," *Artificial Intelligence*, 41 (1), pp. 1-63, 1989.
- [44] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. "From Data Mining to Knowledge Discovery: An Overview," In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA, pp. 1-36, 1996.
- [45] "Flavored Malt Beverages and Related Proposals; Posting of Comments Received on the TTB Internet Web Site," *Federal Register*, 68 (231), pp. 67388-67389, 2003.
- [46] P. Ganesan, H. Garcia-Molina and J. Widom. "Exploiting Hierarchical Domain Structure to Compute Similarity," *ACM Transaction on Information Systems*, 21 (1), pp. 64-93, 2003.
- [47] A. Gardner. *An Artificial Intelligence Approach to Legal Reasoning*, Ph.D. Thesis, Computer Science, Stanford University, Stanford, CA, 1984.
- [48] E. Garfield. "New International Professional Society Signals the Maturing of Scientometrics and Informetrics," *The Scientist*, 9 (16), 1995.
- [49] D. Gentner and A.B. Markman. "Structure Mapping in Analogy and Similarity," *American Psychologist*, 52 (1), pp. 45-56, 1997.
- [50] M.P. Gibbens. *CalDAG 2000: California Disabled Accessibility Guidebook*, Builder's Book, Canoga Park, CA, 2000.

- [51] D. Gibson, J. Kleinberg and P. Raghavan. "Inferring Web Communities from Link Topology," In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, PA, pp. 225-234, June 20-24, 1998.
- [52] C. Glymour, D. Madigan, D. Pregibon and P. Smyth. "Statistical Inference and Data Mining," *Communications of the ACM*, 39 (11), pp. 35-41, November, 1996.
- [53] R. Goldman, J. McHugh and J. Widom. "Lore: A Database Management System for XML," *Dr. Dobb's Journal*, 25 (4), pp. 76-80, 2000.
- [54] G.H. Golub and C.F. Van Loan. *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
- [55] R. Grishman. "Information Extraction: Techniques and Challenges," In M.T. Pazienza (Eds.), *Information Extraction (International Summer School SCIE-97)*, Springer-Verlag, New York, NY, 1997.
- [56] J. Grosjean, C. Plaisant and B. Bederson. "SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation," In *Proceedings of IEEE Symposium on Information Visualization*, Boston, MA, pp. 57-64, October 28-29, 2002.
- [57] C. Gurrin and A.F. Smeaton. "A Connectivity Analysis Approach to Increasing Precision in Retrieval from Hyperlinked Documents," In *Proceedings of Text REtrieval Conference (TREC)*, Gaithersburg, MD, 1999.
- [58] M.A. Hearst. "Untangling Text Data Mining," In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD, pp. 3-10, 1999.
- [59] T. Hofmann. "Probabilistic Latent Semantic Indexing," In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, California, pp. 50-57, 1999.

- [60] E. Hovy. "Using an Ontology to Simplify Data Access," *Communications of the ACM*, 46 (1), pp. 47-49, 2003.
- [61] *HyperText Markup Language (HTML)*, World Wide Web Consortium (W3C), <http://www.w3.org/MarkUp/>, 2004.
- [62] E. Ide. "New Experiments in Relevance Feedback," In G. Salton (Eds.), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.
- [63] *International Building Code 2000*, International Conference of Building Officials (ICBO), Whittier, CA, 2000.
- [64] K.S. Jones and P. Willett. *Readings in Information Retrieval*, Morgan Kaufmann, San Francisco, CA, 1997.
- [65] J. Kepler. *New Astronomy* (W.H. Donahue, Trans.), Cambridge University Press, Cambridge, England, 1992 (Original work published 1609).
- [66] S. Kerrigan. *A Software Infrastructure for Regulatory Information Management and Compliance Assistance*, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 2003.
- [67] F. Kidder and H. Parker. *Kidder-Parker Architects' and Builders' Handbook*, John Willey & Sons, London, UK, 1931.
- [68] J. Kleinberg. "Authoritative Sources in a Hyperlinked Environment," In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, CA, pp. 668-677, 1998.
- [69] C. Lin, P.J. Hu, H. Chen and J. Schroeder. "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, May 18-21, 2003.
- [70] T. McCarty. "Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning," *Harvard Law Review*, 90, pp. 837-893, 1977.
- [71] D. Merkl and E. Schweighofer. "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, 1997.
- [72] P. Mitra. *An Algebraic Framework for the Interoperation of Ontologies*, Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, CA, 2003.
- [73] P. Mitra and G. Wiederhold. "Resolving Terminological Heterogeneity in Ontologies," In *Proceedings of Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence (ECAI)*, Lyon, France, pp. 45-50, 2002.
- [74] M.-F. Moens, C. Uyttendaele and J. Dumortier. "Abstracting of Legal Cases: The SALOMON Experience," In *Proceedings of the 6th International Conference on Artificial Intelligence and Law*, Melbourne, Australia, pp. 114-122, 1997.
- [75] J. Osborn and L. Sterling. "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, 1999.
- [76] L. Page, S. Brin, R. Motwani and T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford University, Stanford, CA, 1998.
- [77] *Portable Document Format (PDF)*, Adobe Systems Incorporated, <http://www.adobe.com/products/acrobat/adobepdf.html>, 2004.

- [78] M.F. Porter. "An Algorithm for Suffix Stripping," *Program: Automated Library and Information Systems*, 14 (3), pp. 130-137, 1980.
- [79] *Potential Drinking Water Contaminant Index*, US Environmental Protection Agency, Washington, DC, <http://www.epa.gov/safewater/swp/vcontam3.html>, 2003.
- [80] *Proceedings of Business Compliance One Stop Workshop*, Small Business Administration, Queenstown, MD, July 24-26, 2002.
- [81] *Proceedings of the National Conference on Digital Government Research*, Los Angeles, CA, May 21-23, 2001.
- [82] *Proceedings of the National Conference on Digital Government Research*, Los Angeles, CA, May 20-22, 2002.
- [83] *Proceedings of the National Conference on Digital Government Research*, Boston, MA, May 18-21, 2003.
- [84] *Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL 1999)*, Oslo, Norway, June 14-17, 1999.
- [85] *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL 2001)*, St. Louis, Missouri, May 21-25, 2001.
- [86] *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003)*, Edinburgh, Scotland, June 24-28, 2003.
- [87] Y. Qiu and H.-P. Frei. "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, 1993.
- [88] R.L. Raskopf and D. Bender. "Cross-Border Data: Information Transfer Restrictions Pose a Global Challenge," *New York Law Journal*, July 29, 2003.

- [89] E.L. Rissland, K.D. Ashley and R.P. Loui. "AI and Law: A Fruitful Synergy," *Artificial Intelligence*, 150 (1-2), pp. 1-15, 2003.
- [90] J.J. Rocchio. "Relevance Feedback in Information Retrieval," In G. Salton (Eds.), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.
- [91] G. Salton. *The Smart Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliffs, NJ, 1971.
- [92] G. Salton and C. Buckley. "Term-Weighting Approaches in Automatic Retrieval," *Information Processing and Management*, 24 (5), pp. 513-523, 1988.
- [93] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY, 1983.
- [94] E. Schweighofer, A. Rauber and M. Dittenbach. "Automatic Text Representation, Classification and Labeling in European Law," In *Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAAIL 2001)*, St. Louis, Missouri, pp. 78-87, 2001.
- [95] *Semio Tagger*, Semio Corporation, <http://www.semio.com>, 2002.
- [96] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura and N. Ziviani. "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, July 24-28, 2000.
- [97] *Technical Standards*, Scottish Executive, Edinburgh, Scotland, UK, 2001.
- [98] R. Thomson, J. Huntley, V. Belton, F. Li and J. Friel. "The Legal Data Refinery," *International Journal of Law and Information Technology*, 8 (1), pp. 87-97, 2000.

- [99] C.M. Tiebout. "A Pure Theory of Local Expenditures," *The Journal of Political Economy*, 64 (5), pp. 416-424, 1956.
- [100] *Tree Chart*, Bob Lee, <http://www.crazybob.org/>, 2003.
- [101] *Uniform Federal Accessibility Standards (UFAS)*, US Architectural and Transportation Barriers Compliance Board (Access Board), Washington, DC, 1997.
- [102] A. Valente and J. Breuker. "ON-LINE: An Architecture for Modelling Legal Information," In *Proceedings of the 5th International Conference on Artificial Intelligence and Law*, College Park, MD, pp. 307-315, 1995.
- [103] P. Wahlgren. *Automation of Legal Reasoning*, Kluwer Law and Taxation Publishers, Deventer, The Netherlands, 1992.
- [104] J. Wang. *Distributed Information Organization and Management for Hazardous Waste Regulation Compliance Checking*, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 2003.
- [105] K. Wang and H. Liu. "Discovering Typical Structures of Documents: A Road Map Approach," In *Proceedings of the 21st Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 146-154, 1998.
- [106] *Xpdf*, Glyph & Cog, LLC, <http://www.foolabs.com/xpdf/>, 2003.
- [107] J. Xu and W.B. Croft. "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, 1996.

- [108] N. Yabuki. *An Integrated Framework for Design Standards Processing*, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 1992.
- [109] J. Zeleznikow and D. Hunter. *Building Intelligent Legal Information Systems*, Kluwer Law and Taxation Publishers, Deventer, The Netherlands, 1994.