Locating Related Regulations Using a Comparative Analysis Approach

Gloria Lau
Research Scientist
Thomson Findlaw, Inc.
Sunnyvale, CA 94086
glau@stanford.edu

Haoyi Wang
Graduate Student
Stanford University
Stanford, CA 94305-4020
haoyiw@stanford.edu

Kincho H. Law
Professor of Civil and Env. Engr.
Stanford University
Stanford, CA 94305-4020
law@stanford.edu

ABSTRACT
The sheer volume and complexity of government regulations make any attempt to locate, understand and interpret the information a daunting task. Other factors, such as the scattered distribution of the regulations across many sources, different terminologies and cross referencing, further complicate the technical issues in developing a regulation information management system. This paper describes a comparative analysis approach and its potential application to assist locating relevant regulations from different sources. Examples from environmental regulations are employed to illustrate the proposed methodology and framework.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models.*

General Terms
Algorithms, Management, Legal Aspects

Keywords
Relatedness Analysis, Regulatory Comparison, Structural Analysis.

1. INTRODUCTION
Regulations are typically specified by Federal as well as State governmental bodies and are often amended by local counties or cities. Regulations emanating from diverse agencies often overlap; because settings and objectives differ they may be difficult to reconcile. As new issues of public safety or fairness arise new regulations are promulgated and must be integrated in the complex existing regulatory framework. The distributed responsibilities for enforcement and compliance assistance increase the complexity of complying with regulations. The scope of concern and the terminology used to express those concerns differs among agencies and industries.

Since environmental regulations have the force of law, it is important that companies be able to locate, understand, and comply with them. It is also advantageous for society to make these regulations as easy to locate and understand as possible so that the environment is protected to the extent provided by the laws in place. However, many have argued that the "complex, evergrowing and oft-adpating … environmental law is becoming more challenging for practitioners and the judiciary alike" [7]. Furthermore, "there is ample reason to believe that a growing percentage of environmental violations result from a misunderstanding of regulatory requirements or are otherwise unintended" [21].

The burden of complying with environmental regulations can fall disproportionately on small businesses, since these businesses may not have the expertise or resources to keep track of regulations and their requirements [15]. That the requirements of these complex regulations change over time further compounds the problem [21]. As noted in the Washington Post, "Deciphering and complying with federal regulations is a legal and paperwork nightmare for many businesses. To keep pace, some hire consultants to keep track of the applicable health, safety, environmental and equal-opportunity rules" [19]. This burden has been recognized and targeted by legislation designed to address the problem. The Regulatory Flexibility Act (RFA) [16], amended by the 1996 Small Business Regulatory Enforcement Fairness Act (SBREFA) [20], clearly recognizes the information problem facing businesses, particularly small businesses, that must comply with environmental regulations. Although many efforts have been initiated, actual changes in regulation management and dissemination remain a fairly slow process. Advanced ICT technologies and innovative, high quality tools are crucial to further move the regulatory information to the public.

Government regulations are now available on-line but these online portals are primarily designed for displaying information (and often usable only for experienced users). The sheer volume and complexity of this information, coupled with its scattered distribution across many different sources, makes any attempt to locate, understand and interpret the information a daunting task. Some primitive searching capabilities may be provided; however, it remains difficult to locate cross-referenced or related information and to link the information with useful applications, such as compliance assistance. Other factors, such as the high density of inter-referencing within a regulation code and intra-referencing between regulatory documents and the heavy reliance on acronyms, contribute to reducing the readability of the

documents that can be located. Our research objective is to systematically develop formal approaches that will aid locating relevant regulations and assist compliance.

This paper presents a comparative analysis methodology that can be used to search and retrieve regulations as well as to compare regulations from different sources. This paper is organized as follows: Section 2 briefly describes the overall system framework and the development of an XML-based regulatory repository. Section 3 discusses the fundamental methodology employed for comparative analysis of regulations. To illustrate the comparative analysis framework, Section 4 describes example applications in identifying similar provisions from different sources of drinking water standards. Section 5 describes briefly a work-in-progress prototype for locating related provisions across Federal and State regulations. Finally, this paper concludes with a brief discussion on future works in Section 6.

2. SYSTEM FRAMEWORK

The purposes of the regulatory information management (RIM) system are as follows:

- To develop a formal repository to handle diverse regulation files and define a representational structure;
- To develop mechanisms for extracting features and concepts from the regulatory documents and tools for assisting user to identify related regulations;
- To retrieve and compare regulations from different sources on a specific domain topic.

Figure 1 shows the overall framework for the regulatory information management system. There are four basic functions implemented: (1) textual parsing and storage, (2) semi-structured, indexed storage, (3) feature and concept extraction, and (4) comparative analysis and retrieval of related regulatory documents.

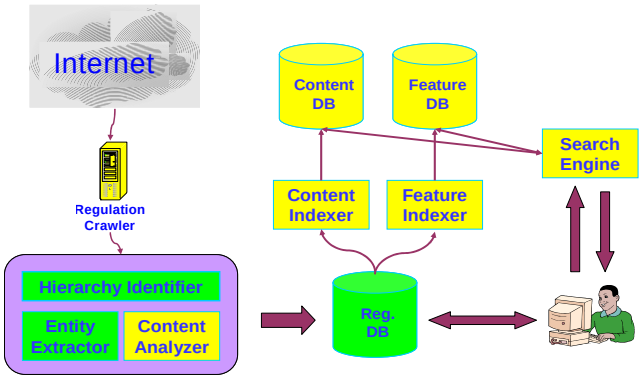

Figure 1. System Framework for the RIM system

To build a repository, the first step is to gather the regulations from diverse sources and transform them into a well-organized XML structure. Since all the State and Federal regulations are now available online, we developed a web crawler to download the raw regulation files from the regulation web sites. Starting with a specific web page containing a regulation, the web crawler is capable of following the links within this web page to further gather other web pages containing the regulations. For each regulation web site, a configuration file is defined that provides the information about the starting URL, the lowest level the crawlers should traverse, the type of links to follow and the downloadable file types. Figure 2 illustrates the configuration file for retrieving the environmental regulations in the State of Hawaii: the "outputDir" asks the crawler to put downloaded files into directory "HI"; "startTOC" indicates the URL of the first web page the crawler should first traverse; "maxDepth" defines the number of levels of "web pages", starting from "startTOC", to be retrieved. For each downloadable web page, the configuration file also defines useful information for traversing the links to find the linked web pages. For example, the variable "linkPattern" defines the pattern to extract useful links from a page; "matchLink" indicates whether the pattern is applied on an embedded link or its anchor text; "filePattern" matches the links to the content; "indexPattern" tells whether a link is an index page. The parameter appended to a variable represents the level number from the start page. Using the web crawler and individually defined configuration file, over a dozen of State

```
outputDir = HI
startTOC = http://www.hawaii.gov/dlnr/AdminRulesIdx.htm
maxDepth=2

linkPattern1 = ^Final.*Rules
matchLink1 = false
filePattern1 = .*/dlnr/.*\\.pdf
indexPattern1 = .*

linkPattern2 = .*
filePattern2 = .*/dlnr/.*\\.pdf
```

regulations have been successfully downloaded and stored in the

```
<regulation id="40.cfr.2" name="PUBLIC INFORMATION"
type="federal">
<regElement id="40.cfr.2.A" name="-- Procedures for
Disclosure of Records Under the Freedom of Information Act ">
<regText></regText>
<paragraph>
Source: 67 FR 67307, Nov. 5, 2002, unless otherwise noted.
</paragraph>
<regElement id="40.cfr.2.100" name=" General provisions. ">
<reqText>
```

repository.

Fig. 2: An Example Configuration File for the Web Crawler

Figure 3: Example of the XML Structure for Regulation

The downloaded regulations are then transformed into an XML structure, which is well suited for representing semi-structured information. The XML structure is designed to map directly to the hierarchical structure inherent in the regulation documents. For example, we use XML tag "regElement" to label a section in the regulation hierarchy. Each "regElement" in XML file may have a parent and/or multiple children elements representing the corresponding sections and subsections. Another advantage of using XML is that metadata can be added easily to the content. To transform the downloaded regulation content (which are typically in HTML, PDF or WORD format), we first convert the file into simple texts, if necessary, using utility tool such as XPDF. A shallow parser, which is written in Perl language, is then employed to transform the text into an XML structure as shown in Figure 3. The hierarchical structure of regulations is preserved by properly structuring provisions as XML elements.

For instance, Section 40.cfr.2.A is a provision in Section 40.cfr.2, and is thus structured to be a child node of the XML element of Section 40.cfr.2. With the hierarchical organization captured in the XML structure, rendering tools can easily be developed to display and view the regulations in its natural organization.

3. COMPARATIVE ANALYSIS

The proliferation of the Internet has led to an extensive amount of research on retrieving relevant documents based on keyword search [2]. Well-established techniques such as query expansions [8, 17] have been deployed to increase retrieval accuracy, with a significant amount of subsequent developments [1, 6, 14, 22] to improve performance. Thus, most repositories are equipped with a search and browse capability for viewing and retrieval of documents. It is reasonable to assume the following in a regulatory repository: at least one relevant document will be located by the user either with keyword search or by browsing through an ontology. In this section, we will discuss the techniques we use to suggest to the users similar provisions from different sources of regulations, starting from a correctly identified section. In essence, we focus on refining the back end comparison technique for documents based on a deep understanding of regulations rather than matching queries at the front end.

Since a typical regulation can easily exceed thousands of pages, a comparison between a full set of regulation and another is meaningless [3]. Instead, a section from one set of regulation is compared with another section from another set, such as a comparison between Section 141.32.e.16 in Code of Federal Regulations Title 40 (40CFR) [5] and Section 64468.1(c) in California Code of Regulations Title 22 (22CCR) [4]. The analysis computes a similarity score, which measures the *degree of similarity* between two documents. The score is defined on a relatedness measurement interval that ranges from 0 to 1, with 0 representing unrelated materials and 1 being the most related or identical materials. The similarity score is denoted by $f(F, C) \in [0, 1]$ per pairs of provisions, for example, pair (F, C) with Section F from the Federal standard 40CFR and Section C from the California code 22CCR. Naturally, the comparison is commutative. In other words, we have $f(F, C) = f(C, F)$.

The similarity score represents the direct content comparison of provisions based on different feature matching. Feature is the evidence of relatedness between two provisions, which could contain domain-specific information. There are generic features that are common across all domains of regulations, such as exceptions, definitions and concepts. For instance, examples of concepts found in drinking water regulations are “ground water” and “levels of exposure.” 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 prescriptive regulations, and chemicals and effective dates specific to environmental regulations.

The similarity score between two sections is computed as a linear combination of the scores obtained using different feature matching, which allows for a combination of generic features, such as concepts, as well as domain knowledge, such as drinking water contaminants in environmental regulations. This design provides the flexibility to add on features and different weighting schemes if domain experts desire to do so. The scoring scheme for each of the features essentially reflects how much resemblance can be inferred between the two sections based on that particular feature. For instance, concept matching is done similar to the index term matching in the Vector model [18], 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.

Here, our usage of the Vector model differs from generic applications in two ways. Our comparison is on extracted features, such as measurements, but not index terms; in addition, we have a much more selective collection of documents, namely regulations in certain domains rather than a general-purpose corpus. If one desires to incorporate domain knowledge, axis independence no longer holds. For instance, some features are characterized by ontologies to define synonyms. Some features simply cannot be modeled as Boolean term matches due to their inherent non-Boolean property, such as measurements, (As an example, a domain expert can potentially define a measurement of “12 inches maximum” as 75% similar to a measurement of “12 inches.”) Some domain-specific features are supplemented with feature dependency information defined by knowledge experts, who do not necessarily agree with a Boolean definition. It is unrealistic to assume that the world can be modeled as a Boolean match, and as a result, domain knowledge is potentially non-Boolean. In essence, the degree of match between two features is no longer limited to only 0% or 100%.

To accommodate a non-Boolean degree-of-match algorithm, we propose a vector space transformation based on the Vector model. For features with defined synonyms or a non-Boolean matching scheme, the feature vectors are mapped onto a different vector space before a cosine comparison. A linear transformation in the form of $\vec{m}' = D\vec{m}$, where D denotes the transformation matrix, is employed to account for axis dependencies introduced by user-defined partial match algorithms. In other words, D captures available domain knowledge, and projects the feature vector $\vec{m}$ onto an alternate space where the resultant vector $\vec{m}' = D\vec{m}$ represents the consolidated feature frequencies. Details and proofs of the formulation are given in [10]. The transformation is shown to produce consistent results when synonymic information are modeled using two different spaces, namely the original n-dimensional space and a reduced vector space with the synonymic feature axes collapsed into one.

4. SIMILAR PROVISIONS FROM DIFFERENT SOURCES – EXAMPLE FROM DRINKING WATER STANDARDS

To illustrate the use of domain knowledge such as ontological information and the associated vector space transformation, we will discuss one particular example of feature extraction and matching here. We focus on drinking water standards in environmental regulations, where 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 [13]. 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 4. 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 dictionary[1].

```
!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 4: Ontology Developed on Drinking Water Contaminants

To incorporate this piece of domain knowledge, our XML parser takes the ontology as a flat list and tags the drinking water contaminants as `<dwc>` subelements in provisions where they appear. As shown in Figure 5, stemming and frequency counting are performed as in `<concept>` and `<index>`.

```
<dwc name=”arsen” times=”1” />
```

The terms or phrases, such as “arsenic”, might be extracted as a concept already, however its sheer presence in the dwc list adds to its importance in this particular domain. Using the ontological information as shown in Figure 4, feature matching can now identify important vocabularies in the domain of drinking water regulations. Similarity computation is enhanced with synonymic information such as phrases like “disinfection byproducts” and “d/dbp”. The transformation matrix *D* would represent the ontology in this example, and the consolidated frequency vector $\vec{m}'$ would contain the consolidated frequency counts of synonyms. The similarity computation would count the frequency of “disinfection byproducts” combined with “d/dbp” on the same feature axis.

[1] 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.

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 5: Drinking Water Contaminant and Effective Date Tags

As a result of the similarity analysis, related provisions can be retrieved and recommended to users based on the resulting scores. Different combinations of features, different feature weights, and different feature scoring schemes can be experimented. Preliminary results are shown using an equal weight of concepts, measurements, drinking water contaminants, and effective dates in the domain of drinking water regulations.

Our comparison system is tested on different groups of regulations, such as comparisons among accessibility regulations, comparisons among drinking water standards, and cross domain comparisons. The average similarity scores among drinking water regulations are relatively small compared to that of accessibility regulations, possibly because of the volume and diversity of coverage of drinking water regulations. Comparing different features among drinking water standards, similarity appears to be captured mostly by concepts. 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 drinking water contaminants, also result in average similarity scores bigger than those obtained using other features such as measurements. Overall, the use of an ontology to help identify synonyms seems to boost the retrieval of similar sections. 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 6, is a top ranked pair of related provisions on drinking water control of the chemical Barium required by the 40CFR and 22CCR. This pair of provisions is actually identical in text except the subject of

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 6: Direct Adoptation of Provisions Across Federal and California State on the Topic of Drinking Water Standards

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.

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 40CFR makes use of many chemical acronyms, such as TTHM, whereas the full term "total trihalomethanes" is always spelled out in the 22CCR. Figure 7 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 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 fail to find anything in 22CCR but only in 40CFR.

To show the dissimilarity between different domains of regulations, we compared drinking water standards 40CFR with fire protection standards in Chapter 9 of the International Building Code (IBC) [9]. 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.

One example is shown below in Figure 8, where provisions from the two separate domains share some remote similarity. Section 141.85.a.1.iv.B.6 from the 40CFR is a small subsection under Section 141.85 on "public education and supplemental monitoring requirements." This section happens to touch on the safety of

Code of Federal Regulations Title 40

141.132.a.2 [No Title; under Monitoring Requirements]

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.

California Code of Regulations Title 22

64823(e) [No Title; under Field of Testing]

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 ploychlorinated 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**.

Figure 7: Terminological Differences Between Federal and State Regulations on the Topic of Drinking Water Standards

Code of Federal Regulations Title 40

141.85.a.1.iv.B.6 [No title; under Public Education and Supplemental Monitoring Requirements]

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.

International Building Code, Chapter 9

907.2.8.1 Fire Detection System

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.

Figure 8: Remotely Related Provisions Identified from a Drinking Water Regulation and a Fire Code

electrical systems in public education. Section 907.2.8.1 from the IBC 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.

5. LOCATING SIMILAR REGULATIONS

As demonstrated in the previous sections, the knowledge-driven comparative analysis approach is potentially capable of discovering similar regulatory provisions. In an ongoing work, we are extending the methodology and framework to develop a "regulatory locator" for domain specific applications. While regulations on a specific domain are mostly grouped under a specific title or part(s), related regulations also exist in other titles or parts. For example, "mercury", a specific chemical, which appears in 40.CFR.141 (Part 141 of Title 40 of CFR), also appears in Title 21 of CFR on Drug and Food Administration. Similarly, the term "mercury", which appears mostly in Title 22 of CCR, also appears in Titles 17 (Public Health), 8 (Industrial Relations), 3 (Food and Agriculture) and other parts of CCR. To fully locate "all" regulations for a specific hazardous substance is a very difficult task. Current attempts to classify industry related regulations are mainly done manually. Our objective is to apply the comparative analysis framework and study the feasibility of extending the tool to facilitate the development of "regulatory locators" (RegLocator) for different domain specific application.

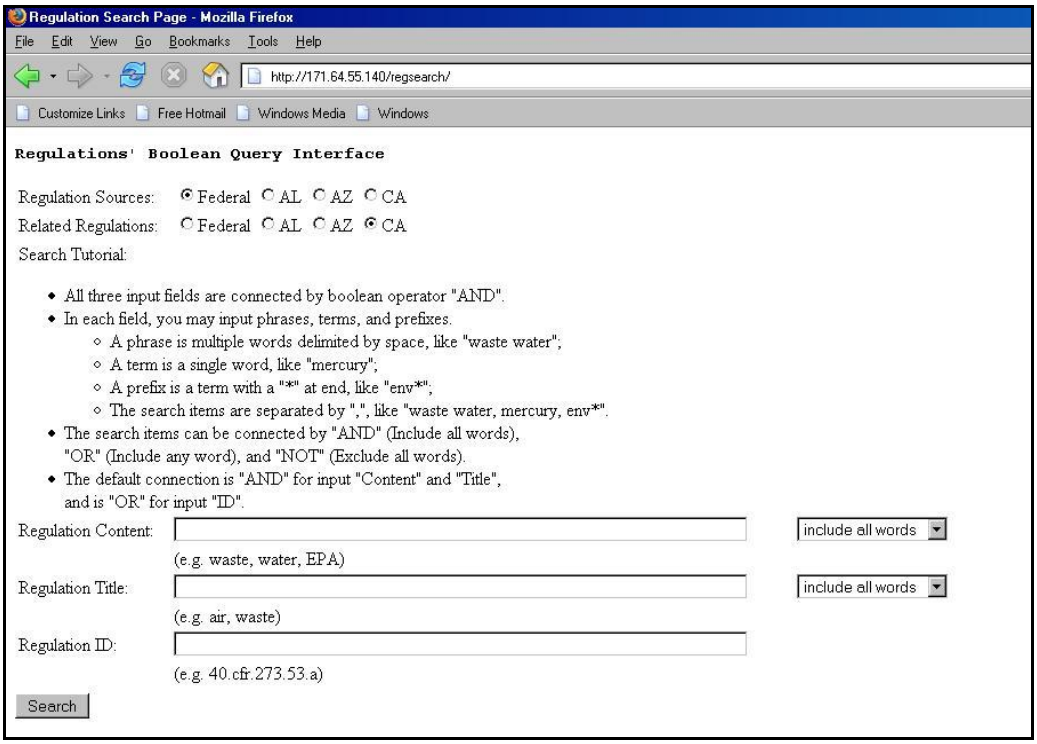

Figure 9: GUI for RegLocator – A Prototype System for Regulation Locator

To enable users to quickly search and find regulations of interests, a search system that utilize terms, concepts and structural relationships, has been implemented. Figure 9 shows the GUI for the RegLocator prototype system. The user can define the primary source of interest for finding the regulations for a particular subject. At the same time, the user can also specify a secondary source to locate possibly related regulations. For instance, if the user is interested in "waste water" in the Federal code, the user can also find related regulations from the secondary source, such as California. This feature is implemented through both a search mechanism and the comparative analysis system. Currently, the environmental codes (which were obtained using the web crawler and text parser discussed earlier) in the RegLocator repository include the Federal (CFR) and three States (Alabama, Arizona and California) regulations.

Figure 10 shows the search results from the query "waste water" from the Federal regulations. Additionally, a set of related (domain) concepts are also shown that also indicate their "relevance" with the query's key words. The related concepts could potentially be useful for appending the terms to the previous query to form a new query or for providing hints when issuing a new query.

From the search results, the user can browse and retrieve a specific provision of interest (see Figure 11). Besides the text of the provision, other provisions located in the "vicinity" of the retrieved provision are also shown in a hierarchical structure (which reflects the typical structure of a regulatory document). Furthermore, related concepts and terms for the provision are also shown to enable searching for further results. Last but not least, related regulations from the secondary source are also shown. User can then search and browse the related regulations, possibly for comparison purposes.

6. DISCUSSION

In developing a regulation information management (RIM) system that would allow searching, retrieving and comparing regulations, domain knowledge plays a very important role in understanding regulations and the relationships between them. We believe a knowledge driven approach, combining with similar analysis, is a powerful way to develop the RIM system. In particular, distinct knowledge sources or regulations do not have to be made completely consistent, only the terms and the concepts that *articulate* their application connections are involved. In this study, we demonstrated the use of an ontology to match features in drinking water standards. With our current XML repository of environmental regulations from the Federal and States, the RegLocator can be enhanced to incorporate domain knowledge to help retrieval of related provisions from different states as well as matching keywords. For instance, users typing in "disinfection byproduct" will now be able to locate provisions written using the acronym "dbp." To this end, we plan to collect and to develop, by way of collaboration with industry experts, ontological information relating to other sub-domains within environmental regulations. We also plan to study and to implement ontological composition [12] once a satisfactory set of ontologies is developed.

We have partially evaluated the performance of the similarity analysis system by comparing results from our system to that of a traditional retrieval system. Preliminary study shows that our system outperformed a traditional index term analysis, especially with the use of domain knowledge [10,11]. A formal evaluation

is planned to estimate the precision and recall of the RegLocator system. However, due to the size of the regulatory repository and the complexity of the law, we plan to scope the evaluation to drinking water standards in a few selected states. A traditional bag-of-word Vector model will serve as the baseline, and we will explore different combinations of related concepts and related provisions to improve the accuracy of a keyword search.

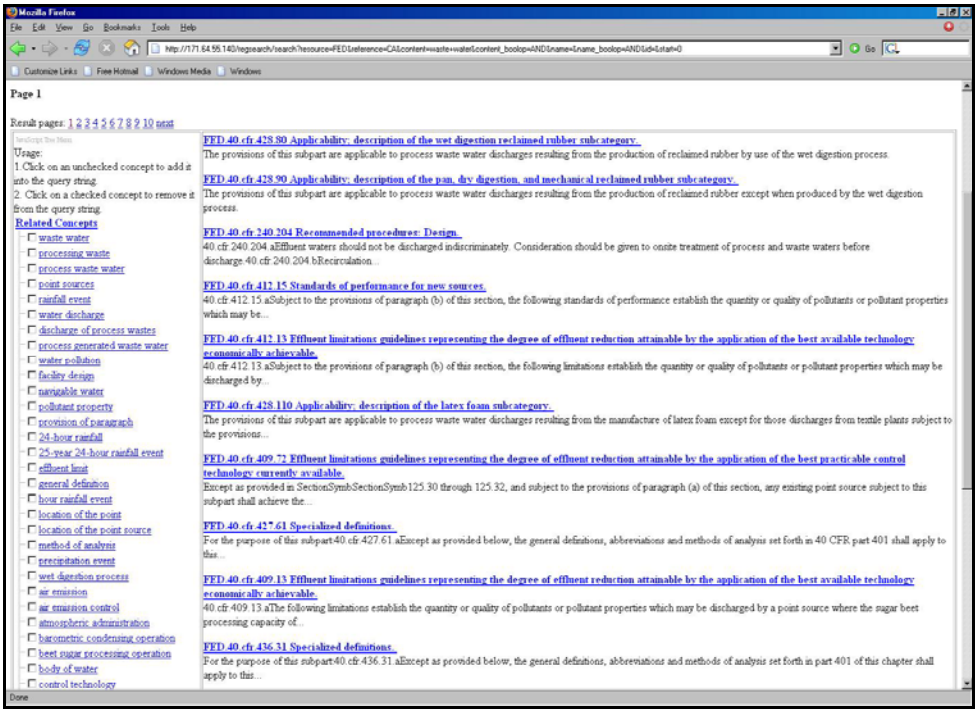

Figure 10: Search Results and Related Concepts

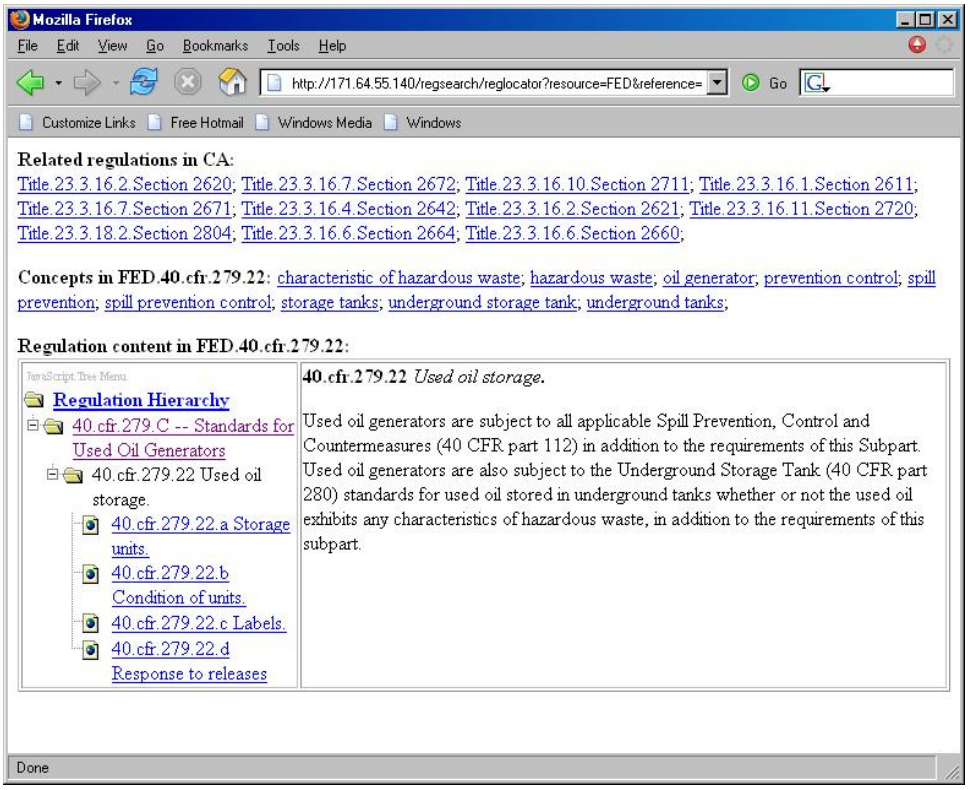

Figure 11: Regulation Displayed and Related Provisions in Secondary Source

7. ACKNOWLEDGMENTS

This research project is sponsored by the National Science Foundation, Grant Numbers EIA-9983368 and EIA-0085998. The authors would like to thank Mr. Bill Labiosa for developing the ontology for the drinking water contaminants. The authors would also like to acknowledge an equipment grant from Intel Corporation.

8. REFERENCES

[1] R. Attar and A.S. Fraenkel. "Local Feedback in Full-Text Retrieval Systems," *Journal of the ACM, 24 (3)*, pp. 397-417, 1977.

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

[3] L.K. Branting. "Reasoning with Portions of Precedents," In *Proceedings of the 3rd International Conference on Artificial Intelligence and Law (ICAIL 1991)*, Oxford, England, pp. 145-154, June 25-28, 1991.

[4] *California Code of Regulations (CCR)*, Title 22, California Office of Administrative Law, Sacramento, CA, 2003.

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

[6] 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.

[7] E. L. Dawson and L.L. Davies, "Book Review: Environmental Law And Policy: Nature, Law, And Society. By Zygmunt J.B. Plater, Robert H. Abrams, William Goldfarb, And Robert L. Graham," *Stanford Environmental Law Journal*, Volume 19, Number 2, pp. 469-478, May 2000.

[8] 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.

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

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

[11] G. T. Lau, K. H. Law, and G. Wiederhold. "A Relatedness Analysis of Government Regulations using Domain Knowledge and Structural Organization," (submitted for publication) *Information Retrieval*.

[12] P. Mitra and G. Wiederhold, "Resolving Terminological Heterogeneity in Ontologies," *Proceedings of Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence* (ECAI), Lyon France, 2002.

[13] *Potential Drinking Water Contaminant Index*, US Environmental Protection Agency, Washington, DC, 2003.

[14] 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.

[15] C. Rechtschaffen, "Competing Visions: EPA And The States Battle For The Future Of Environmental Enforcement," *Environmental Law Reporter*, 30 Envtl. L. Rep. 10803, September 2000.

[16] *Regulatory Flexibility Act (RFA)*, 5 U.S.C. §§ 601 et seq, 1980.

[17] 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.

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

[19] C. Skrzycki, "The Regulators; Compliance Education Goes Self-Service", The Washington Post, May 23rd, 2000.

[20] *Small Business Regulatory Enforcement Fairness Act (SBREFA)*, Pub Law No. 104-121, March 29 1996 (available at http://www.epa.gov/sbrefa/statute.htm).

[21] D. B. Spence, "Paradox Lost: Logic, Morality, and the Foundations of Environmental Law in the 21st Century," *Columbia Journal of Environmental Law*, Volume 20, Issue 1, pp. 145-182, 1995.

[22] 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.