

DURIAN: A Demo for Near-Duplicate Detection

Hui Yang

Language Technology Institute
School of Computer Science
Carnegie Mellon University
huiyang@cs.cmu.edu

Jamie Callan

Language Technology Institute
School of Computer Science
Carnegie Mellon University
callan@cs.cmu.edu

Stuart Shulman

Library and Information Science
School of Information Sciences
University of Pittsburgh
shulman@pitt.edu

ABSTRACT

Recently, the move from paper to electronic public comments makes it much easier for individuals to customize form letters while harder for agencies to identify substantive information since there are many near-duplicate comments that express the same viewpoint in slightly different language. The identification of exact- and near-duplicate texts, and recognition of unique text within near-duplicate documents, is an important component of data cleaning and integration processes for eRulemaking.

This brief paper describes a demonstration of a near-duplicate detection system, DURIAN (DUPLICATE Removal In lARge collectioN), that identifies and organizes the near-duplicates for eRulemaking applications.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval] Clustering.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Duplicate detection, clustering, eRulemaking, text analysis.

1. SYSTEM OVERVIEW

DURIAN (DUPLICATE Removal In lARge collectioN) [1] is a system to detect and organize exact- and near-duplicates in notice and comment rulemaking. Near duplicates are generated in large volume when an advocacy organization distributes a form letter by email or posts the form letter on a web page that allows or encourages customization before submission. A regulatory agency would likely want these comments grouped together even though the amount of modified text varies greatly. Our goal in this work is to greatly improve near-duplicate detection accuracy for notice and comment rulemaking as well as to maintain efficiency.

Our research was conducted with two public comment datasets. One dataset, for the EPA's Proposed National Emission Standards for Hazardous Air Pollutants for Utility Air Toxics rule (USEPA-OAR-2002-0056, "Mercury rule"), contains 536,975 email messages. The second dataset, for the DOT's Proposed Average Fuel Economy Standards for Light Trucks rule (USDOT-2005-22223), contains 45,979 comments, including a mixed format of email messages, pdf files and scanned image files. Experimental evaluation of DURIAN's accuracy requires human assessment, which is impractical for the full datasets. Manual evaluation of near-duplicate detection accuracy on samples of each dataset

shows that system-human intercoder agreement is comparable to human-human intercoder agreement [Yang, et al, 2006].

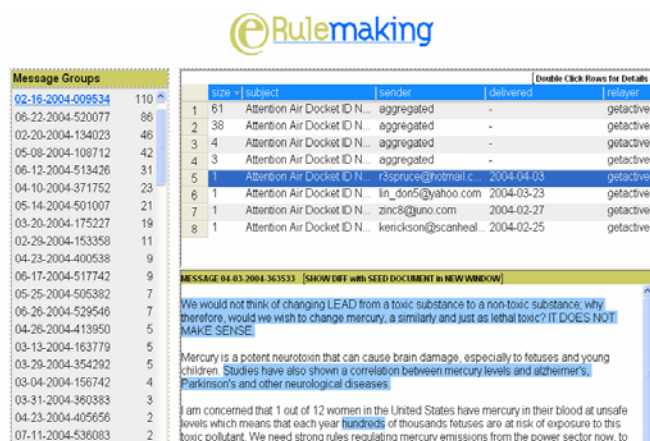


Figure 1: The DURIAN software.

Figure 1 shows a DURIAN window. DURIAN is web-based and publicly accessible with password protection. The system identifies reference copies of form letters, modified copies of form letters (near-duplicates) and how they were modified, and unique comments. The left pane lists the document IDs of reference copies. When a reference document ID is clicked, the upper right pane shows message metadata, such as, subject, author, submitted date, and relayer information. The lower right pane shows the message contents and highlights the modified part of a form letter.

ACKNOWLEDGMENTS

We thank the DOT and EPA for providing the public comments that made this research possible. This research was supported by NSF grant IIS-0429102. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Hui Yang, Jamie Callan, Stuart Shulman, "Next Steps in Near-Duplicate Detection for eRulemaking", In *Proceedings of the 6th National Conference on Digital Government Research (DG.O2006)*, San Diego, California, May 21-24 2006.