Analyzing Government Regulations Using Structural and Domain Information

To address the difficulties encountered in comparing regulatory documents with multiple authoritative sources, the Regnet project is developing a relatedness analysis system that exploits such documents' unique computational properties.

Gloria T. Lau
Kincho H. Law
Gio Wiederhold
Stanford University

Government regulations, by extending laws with specific guidance for corporate and public actions, provide an important societal benefit. Ideally, they should be intelligible to ordinary citizens as well as rule makers, but the volume of regulations coupled with heavy referencing between provisions limit their accessibility.

Apart from the difficulties in locating and understanding a particular regulation, users often must consult and reconcile multiple authoritative sources. For example, US companies frequently must comply with overlapping federal, state, and local regulations; in addition, some nonprofit organizations publish their own codes of practice. The problem is exacerbated in the European Union, where regulators must harmonize legislation across countries with different languages and traditions.[1]

For enterprises involved in global commerce, regulatory compliance presents a major challenge. For example, a 2003 survey of cross-border data-protection laws revealed that "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."[2]

Retrieving and interpreting particular US government regulations have become easier in recent years. For example, Business.gov, a presidential e-government initiative, aims to guide users "through the maze of government rules and regulations and provides access to services and resources to help you start, grow, and succeed in business." In addition, Regulations.gov provides a national forum for users to comment on existing and pending federal rules.

However, what is needed is a framework that enables individuals and small companies with limited resources to retrieve related regulations from multiple governing copies and then perform comparative analysis. Stanford University's Regnet project (http://eig.stanford.edu/regnet) seeks to develop such a framework, with a current focus on US national and regional codes in the domains of disabled access and environmental standards.

The project's components include an XML repository, a reference extractor, a concept ontology framework, a logic-based compliance assistance system, and a relatedness analysis system.[3,4] We present an overview of the relatedness analysis prototype along with an e-rulemaking example to demonstrate the system's applicability to existing digital government problems.

Published by the IEEE Computer Society

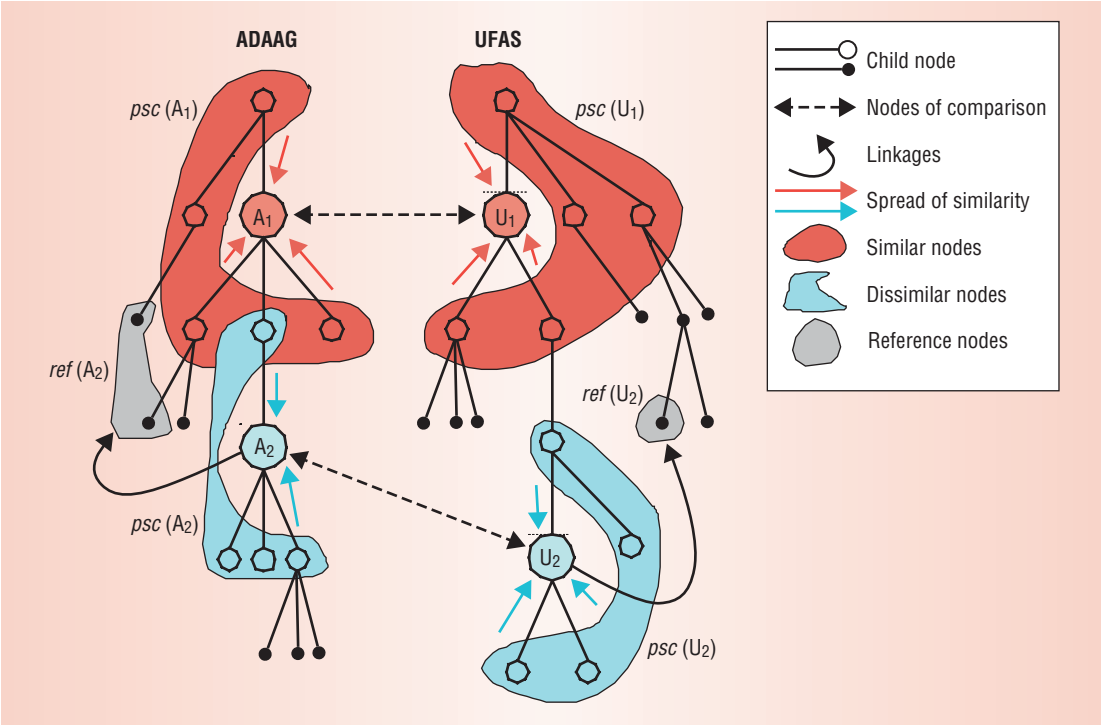

Figure 1. Regulation tree structures. The operator* psc *returns a provision's parent, siblings, and children, while* ref *returns a provision's references. Relatedness between* psc*(A_1) and* psc*(U_1) implies a resemblance between A_1 and U_1, while a comparison of* ref*(A_2) and* ref*(U_2) reveals hidden similarity between A_2 and U_2.

RELATEDNESS ANALYSIS SYSTEM

A typical regulation can easily exceed thousands of pages, making it impractical to compare entire sets of regulations.[5] Instead, our relatedness analysis system compares a provision from one set of regulations with similar provisions from other sets. Given the Internet's ubiquity and the availability of increasingly accurate information-retrieval algorithms, we assume that the average user can locate at least one relevant provision from a regulatory repository either through a keyword search or an ontology. The user inputs this data into the system, which identifies and retrieves related provisions from other regulations.

Regulation properties

In contrast to commercial tools that recommend relevant case laws, such as those offered by online legal resource vendors LexisNexis (www.lexisnexis.com) and Westlaw (www.westlaw.com), our system exploits the intrinsic tree structure, heavy cross-referencing, and domain-centric knowledge shared by most regulations to perform an in-depth comparison.

Unlike typical documents found in generic free-form text corpora, regulations are semistructured documents organized into a tree structure. For example, Section 11.4.5(a) is a subpart or child node of Section 11.4.5, and it is also a sibling of Section 11.4.5(b). This tree hierarchy is crucial to understanding contextual relationships among sections.

Figure 1 shows partial regulation trees for two US federal codes, the Americans with Disabilities Act Accessibility Guidelines (www.access-board.gov/adaag/html/adaag.htm) and the Uniform Federal Accessibility Standards (www.access-board.gov/ufas/ufas-html/ufas.htm). Our system compares pairs of nodes, such as Section 1.2 in the ADAAG, represented as node A_1, and Section 1.1 in the UFAS, denoted as U_1. The operator *psc* returns a provision's parent, siblings, and children—for example, *psc*(A_1) returns the set of immediate neighboring nodes for A_1, while *psc*(U_2) returns the set of immediate neighboring nodes for U_2.

In addition, regulations employ heavy cross-referencing. For example, Section 11.4.5(a) may refer to Section 8.2 for compliance requirements under other conditions. In analyzing and comparing provisions, this type of linkage is important because rules prescribed in one section are only complete if they include references. In our system, the operator *ref* returns a provision's references—for example, in Figure 1, *ref*(U_2) returns the references within U_2.

Finally, regulations are domain centered. For example, the UFAS focuses exclusively on disabled access. In understanding regulations, common sense or dictionary knowledge cannot replace domain knowledge—"lift" and "elevator" are syn-

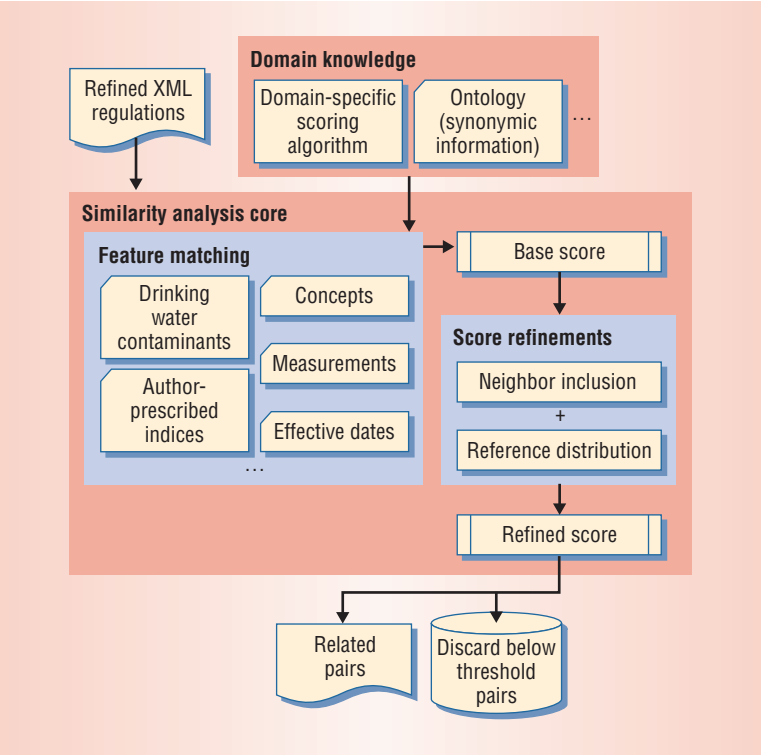

Figure 2. Relatedness analysis system. A combination of information-retrieval techniques, feature matching, and document structure analysis identifies the most-related provisions across different regulation trees.

onymous in normal English usage, but each have specific legal meanings in the US.[6]

Similarity computation

As Figure 2 shows, our system uses a combination of information-retrieval techniques, feature matching, and document structure analysis to identify the most-related provisions across different regulation trees.

Feature extraction. Using handcrafted rules and commercial text-mining algorithms, the system semiautomatically extracts nonstructural characteristics from regulations that signal relatedness between provisions. These *features* can be either generic, such as concept phrases ("access aisle"), or domain specific, such as measurements ("12 inches maximum"). The flexible design enables domain experts to add new features and different feature-weighting schemes.

Base score computation. The system then computes the relatedness between two provisions based on their shared features using vector matching. For each feature, it calculates a base score, between 0 and 1, that measures the cosine similarity between matching vectors. For example, a score of 0.80 would indicate relatively high similarity between provisions for a given concept match.

The system associates some features with ontologies to define synonyms, which cannot always be modeled as Boolean term matches. For example, defining "5 inches minimum" as 80 percent similar to "5 inches" would result in a non-Boolean index. To account for non-Boolean domain knowledge, the system employs *vector space transformation*—it maps feature vectors onto an alternate space to obtain a new set of consolidated frequency vectors prior to cosine computation.[4]

Score refinement. The system extracts regulations' structural properties—namely, the tree hierarchy and references between provisions—and uses these to refine the base score. It compares the parent, siblings, and children of the interested sections to include similarities not previously accounted for based on direct comparison. For example, in Figure 1, similarities between $psc(A_1)$ and $psc(U_1)$ imply resemblance between the interested pair A_1 and U_1 on the basis of neighbor inclusion.

Two sections referencing related sections are more likely to be related. Akin to citation and link analysis,[7] our system utilizes regulations' heavy self-referencing structure to further refine the similarity score between two interested sections. For example, in Figure 1, a comparison of $ref(A_2)$ and $ref(U_2)$ reveals hidden similarity between A_2 and U_2. Thus, the final score for two given provisions includes similarities based on common near-tree neighbors and references as well as content.

COMPARING RESULTS AMONG REGULATIONS

We have tested the relatedness analysis system on different sets of regulations and obtained some preliminary results.[4]

Prediction accuracy

One test compared our system with latent semantic indexing (LSI),[8] a traditional information-retrieval model that employs singular value decomposition to capture documents' conceptual content. We analyzed 10 randomly chosen provisions from the ADAAG and 10 from the UFAS, then evaluated the results against a user-survey-based similarity ranking.

Overall, our system outperformed LSI, with a root mean-square prediction error of 22.9 versus 27.4. Individual combinations of features and structural matching resulted in prediction errors ranging from 12.0 to 29.1, most of which are smaller than the error rate produced via LSI. Among accessibility features, measurement results in the lowest error rate. This reinforces our belief in the value of incorporating domain knowledge

into relatedness analysis—especially in the case of regulations such as ADAAG and the UFAS, which prescribe heavily quantified requirements that only measurement features can capture.

On the other hand, structural matching did not noticeably affect prediction error. This may be due to the fact that the 10 randomly selected pairs of provisions happen not to contain heavy referencing—the *ref* operation returned mostly empty sets. Another possibility is that the survey-based "correct" answers did not use the structures either. Time constraints prevented us from requiring participants to comprehend extensive contextual (*psc* nodes) or referential (*ref* nodes) information on the two regulations.

Neighbor inclusion

We analyzed relatedness between two specific sections in the same pair of codes. Section 4.1.6(3)(d) in the ADAAG, which deals with doors, reads as follows: "(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 …" Section 4.14.1 in the UFAS, which deals with entrances, reads as follows: "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 …"

As expected, a pure concept match resulted in a zero base score. However, with nonzero similarities between their *psc* nodes, the system was able to infer some relatedness between the two sections. As Figure 3 shows, it identified related accessible elements, namely doors and entrances, indirectly through neighbor inclusion. The regulations' referential structure likewise revealed hidden relatedness between the provisions.

Domain knowledge

We also analyzed a pair of related provisions on drinking water standards. The US Code of Federal Regulations Title 40 (www.epa.gov/epahome/cfr40.htm) uses many chemical acronyms and abbreviations, such as TTHM for total trihalomethanes, whereas Title 22 of the California Code of Regulations (www.calregs.com) always spells out the full phrase.

A pure concept match resulted in a zero similarity base score between the pair of provisions. However, the system refined the score to 0.49 using a domain-specific feature—drinking water contaminants, with the following associated ontology:

```
!Disinfectants and Disinfection-
byproducts
...
!Chlorine
    +chlorine
    +cl2
    +hypochlorite
    +hypochlorous acid
!Haloacetic Acids
    +haa
!Disinfection Byproducts
    +d/dbp
    +dbp
...
!Total Trihalomethanes
    +trihalomethane
    +tthm
...
```

The ontology identified TTHM as a match to "total trihalomethanes," as well as HAA to "haloacetic acids." Without the inclusion of such domain knowledge, a user searching for TTHM or HAA would find the abbreviations in 40CFR but not in 22CCR.

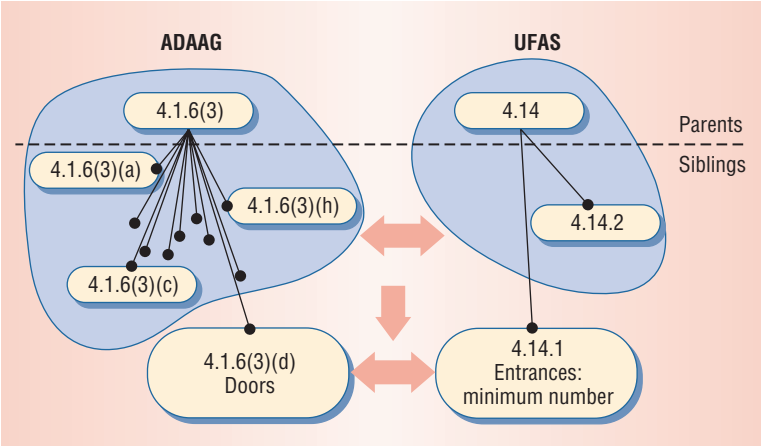

Figure 3. Neighbor inclusion. Despite a zero base score concept match, the relatedness analysis system infers similarity between provisions in the ADAAG and UFAS by identifying related accessible elements, namely doors and entrances, through contextual comparisons.

E-RULEMAKING

Apart from comparing regulatory documents, we have applied the relatedness analysis tool to electronic rulemaking. US government agencies are required to invite public comment on proposed rules, and increasing use of the Internet for this process has resulted in vast quantities of electronically submitted responses. For example, the Alcohol and Tobacco Tax and Trade Bureau (TTB)

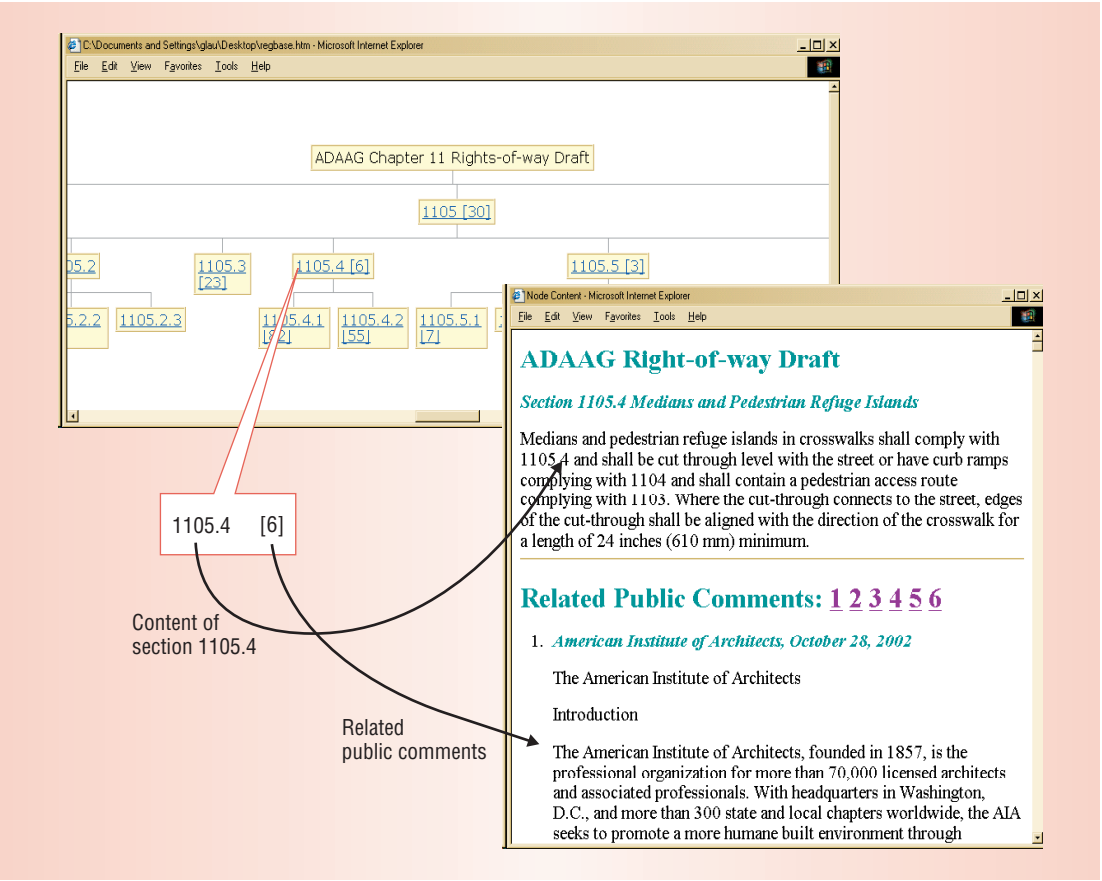

Figure 4. Comparing draft guidelines with public comments in e-rulemaking. Users can follow the links in the tree structure to view both content and related commentary.

received more than 14,000 comments on a recent flavored malt beverages proposal, mostly e-mail, within seven months.[9]

Sorting through and organizing this much data is a nontrivial task. For example, the call for public input on the malt beverages proposal included the 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, later the TTB found it infeasible to individually delete this information from so many submissions, and instead reviewers concerned about their privacy had to formally request its removal. As such, a seemingly effortless electronic submission process turned into a massive data-processing problem.

To assess whether our relatedness analysis system can help agencies process public comments on proposed regulations, we compared a draft chapter in the ADAAG on rights of way with public comments received by the US Architectural and Transportation Barriers Compliance Board. As Figure 4 shows, system users view the draft regulation's tree structure, with each node representing a different section or subsection. Bracketed numbers indicate related public comments—in this case, Section 1105.4 has six comments. Users can follow the links to view both content and commentary.

The following example illustrates a typical section in the same draft chapter, which establishes the requirements for 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.

Below is a related public comment:

> This letter is in response to the draft of the Access Board regarding the new ADA rules....

2. One ADA parking space per block face: In Illinois, motorists with ADA placard or license plates can park free of charge at any meter and can exceed the parking duration in time limit zones. High turn-over meters offer excellent opportunities for ADA access. Municipalities should be allowed flexibility in providing ADA on-street parking spaces. An average of space per block face is more flexible.

3. Parallel ADA spaces to have 5 ft....

Despite the lack of structural features in public comments, such as neighbors and references, the relatedness analysis system was able to retrieve several reviewer suggestions regarding accessible parking spaces based on shared nonstructural features. For example, as the above comment illustrates, respondents often adopted the draft guideline's technical vocabulary, such as the term "block face."

These and other tests have clearly revealed the benefits of automating the e-rulemaking process. Otherwise, conducting full content comparisons of proposed rules and public comments is extremely labor-intensive, especially when each comment might contain several points related to different provisions in the draft.

Information technology can help streamline regulatory policy development in many ways. For example, Harvard's Cary Coglianese has suggested integrating rules with other laws and then using IT to "link all the traces of a rule's history, both back to the underlying statute and back to past or related rules, facilitating improved understanding of legal requirements."[10]

The Regnet relatedness analysis system takes a major step in this direction by linking provisions to relevant counterparts in other regulations as well as draft provisions to related public comments. Apart from assisting rule makers and interested citizens in understanding regulations, the tool can also serve as a research aid. Large corporations often conduct a 50-state survey to identify different legal requirements,[11] and lawyers can use the system to analyze related data from various jurisdictions. It also facilitates historical research on legislation, a common but laborious task that involves tracing a particular provision's evolution over time.

In developing the Regnet framework, we have observed the need to capture differences as well as relatedness between provisions. For example, our tool identified similarities between US and California regulations regarding barium in drinking water, but the state agency enforces a more stringent requirement than the federal government. Incorporating this capability will require a formal definition and formulation of a difference operator between provisions, and we plan to study semantic overlaps, completeness, and conflicts of regulations as the next step. ■

Acknowledgments

The Regnet project is sponsored by the National Science Foundation, under contract numbers EIA-9983368 and EIA-0085998. The authors acknowledge an equipment grant from Intel Corporation and the support of Semio Corporation in providing software for this research.

References

1. E.L. Rissland, K.D. Ashley, and R.P. Loui, "AI and Law: A Fruitful Synergy," *Artificial Intelligence,* vol. 150, nos. 1-2, 2003, pp. 1-15.
2. R.L. Raskopf and D. Bender, "Cross-Border Data: Information Transfer Restrictions Pose a Global Challenge," *New York Law J.,* 29 July 2003; www.whitecase.com/publications/pubs_detail.aspx?pubid=2358&type=Articles.
3. S. Kerrigan, "A Software Infrastructure for Regulatory Information Management and Compliance Assistance," doctoral dissertation, Dept. Civil and Environmental Eng., Stanford Univ., 2003.
4. G. Lau, "A Comparative Analysis Framework for Semi-Structured Documents, with Applications to Government Regulations," doctoral dissertation, Dept. Civil and Environmental Eng., Stanford Univ., 2004.
5. L.K. Branting, "Reasoning with Portions of Precedents," *Proc. 3rd Int'l Conf. Artificial Intelligence and Law,* ACM Press, 1991, pp. 145-154.
6. D.C. Balmer, "Trends and Issues in Platform Lifts," *Proc. Space Requirements for Wheeled Mobility Workshop,* Center for Inclusive Design and Environmental Access, 2003; www.ap.buffalo.edu/idea/space%20workshop/papers.htm.
7. E. Garfield, "New International Professional Society Signals the Maturing of Scientometrics and Informetrics," *The Scientist,* vol. 9, no. 16, 1995, p. 11.
8. S. Deerwester et al., "Indexing by Latent Semantic Analysis," *J. Am. Soc. Information Science,* vol. 41, no. 6, 1990, pp. 391-407.
9. "Flavored Malt Beverages and Related Proposals; Posting of Comments Received on the TTB Internet Web Site," *Federal Register,* vol. 68, no. 231, 2003, pp. 67388-67389.

10. C. Coglianese, "Information Technology and Regulatory Policy," *Social Science Computer Rev.*, vol. 22, no. 1, 2004, pp. 85-91.
11. R.A. Leiter, ed., *National Survey of State Laws*, 5th ed., Thomson Gale, 2005.

Gloria T. Lau *is a research scientist at Thomson FindLaw and formerly a research assistant in Stanford University's Engineering Informatics Group. Her research interests include legal informatics, knowledge discovery, and data mining with applications to text documents and semistructured data. Lau received a PhD in civil engineering from Stanford University. Contact her at glau@stanford.edu.*

Kincho H. Law *is a professor in the Department of Civil and Environmental Engineering at Stanford University. His research focuses on the application of advanced computing principles and techniques to structural and facility engineering. Law received a PhD in civil engineering from Carnegie Mellon University. He is a member of the ACM, the American Society of Civil Engineers, and the Society of Industrial and Applied Mathematics. Contact him at law@stanford.edu.*

Gio Wiederhold *is a professor emeritus of computer science, medicine, and electrical engineering at Stanford University. His research interests include the design and operation of large-scale software systems; the design of databases, knowledge bases, distributed systems, and applications in medicine, planning, and business; and privacy and security. Wiederhold received a PhD in medical information science from the University of California, San Francisco. He is a fellow of the IEEE, the ACM, and the American College of Medical Informatics. Contact him at gio@stanford.edu.*