

SGER Collaborative: A Testbed for eRulemaking Data

Project Highlight prepared for dg.o 2004

Stuart Shulman <stuart.shulman@drake.edu>

Drake University

Jamie Callan <callan@cmu.edu>

Carnegie Mellon University

Eduard Hovy <hovy@isi.edu>

USC-ISI

Stephen Zavestoski <smzavestoski@usfca.edu>

University of San Francisco

As text production and communication becomes easier, people gathering and reviewing information-intensive commentary face an increasing challenge. For news analysts, opinion pollsters, advertising planners, and, not least of all, government regulation writers, it is no longer easy to obtain a bird's-eye view of the world, given the rich, interconnected, and sometimes very subtly different texts aimed at them by the public. Thus, where straightforward information retrieval in the past might have been sufficient to enable knowledge workers to manage large amounts of textual information, increasing volumes of text, coupled with increasing expectations on the depth and complexity of analysis, make sophisticated text processing techniques and tools highly desirable, and in some cases imperative. Of particular interest are techniques to identify peoples' opinions, pinpoint groups of stakeholders such as mothers or lawyers, or segment and cross-relate comments according to various criteria of similarity. To a large extent, such tools exist only in embryonic form, and have not yet been integrated or tested in real-world settings.

The process of government regulation writing is an ideal "testbed" in which to experiment with the development, integration, and deployment of such techniques. Government regulation writers are a class of knowledge workers largely unrecognized, yet highly important to modern society. Their function is to formulate, in a tightly scripted procedure, the rules and regulations that define the details of our laws. Thousands of regulation writers employed in approximately 200 federal agencies and other entities have been granted the authority to perform an increasingly complex and vital task. After issuing proposals, agencies have to consider feedback from interest groups, commercial enterprises, and the public at large. If the regulations are to withstand legal challenges, they must incorporate substantive comments accurately.

This project is a small exploratory grant for research (SGER), in which we explore the possibility of a larger grant that would focus on developing and evaluating a **Rule-Writer's Workbench**, in which a variety of text analysis tools can be applied to help analysts, singly or jointly, master a mountain of incoming text to produce comprehensive, detailed, and densely cross-indexed analyses. Our initial work with federal agency personnel suggests that the following capabilities are important:

- **basic information retrieval**, for gathering relevant texts both within and outside the docket;
- **text classification**, for channeling comments to the appropriate rule writer's desk;
- **overall text characterization using word frequency counts**, for identifying key issues;
- **duplicate detection**, for quickly identifying form letters;
- **near-duplicate detection**, for identifying and extracting text changes to form letters;
- **text summarization**, for creating a first rough cut through the data;
- **author typing**, for stakeholder analyses during and after a public comment period;
- **opinion/affect determination**, for determining what stakeholder concerns exist; and
- **document partitioning and cross-indexing**, for connecting comments to sections of regulations.

Controversial regulatory actions invariably result in massive amounts of public comment. Electronic public comment opens the floodgates wider, but in ways that can inadvertently undermine the intent of public participation. In one instance—the USDA's national organic standard—a small team of rule writers in the late 1990s faced the task of sorting manually over a

quarter million public comments. The number of comments for such controversial rules is likely only to increase.

Interest groups have implemented their own information technology (IT) applications to: educate and mobilize supporters, bombard agencies with comments, delay regulatory action, gain favorable publicity, and recruit members. However, such efforts may dilute the voice of the public as agencies face statutory and administrative deadlines to incorporate public input into a defensible final rule. The tools we propose to develop and evaluate may increase the likelihood that all public input is analyzed in a way that does not compromise the importance of public involvement. End users in and out of government will have a rare chance to beta test the technologies that will help make sense of public commentary.

Our preliminary funding (“SGER Collaborative: A Testbed for eRulemaking Data”) produced several new resources that support our research, and are available for use by others studying text mining, text analysis, and/or the policy and political impact large, public comment datasets.

The **eRulemaking Testbed** web site¹ distributes public comment process data for use by the research community. As of this writing, it contains data for eight regulations proposed by the Department of Transportation, the Environmental Protection Agency, and the Department of Agriculture; we constantly solicit new datasets. Proposed regulations, public comments, and final regulations (if any were issued) are available on the site. The site also provides search access to the datasets (requested by some of the agencies that provided the comments) and several simple text analysis tools. Access is controlled, to avoid privacy problems caused by search engines indexing the public comments, but access is granted to any researcher who requests it.

The **eRulemaking Research** web site² is a clearinghouse for conference papers, manuscripts, workshop information, and presentations related to eRulemaking research. Also, the group conducted an **eRulemaking Workshop** at the National Science Foundation in September 2003 with government rule writers (first day) and interest groups (second day) produced a pair of evaluation reports disseminated via the Internet. Other workshops have focused on the government perspective; we believe this was the first time anyone solicited the opinions of the activists and interest groups (that generate most of the form letters). Broader follow up sessions are planned for early June, 2004.

Collaboration between computational and social scientists raises unique challenges, as does work between academic and governmental personnel. Nonetheless, significant inroads have been made toward attaining all these goals over the past two years. Digital government requires interdisciplinary collaboration, to ensure that technological innovation meets the requirements of democratic institutions and traditions. This project fosters collaboration not only through research and publications, but also through regular presentations and workshops that bring together advisory group members, academic researchers, private sector technology designers, and federal agency personnel. According to a National Research Council report, *Making IT Better*, “[n]ontraditional research mechanisms may be needed that will encourage the participation of end user organizations in research, broaden the outlook of IT researchers, and/or overcome disciplinary boundaries in universities.” Our research group activities are one such mechanism.

We intend to continue on this path, recognizing from prior experience the importance of devoting time not only to IT development and evaluation, but to constant and thorough liaison with government partners. The group recently submitted a full proposal to the NSF to perform the work outlined here, with support from the DOT, USDA (APHIS & USFS), EPA, and BLM.

¹ <http://www.cs.cmu.edu/~callan/Data/eRulemaking/>

² <http://www.drake.edu/artsci/faculty/sshulman/eRulemaking/>